AFRL-RI-RS-TR-2015-234

ROBUST SPEECH PROCESSING & RECOGNITION: SPEAKER ID, LANGUAGE ID, SPEECH RECOGNITION/KEYWORD SPOTTING, DIARIZATION/CO-CHANNEL/ENVIRONMENTAL CHARACTERIZATION, SPEAKER STATE ASSESSMENT

UNIVERSITY OF TEXAS AT DALLAS

OCTOBER 2015

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2015-234 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/ S /
MICHELLE R. GRIECO
Work Unit Manager

/ S /
WARREN H. DEBANY, JR.
Technical Advisor, Information
Exploitation and Operations Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
OCT 2015	FINAL TECHNICAL REPORT	APR 2011 – APR 2015

4. TITLE AND SUBTITLE

ROBUST SPEECH PROCESSING & RECOGNITION: SPEAKER ID, LANGUAGE ID, SPEECH RECOGNITION/KEYWORD SPOTTING, DIARIZATION/CO-CHANNEL/ENVIRONMENTAL CHARACTERIZATION, SPEAKER STATE ASSESSMENT

5a. CONTRACT NUMBER
FA8750-12-1-0188

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
35885G

6. AUTHOR(S)

John H. L. Hansen

5d. PROJECT NUMBER
COMB

5e. TASK NUMBER
JU

5f. WORK UNIT NUMBER
TD

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Texas at Dallas
Erik Jonsson School of Engineering & Computer Science EC32
P.O. Box 830688
Richardson, Texas 75083-0688

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIGC
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2015-234

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This study has focused on five complementary research tasks in the domain of audio, speech, language, and speaker recognition and processing. In the area of speaker recognition/identification (SID), advancements have been realized to address acoustic mismatch due to speaker overlap, language mismatch, channel/microphone/additive noise, speaker style (spoken vs. singing), speaker state (physical task stress), distant speech, and environment based (room reverberation). In language ID (LID), advancements have been shown for improved out-of-set language rejection, as well as integrated spectral and prosody based LID solutions. For co-channel and diarization, new algorithms based on gammatone subband frequency modulation was achieved. In diarization, robust speech activity detection based on a combination (Combo-SAD) feature stream was developed. New keyword spotting technology using phonological features as well as audio stream assessment for peak clipping and speaker height estimation were also developed. All algorithms were evaluated on various speech corpora from AFRL, CRSS-UTDallas, and publicly available.

15. SUBJECT TERMS
Speaker recognition, Language Recognition, Overlap Speech Detection, Diarization, Automatic Speech Recognition, Keyword Spotting, Speaker Analysis, Speech Analysis, Audio Waveform Analysis, Machine Learning, Audio Analysis and Information Extraction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**MICHELLE R. GRIECO**
U	U	U	SAR	134	19b. TELEPHONE NUMBER (*Include area code*) **N/A**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 SUMMARY

The overall goal of this 36 month project has been to formulate advanced speech processing algorithms and develop functional sub-systems that address the following five project goals which have been delivered to the USAF:

(i) **Speaker ID (SID) Robustness**: to address SID for noise, channel, and speaking style including additive noise, communication system/HF modulations issues, channel/microphone mismatch, differences in speaking style (read, spontaneous, distant, whisper, singing), environmental mismatch (room reverberation, microphone placement), SID in language mismatch conditions

(ii) **Open-Set Language ID (LID) / Dialect ID (DID)**: to improve LID/DID based solutions that capitalize on phonological feature and articulatory based speech modeling as well as prosodic structure; to focus on Open-Set LID where the languages to reject could be close or far apart (i.e., either closely spaced reject such as Urdo-Hindi or Russian-Ukrainian, or closely spaced dialects such as Arabic (UAE, Egypt, Iraqi, etc.); wide open language rejection (rejecting Mandarin when trying to recognize Indian languages such as Kanneda, Tamil, Teleglu, etc.).

(iii) **Co-Speaker Diarization/Environment (CoSpkrD):** the ability to detect the presence of co-channel speech for usable speech detection as well as speaker-separation for speech systems; diarization of the same audio streams to tag speaker identify, as well as develop advanced environmental sniffing algorithms to contribute to situation awareness of the voice communications (i.e., number of subjects in the room, size of the room, etc.) as well as system adaptation robustness. This domain would focus on unsupervised methods that could be incorporated into speech and speaker diarization schemes.

(iv) **Automatic Speech Recognition/Keyword Spotting(ASR/KWS)**: formulate next generation techniques for speech recognition which include articulatory, phonological, and prosody based detectors. Such knowledge would be integrated into systems to address problems in speech recognition, as well as keyword spotting (KWS) in new and emerging domains. Robustness methods will be explored to improve performance in real audio data scenarios.

(v) **Speaker State Assessment/Environmental Sniffing (SSA/ EnvS):** develop algorithms to assess new knowledge regarding speakers within audio streams – sub-areas would include physical speaker traits (i.e., height, weight), speaking style scenario/room (i.e., read, spontaneous, whisper, distant based speech), situation space (i.e., environment for which speech is originating – size of the room, number of speakers in the room, etc.), or situational speaker state (i.e., stress state, or general emotional outlook). Such knowledge would contribute to improved speaker ID systems, help partition speech recognition or language ID acoustic models to improve overall classification performance, as well as contribute to improved forensic categorization of subjects.

The 36 month project has been partitioned into four project year periods:

Year 1	Year 2	Year 3	Year 4
(4/24/12 – 9/23/12)	(9/24/12 – 9/23/13)	(9/24/13 – 9/23/14)	(9/24/14 – 4/23/15)
5 months	12 months	12months	7 months

Research has focused on each of the five tasks, and included regular monthly/bi-monthly project summaries and monthly teleconference calls (including a (i) man hours report, and (ii) monthly financial status report), bi-yearly site visits (one at UTDallas and one at USAF) with completed presentations and code/system deliveries every six months. This represents the final technical report of the project.

2 INTRODUCTION:

A. ORIGINAL PLANNED OBJECTIVES: (overall summary of Tasks)
This section highlights the original planned objectives across the 5 research tasks for the 36month period.

Objective Name: Task 1 - Speaker ID (SID) Robustness: Speaker identification is known to be sensitive to mismatch in train/test conditions. Algorithm advancements which capitalize on GMM-UBM supervector, i-Vector, GMM-SVM based modeling/processing will be employed. The major advancements here to address robustness will center on speech signal processing breakthroughs to address the sources of mismatch. These contributions will include novel speech feature development, feature processing, feature and model normalization, alternate universal background modeling (UBM) for out-of-set speaker rejection, speech enhancement and noise suppression schemes for additive and channel or distance based interference. Speaking style differences including read, spontaneous, distant, whisper, singing will be addressed. Since noise is generally assumed to be additive, and channel/microphone affects are generally assumed to be convolutional, one major technical advancement here will be acoustic factor analysis (AFA), where factor analysis is performed in the speech feature domain. This has strategic advantages since it allows one to directly address the two broad sources of distortion simultaneously without being concerned on the interoperability of separate noise versus channel/microphone processing. The additional advantage is that it is expected that the formulation of such a scheme could be directly incorporated into an i-Vector processing scenario. The performance metrics to advance include DET (detection error trade-off) curves, minDCF (minimum decision cost tradeoff scores), EER (equal error rates), and sustained performance as noise, channel, room-reverberation is introduced. The specific quantitative improvements have been based on the evaluation data employed, including the amount of training data, as well as duration of the test data.

Objective Name: Task 2 - Open-Set Language ID (LID) / Dialect ID (DID): The focus here has been to develop new features, mismatch compensation schemes, and classification strategies for language identification. These methods will consider existing audio data provided by USAF, in addition to data available from NIST LRE. Based on the proposed automatic speech recognition advancements from Task#4, we will formulate improved LID/DID solutions that capitalize on phonological feature and articulatory based speech modeling as well as prosodic structure. Historically, LID has been based on a 1-of-N classification task where all potential languages would be known in advance. However, in practical scenarios it is more typical to have a desired language or set of languages which could appear within a larger set of unknown languages. As such, the focus in this domain will be on *Open-Set LID* where the languages to reject could be close or far apart (i.e., either closely spaced rejection languages such as Urdo-Hindi or Russian-Ukrainian, or closely spaced dialects such as Arabic (UAE, Egypt, Iraqi, etc.)). In the case of wide open language rejection, the task would be much easier, such as rejecting Mandarin when trying to recognize Indian languages. Algorithm development will focus on (i) i-Vector systems, (ii) GSM-SVM supervector based LID, (iii) combined Articulatory Feature and Prosody Based System, and (iv) PPRLM systems employing phonotactical speech recognition sub-systems.

Objective Name: Task 3 – Co-Speaker Diarization/Environment (CoSpkrD): the ability to detect the presence of co-channel speech for usable speech detection as well as speaker-separation for subsequent speech systems; formulation of diarization advancement of the same audio streams to tag speaker identify, as well as develop advanced environmental sniffing algorithms to contribute to situation awareness of the voice communications (i.e., number of subjects in the room, size of the room, etc.) as well as system adaptation robustness. This domain will focus on unsupervised methods that could be incorporated into speech and speaker diarization schemes.

Objective Name: Task 4 – Automatic Speech Recognition/Keyword Spotting: (ASR/KWS): formulate next generation techniques for speech recognition which include articulatory, phonological, and prosody based detectors. Such knowledge would be integrated into systems to address problems in speech recognition, as well as keyword spotting (KWS) in new and emerging domains. English as well as other

languages of interest (Arabic, Farsi, south Indian, etc.) will be addressed, as well as robustness methods to improve performance in real audio data stream scenarios. While the emphasis here will be on KWS, these advancements will also be considered for unsupervised open-word-set based Speaker ID (SID) and Language/Dialect ID (LID/DID).

Objective Name: Task 5 - Speaker State Assessment/Environmental Sniffing (SSA/EnvS):
advancements has focused on the development of new algorithms to assess unique knowledge regarding speakers within audio streams – sub-areas would include physical speaker traits (i.e., height, weight), speaking style scenario/room (i.e., read, spontaneous, whisper, distant based speech), situation space (i.e., environment for which speech is originating – size of the room, number of speakers in the room, etc.), or situational speaker state (i.e., stress state, or general emotional outlook). Such knowledge would be used to contribute to improved speaker ID systems, help partition speech recognition or language ID acoustic models to improve overall classification performance, as well as contribute to improved forensic speech analysis and categorization.

Since this project has focused on five parallel tasks, project "Intro", "Methods", "Assumptions/Procedures", "Results/Discussion", "Conclusions", and "References" are separated within each project task.

3 METHODS:

Task#1: SID

In the area of speaker ID, the main methods employed have been i-Vector, GMM-UBM supervector, and GMM-SVM based classifiers. Since mismatch is a primary challenge, various features and compensation schemes have also been employed. These include MFCCs, PMVDR, PLP, etc. as features, and CMN CMVN as cepstral mean and variance normalization. Compensation to address language mismatch in SID, noise/channel issues, session variability, speaking style such as speaking/singing, whisper/vocal effort, and physical task stress have all been considered. Blind spectral weighting is one proposed method which has addressed a number of challenges in the noisy reverberant space for SID. In general, these domains consider a range of methods for analysis of the speech production differences, as well as developing models and compensation of the mismatch. Because the specifics of each area are specialized, the details of these methods are highlighted more specifically in the section entitled Results & Discussion.

Task#2: LID

In the area of language ID, our effort has concentrated on two perspectives, close-set task and open-set task. For the open-set task, the emphasis has been on out-of-set language rejection for LID. The methods employed include phonotactic modeling techniques such as Parallel Phone Recognition with Language Modeling (P-PRLM) and Phone Recognition-SVM (PR-SVM). A state-of-the-art i-Vector approach is also explored given its success in the SID domain. In the domain of phonotactic systems, Deep Neural Networks (DNN) have been popular, and in this area the specific approach explored has been deep belief network (DBN) as the back-end, instead of an SVM, which highlighted clear benefits in system performance. Methods also included various features such as MFCCs, PLP, LFCC, GFCC, PNCC, PMVDR, RASTA-PL, RASTA-LFCC, Multi-peak MFCC, Thomson MFCC, and sine-weighted cepstral estimator (SWCE) (see Table 17). An extensive toolkit, named MS-AcID for "multi-session acoustic ID", is also developed which contains a range of methods from existing approaches (i.e., PLDA, Gaussian Cosine Distance Scoring, Gaussian Backend, etc.) as shown in Fig. 39. The methods in this domain also emphasized the ability to perform data purification for both training and test phases in LID and SID. This is more critical in the domain of LID, since it is quite common to employ "found" data when building models for low resources languages for either in-set or out-of-set scenarios. Such found data will have major mismatch issues associated with microphone, channel, environment, as well as speaking style (monologue, 2-way conversation, etc.) and environmental noise/reverb issues. The methods employed in data purification are

intended to catalog the types of mismatch/noise/conflicting audio content (i.e., commercials, music, etc) contained in audio streams which may be used to build acoustic models for LID/SID applications.

Task#3: Co-Channel/Diarization

For co-channel/overlap speech, there has been far less algorithmic advancements in the past. Therefore, the methods employed from past work is limited. There are perhaps two aspects where overlap speech occurs – cases where there are back-channel confirmations during 2-way conversations (i.e., small intermittent verbalizations, but little long sustained overlap), and the second is when there are clearly two speaker talking simultaneously which confounds virtually all speech technology (i.e., SID, LID, ASR, KWS, diarization, etc.). In this area, the methods explored include gammatone subband frequency modulation features, pychnogram based time-frequency analysis, and GMM based phoneme tagging for overlap. The task focuses on first detection of the presence of overlap speech, followed by the modeling and separation of the overlap speech. In this domain, PLDA based speaker processing for overlap speech analysis is considered. A related aspect in the domain of co-channel speech is in massively long real-world audio stream processing. Here, 8-16hr duration audio streams are considered from the Prof-Life-Log corpus. Methods employed here include effective speech activity detection (SAD), including the CRSS-UTDallas Combo-SAD unsupervised solution, as well as threshold optimized variation for addressing long duration silence intervals. Aspects in knowledge extraction from these audio streams is also important, and in this regard the methods employed include word-count estimation based on primary/secondary speaker estimation, front-end speech enhancement, syllable detection using phonological feature based ASR, and MMSE based estimation of word count using phoneme/syllable content.

Task#4: ASR/KWS

For speech recognition and keyword spotting, there is a rich set of methods that have been considered. In the automatic speech recognition (ASR) front-end, feature processing using LDA/MLLT (Linear Discriminant Analysis/Maximum Likelihood Linear Transform) has been used. Deep learning methods for feature processing such as bottleneck features have also been employed and shown to be helpful in reducing errors. For acoustic modeling, progressive decrease in error rates were shown as the system configuration/solution moved from Speaker Adaptive Training (SAT) using fMLLR to Subspace Gaussian Mixture Models (SGMMs). Additional improvement in performance was seen when moving to Maximum Mutual Information (MMI) criterion for training (MMI is a discriminative method for training). Finally, deep learning methods such as convolutional neural networks (CNNs) provide further gains on top of the best MMI systems developed by CRSS-UTDallas. The solution developed by CRSS-UTDallas has been based on phone-confusion networks (PCN) for keyword spotting. The PCN-KWS algorithm attempts to search the keyword in the PCN structure. By allowing flexibility for phoneme deletion and insertion, the algorithm is more robust to ASR errors. Using part-of-speech (POS) tagging, the algorithmic advancements have been demonstrated for keyword spotting when exploring noun terms versus verbs. Also, technical keywords versus common keywords are also considered. Another aspect of ASR explored include distance based ASR, where the subject speaking is at some physical distance from the microphone, but not necessarily speaking to the microphone. Here, the methods explored include nonnegative matrix factorization (NMF), coupled with voice activity detection (VAD) and Sparse Decomposition (SD) by supervised NMF. Again, deep learning in speech recognition with DNNs are also explored, including tandem and bottleneck features. For ASR, a DNN-HMM system with and without any explicit enhancement or dereverberation processing is explored. While distant speech recognition is challenging, the related domain of vocal effort and whisper ASR is a major challenge. Here, vocal tract length normalization (VTLN) is considered as well as vector Taylor series (VTS) for compensation of the whisper style. A denoising autoencoder (DAE) solution is also considered.

Task#5: Speaker State

For the domain of speaker state, and environmental sniffing, several aspects are addressed. These include (i) speaker height estimation, (ii) nonlinear signal distortion estimation and modeling with application to peak

clipping, and (iii) gender ID in high noise/distortion environments. In the speaker height estimation domain, the methods which have been employed include GMM-UBM, as well as features from MFCCs and direct formant estimation from LPC analysis, line spectral pair (LSP) processing, and fusion strategies to leverage direct height estimation from spectral content versus statistical based height estimation using GMM classification. For the domain of nonlinear signal distortion/environmental sniffing, the methods again considered various features for detection and modeling of peak clipping. Various noise content measures including the NIST STNR (speech to noise ratio), WADA, and speech quality based on PESQ are explored. A toolkit is developed entitled "ClipDaT" which allows for automatic analysis of audio corpora for detection and assessment of clipping content in massive audio data sets. In the third domain, gender ID in high noise conditions is addressed. This focus included methods such as i-Vector based gender ID using PLDA, Maximization of Mutual Information (MMI) based projection, GMM-UBM system formulation, and Maximum-A-Posteriori (MAP) adaptation

4 ASSUMPTIONS AND PROCEDURES:

Due to the wide scope of concepts addressed in the five research topics, it is not to detail all assumptions and procedures in this section. As such, many of the specifics are highlighted in Sec. V Results and Discussion section. Here, the primary assumptions dealing with each research topic are highlighted.

Task#1: SID

The assumptions here in SID are that (i) the audio for training the acoustic speaker models are sufficient to cover a speakers acoustic space (i.e., from CRSS-UTDallas previous work on limited train/test data scenarios, 30sec to 3min of train data would be available); (ii) in the mismatch scenarios, knowledge of what mismatch is known and that at least for training speaker models, that pure "neutral" speech data would be available to build the models; (iii) compensation or normalization methods would require a limited amount of development data to be effective; (iv) for overlap speech scenario, that the level of each speaker could be different but that it would remain relatively constant across the audio stream; (v) for language mismatch, that the two separate scenarios of code-switching between two languages versus a full contact-switch to an alternative language other than American English, would be known; (vi) for the speaking and singing scenario, that the content of what is being spoken is the same in both speaking and singing, and that no music is captured on the audio track (i.e., played through open-air headphones so the subject is able to sing to the music while capturing a pure speech spoken/sung audio stream with no music corruption; (v) that for GPU based system development and experiments using Lombard effect in SID, that the specific Lombard flavor is known during the training and test conditions; (vi) for SID evaluation using Apollo audio streams, that knowledge of the noise/environment is provided for training and test conditions.

Task#2: LID

The assumptions here in LID are that (i) the available training data for each in-set language is sufficient to cover the acoustic space without being speaker dependent; (ii) for the MS-AcID toolkit development – that noise, music, and other distortions can be present, but that a minimum of overlap speech be present and that the speakers be consistent with one language spoken per audio stream; (iii) for work on "Secret In the Silence?" sub-task, that available audio corpora for Arabic be consistent for each dialect; (iv) for open-set language rejection, that the number of out-of-set languages be fixed and known, but that not all out-of-set languages would have available data for training rejection models.

Task#3: Co-Channel/Diarization

The assumptions here in co-channel and diarization domain is that (i) only two speakers be present within a given audio stream where potential overlap speech is being detected; (ii) that knowledge of the identity of each speaker is not known prior to overlap speech detection process; (iii) for Prof-Life-Log audio stream analysis for diarization, that only the primary speaker would be analyzed for content, and that secondary speakers would be identified or content extracted as per IRB protocol established.

Task#4: ASR/KWS

The assumptions here in speech recognition and keyword spotting are that (i) a list of potential keywords are known beforehand for at least some of the experiments; (ii) that performance would be documented across a range of audio stream styles including noise/channel/speaker state conditions; (iii) for word count and sentiment analysis – that only the primary speaker would be analyzed; (iv) for distant based ASR, that only a single speaker is talking with no overlap in the speech content; (v) for whisper based ASR, that a single speaker is talking and there is available neutral speech data for the same speaker for model development.

Task#5: Speaker State

The assumptions here in speaker state and environment sniffing are (i) for speaker height estimation, that no background noise or distortion are present; (ii) that the speakers are speaking American English in a neutral context for height estimation; (iii) that accuracy would be documented with a confidence measure associated with the output; (iv) for nonlinear distortion measurement, that in peak clipping only a single distortion is present at any one time (i.e., no multiple types of nonlinear distortion are competing); (v) that for Gender ID, distortion can be present, but that it needs to be consistent across the audio stream (i.e., for DARPA RATS data, and other test/train corpora); (vi) for normalization processing that knowledge of the distortion type would be assumed initially to assess how effective detection and compensation would be, and subsequent processing would explore multiple conditions/scenarios of distortion (i.e., peak clipping is selected as the first domain for detection, modeling, and analysis).

5 RESULTS & DISCUSSION:

In this section, project progress in terms of procedures, results, and discussion are presented for the 36 month period. These represent the core technical accomplishments for the five research tasks.

Task 1 - Speaker ID (SID) Robustness:

This task is partitioned into a set of sub-tasks, all dealing with robustness issues for speaker recognition/identification (SID). Here, highlights of the research advancements over the past 36month period are highlighted. Further details can be found the following publications:

[1] G. Liu, J.H.L. Hansen, "An Investigation into Back-End Advancements for Speaker Recognition in Multi-Session and Noisy Enrollment Scenarios," IEEE Trans. on Audio, Speech & Lang. Proc., vol. 22, no. 12, pp. 1978-1992, Dec. 2014

[2] S.O. Sadjadi, J.H.L. Hansen, "Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch Conditions," IEEE Trans. Audio, Speech and Language Processing, vol. 22, no. 5, pp. 935-943, May. 2014

[3] G. Liu, C. Yu, N. Shokouhi, A. Misra, H. Xing, J.H.L. Hansen, "Utilization of Unlabeled Development Data for Speaker Verification," IEEE SLT-2014: Spoken Language Technology Workshop, paper PT3.10, Lake Tahoe, Dec. 7-10, 2014.

[4] A. Misra, J.H.L. Hansen, "Spoken Language Mismatch in Speaker Verification: An Investigation with NIST-SRE and CRSS Bi-Ling Corpora," IEEE SLT-2014: Spoken Language Technology Workshop, paper PT3.2, Lake Tahoe, Dec. 7-10, 2014

[5] M.K. Nandwana, J.H.L. Hansen, "Analysis and Identification of Human Scream: Implications for Speaker Recognition," ISCA Interspeech-2014, Paper #974, Singapore, Sept. 14-18, 2014.

[6] C. Yu, G. Liu, J.H.L. Hansen, "Acoustic Feature Transformation using Unsupervised LDA for Speaker Recognition," ISCA Interspeech-2014, Paper #1288, Singapore, Sept. 14-18, 2014.

[7] G. Liu, C. Yu, A. Misra, N. Shokouhi, J.H.L. Hansen, "Investigating State-of-the-Art Speaker Verification in the case of Unlabeled Development Data," ISCA Odyssey-2014: Workshop on Speaker & Lang. Recog., Joensuu, Finland, June 16-29, 2014

[8] C. Yu, G. Liu, S.-J. Hahm, J.H.L. Hansen, "Uncertainty Propagation in Front End Factor Analysis For Noise Robust Speaker Recognition," IEEE ICASSP-2014, pp. 4050-4054, Florence, Italy, May 4-9, 2014

[9] H. Xing, J.H.L. Hansen, "Frequency offset correction in single sideband speech for speaker verification ," IEEE ICASSP-2014, IEEE Inter. Conf. Acoustics, Speech, Signal Proc., pp. 4050-4054, Florence, Italy, May 4-9, 2014

[10] T. Hasan, J.H.L. Hansen, "Maximum Likelihood Acoustic Factor Analysis Models for Robust Speaker Verification in Noise," IEEE Trans. Audio, Speech and Language Processing, vol. 22, pp. 381-391, Feb. 2014

[11] T. Hasan, J.H.L. Hansen, "Acoustic Factor Analysis based Universal Background Model for Robust Speaker Verification in Noise," ISCA INTERSPEECH-2013, pp. 3127-3131, France, August 25-29, 2013

[12] K. Godin, S.O. Sadjadi, J.H.L. Hansen, "Impact of Noise Reduction and Spectrum Estimation on Noise Robust Speaker Identification," ISCA INTERSPEECH-2013, pp. 3656-3660, France, August 25-29, 2013

[13] O. Sadjadi, J.H.L. Hansen, "Robust Front-End Processing For Speaker Identification Over Extremely Degraded Communication Channels," IEEE ICASSP-2013, Vancouver, Canada, May 26-31, 2013

[14] T. Hasan, S.O. Sadjadi, G. Liu, N. Shokouhi, H. Bořil, J.H.L. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," IEEE ICASSP-2013, Vancouver, Canada, May 26-31, 2013

[15] T. Hasan, R. Saeidi, J.H.L. Hansen, D. Van Leeuwen, "Duration Mismatch Compensation for i-Vector based Speaker Recognition," IEEE ICASSP-2013, Vancouver, Canada, May 26-31, 2013

[16] G. Liu, T. Hasen, H. Bořil, J.H.L. Hansen, "An Investigation on Back-End for Speaker Recognition in Multi-Session Enrollment," IEEE ICASSP-2013, Vancouver, Canada, May 26-31, 2013

[17] O. Sadjadi, J.H.L. Hansen, "Unsupervised Speech Activity Detection using Voicing Measures and Perceptual Spectral Flux," IEEE Signal Processing Letters, vol. 20, no. 3, pp. 197-200, March 2013

[18] T. Hasan, J.H.L. Hansen, "Acoustic Factor Analysis for Robust Speaker Verification," IEEE Trans. Audio, Speech and Language Processing, vol. 21, no. 4, pp. 842-853, April 2013

[19] J.H.L. Hansen, J.-W. Suh, M. R. Leonard, "Leveraging the Speaker and Noise Space for Effective In-Set/Out-of-Set Speaker Recognition," Speech Communication, vol. 55, pp. 769-781, April 2013.

- **Task 1.1. Multi-Session for SID:** Over the past six months, this specific task project was defined to investigate within- and between-session variability in speaker verification. Using the MARP corpus [Lawson09], this project has examined how characteristics of speech change throughout a recording (plus over time), and how this influences verification performance. The change in speech characteristics *between* recordings, over a 3 year period, has been initially examined in a similar way. While this investigation has been preliminary, the analysis is expected to enable some 'best practice' guidelines for training speaker verification systems to be defined.

  GOAL: The goal of this project is to investigate how speaker variability, within and between recordings, affects speaker identification (SID) performance, and to uncover the sources of this variability at a speaker level. The MARP corpus [Lawson09] is the primary source of data for this project.

  Project Progress: Several aims of the project have been fulfilled to date: the design and implementation of a SID evaluation using the MARP corpus; analysis of the effect of within and between session variability on SID performance; analysis of vocal characteristics across sessions; proposal of a novel calibration scheme to reduce aging-related performance degradation; application of a zoo classification for aging speakers and the proposal of novel 'migration' metrics. The following report outlines this progress in more detail, and summarizes ongoing and future directions of work on this project.

SID Experimental Design: A gender-dependent, all-vs-all, SID protocol was designed, beginning with a subset of 15 male and 15 female (15m/15f) MARP speakers, and then increasing to a set of 35m/25f. The all-vs-all design enabled the analysis of within- and between-session effects, at a global and individual speaker level. Figure 1 is a visualization of the 15m/15f subset of the MARP database. Figure 2 illustrates the distribution of trials (comparisons) the result from the application of the all-vs-all protocol to the 15 males.

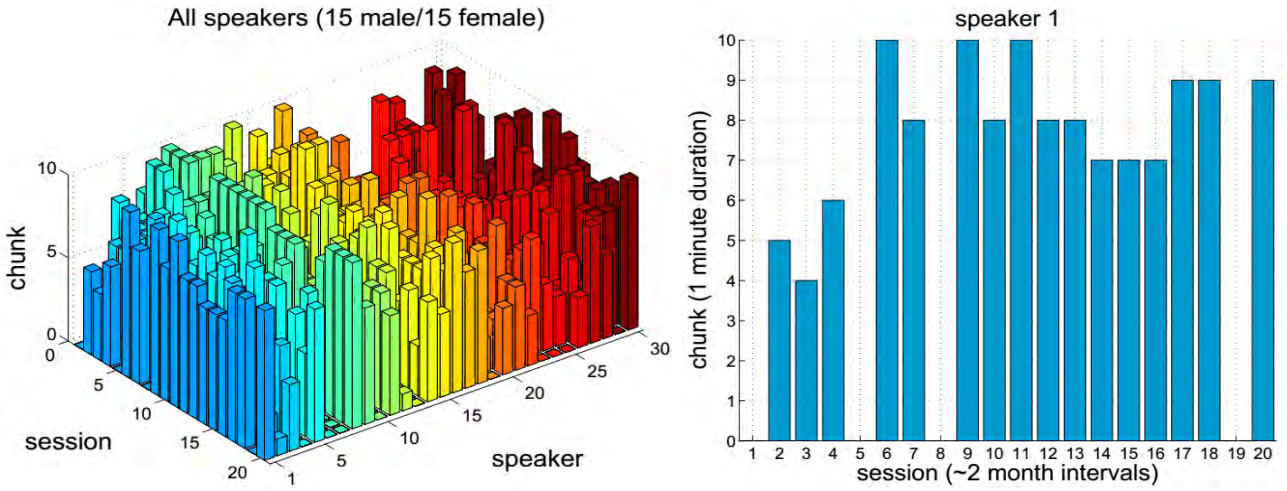


***Figure 1**: **Left:** the audio data present in the MARP corpus for a subset of 15 male and 15 female speakers, where each session is a different recording, and the chunk value is the duration of the recording in minutes. **Right:** a detailed view of speaker 1 from the plot on the left.*

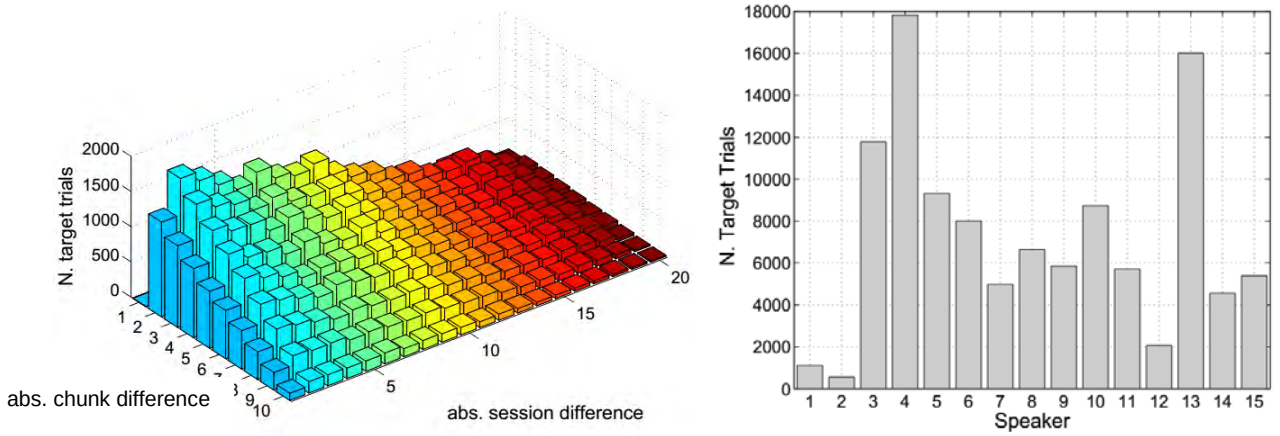


***Figure 2**: The distribution of trials with the all-vs-all protocol, given the recordings of the 15 male speakers shown in Figure 1 (speakers 1-15). **Left:** the number of target trials (genuine trials) according to the absolute difference (in time) between the corresponding sessions, and the absolute difference (in time) between corresponding chunks within sessions. **Right:** shows the total number of target trials for each of the 15 male speakers.*
the audio data present in the MARP corpus for a subset of 15 male and 15 female speakers, where each session is a

SID system: After exploring several baseline SID systems, a suitable approach utilizing an i-vector PLDA system [Sadjadi13a] was adopted. System specification:

- Front-end: 13-dim MFCCs, extracted over 20 ms windows at 10 ms intervals, with first and second derivatives. Mean and variance normalization and RASTA filtering were applied.

- Speech activity detection (SAD) was applied via Combo-SAD [Sadjadi13b].
- Gender-dependent UBMs of 1024 components were trained with a ≈30 hours microphone speech from speakers of US English from NIST SREs 2008 and 2010. A maximum of 30 seconds of speech post-SAD was included from each session.
- i-vector extractor matrix T of rank 400 was estimated with UBM data.
- Linear discriminant analysis (LDA) was applied, reducing the i-vector dimensionality to 200.
- i-vectors were mean and length normalized, and whitened.
- PLDA training used UBM data.

SID Evaluation Overview: Detection error trade-off (DET) curves showing overall system performance for 35m/25f subset, with the duration of all samples fixed at 60 seconds post-SAD, is illustrated in Figure 3. The effect of trial weighting [Leeuwen08], to account for the unbalanced distribution (as in Figure 2) can be observed.

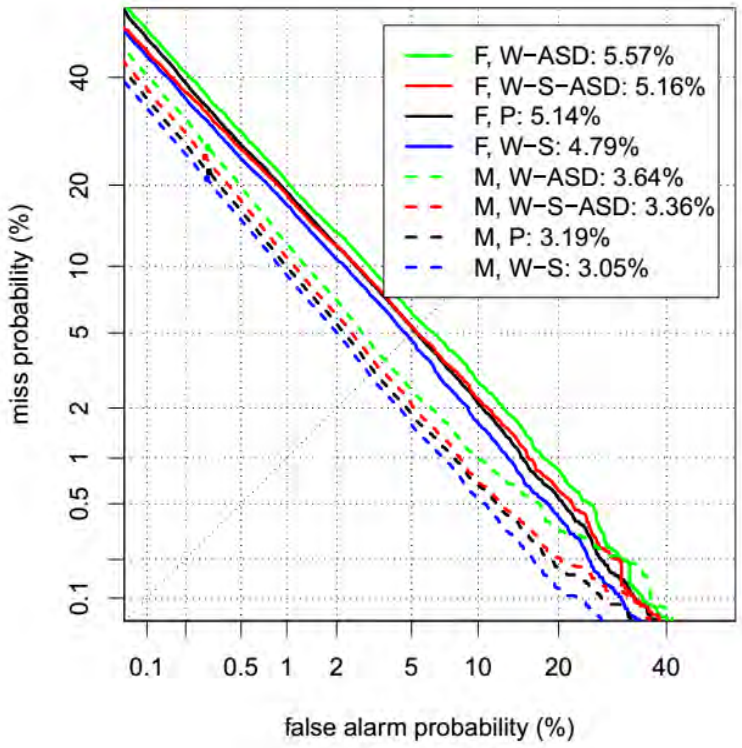


***Figure 3**: DET curves for 'F' (female) and 'M' (male) speakers. Equal error rates (EERs) are indicated in each case.*

- 'P': pooled (unweighted)
- 'W-S': speaker-weighted
- 'W-ASD': ASD-weighted
- 'W-S-ASD': speaker- and ASD-weighted.

An analysis of the output scores from the all-vs-all experimental protocol demonstrated a progressive increase in verification error as the train and test sessions moved further apart in time. In addition, a relationship between verification error and the position of train and test samples *within* their respective sessions was observed. These observations are summarized in Figure 4.

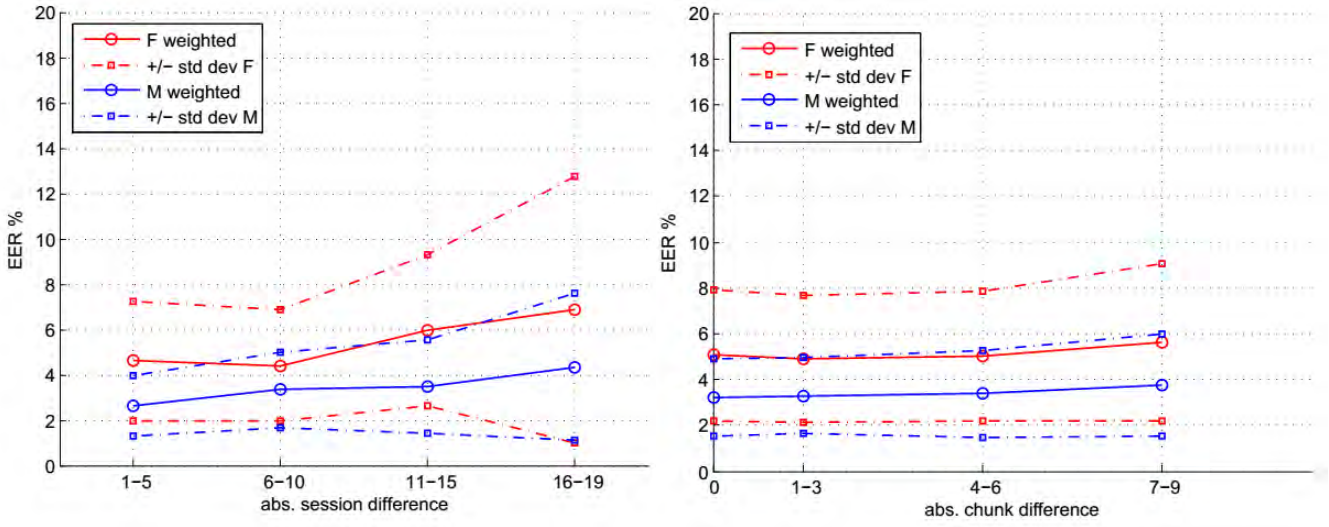


***Figure 4**: EERs (%) for trials grouped according to the difference (in time) between: **Left**: sessions, **Right**: chunks (from different sessions) A chunk is a 1 minute subset of a session.*

Main observations from this Multi-Session SID evaluation:

- The dominant variability in MARP affecting SID is between-session (rather than within-session). The main focus of subsequent work has therefore been on this variability, which is assumed to result from the influence of short-term aging.
- SID performance is gender-dependent: male EER is consistently lower than female EER (Figs. 3 & 4). This is in keeping with results from other evaluations, e.g. NIST SRE 2012 [Saeidi13]. There are also indications that SID performance trends across short-term aging are gender-specific.
- SID performance is speaker-dependent, with speaker variability increasing with absolute session difference, e.g. std. deviations in Fig. 4 (left plot).
- The duration of the train/test samples has a strong influence on the EER. This is in keeping with results from other evaluations, e.g. [Mandasari13]. However, the relative trends in Fig. 4 are consistent across a range of durations (30, 60, 90 and 120 sec. have been evaluated). The most current results are presented here, where there duration is fixed at 60 sec. after SAD, ensuring that the exact same duration of active speech is present in all train/test samples.

SID Score Calibration: Calibration of SID output scores by applying a linear transformation (learned from a development dataset) allows the system to be evaluated in an application-independent way [Leeuwen07]. Conventional score calibration uses only the labels (target/non-target) of the trials to scale and shift the score distribution. This procedure can be extended by considering additional information for every trial. This idea was adopted in [Mandasari13], where the *duration* of the samples corresponding to each trial were incorporated in the score transformation. Based on Figure 1-4 (left plot), there is a relationship between absolute session difference (ASD) and EER. Thus, we proposed to include aging information, via the ASD value for every trial, into the score calibration operation. This was achieved with the inclusion of an extra term in the linear transformation:

$$x = w_0 + w_1 s + Q(w_2, ASD) \quad (1)$$

where w_0 and w_1 are offset and scaling parameters respectively. Q denotes a Quality Measure Function (QMF) defining the way in which ASD (and any additional parameters) are incorporated into the calibration, and w_2 is a new calibration parameter to be optimized on the development set. The performance of different QMFs were evaluated on the scores of male speakers. Performance was evaluated via the log-likelihood ratio cost, C_{llr} [Leeuwen07], which provides a measure of discrimination and calibration over all effective priors, and the relative miscalibration, R_{mc}, defined as:

$$R_{mc} = (C_{llr} - C_{llr}^{min}) / (C_{llr}^{min}) \quad (2)$$

where C_{llr}^{min} is the minimum value of C_{llr} obtained via an optimal score transformation. Results based on the self-calibration of male scores are shown in Table 1.

	ASD range	1-3	4-7	8-11	12-15	16-19
MM	C_{llr}	.094	.121	.150	.153	.226
	R_{mc}%	2.17	7.08	13.64	15.04	34.54
F	C_{llr}	.099	.116	.136	.139	.195
	R_{mc}%	7.61	2.65	3.03	4.50	16.07
M	C_{llr}	.094	.115	.134	.137	.184
	R_{mc}%	2.17	1.77	1.52	3.01	9.53
Q_1	C_{llr}	.095	.115	.135	.137	.185
	R_{mc}%	3.26	1.77	2.27	2.24	10.12
Q_2	C_{llr}	.095	.115	.135	.137	.187
	R_{mc}%	3.26	1.77	2.27	2.24	11.31
Q_3	C_{llr}	.095	.115	.135	.137	.186
	R_{mc}%	3.26	1.77	2.27	2.24	10.71

Table 1: *Performance of score-only calibration approaches: Full (F), Mismatched (MM), and Matched (M) and proposed score-aging calibration approaches Q1-Q3 (each with a different QMF). Linear, log and exponential functions are represented by Q1-Q3 respectively. The score-only calibration 'M' represents an ideal, but unrealistic scenario. Performance comparisons of interest are therefore between 'F', 'M', and each of Q1-Q3.*

In summary, Table 1 indicates that including aging information into calibration via a QMF reduces discrimination and calibration error compared with using the score alone, and the relative improvement increases with ASD.

Analysis of Vocal Features: In addition to speaker ID performance analysis, it was also of interest to understand the changes in speech production changes across the sessions. Therefore, for each of the 35m/25f speakers considered in the SID experiment, a set of 'vocal features' were extracted, including: F0, local jitter, local shimmer, speech rate and HNR (harmonic-to-noise ratio). Some initial findings of interest:

- Speech rate, for almost all speakers increases significantly from session 1, stabilizing around session 6. This is potentially due to conversations between partners becoming more engaged as the sessions progress.
- Female HNR is generally higher than male HNR, however females with lower HNR have generally more stable EERs across sessions.

- Speakers with higher local jitter variability tend to have higher EER variability across sessions. For several speakers there is a correlation between local jitter increase and EER degradation across sessions.
- Male local shimmer is general higher than female local shimmer. Males with high vocal shimmer have greater EER variability across sessions.
- There is a small, but progressive decrease in pitch for some female speakers across session. However, there is no noticeable correlation with EER.

Speaker-level Performance Assessment: Speaker variability is speaker-dependent (observed in the vocal feature measurements, for example), as too is its effect on SID (see standard deviation of scores, Figure 4). Thus, a drawback of using global evaluation metrics, like EER or Cllr, is that the behavior of individual speakers, or subsets of speakers, may be masked.

A means of analyzing speaker recognition performance at a speaker-level was put forward in [Doddington98]. The so-called 'Doddington Zoo' is a categorization of individual speakers as different animals based on the statistics of their recognition scores. Speakers that are inherently difficult to classify are assigned a different category than typical speakers, whose scores are generally well-behaved. This idea was extended in [Yager10] with the introduction of new animals (or categorizations) taking the scores of non-targets into account, and again in [Alexander14] by taking the variance of scores into account.

In this current project, a zoo classification based on [Alexander14] has been applied to the speakers in MARP. Firstly, this provides a visual representation of the relative recognition tendencies of different speakers for a given recording session. More importantly, by tracking the movement of speakers within the zoo classification space across short term aging, the speakers most and least affected by aging can be identified. Measurements of 'migration' in the zoo space, based on various distance metrics (Euclidean, Chebyshev, Cityblock etc.) between speaker points have been used to rank speakers. A correlation analysis between migration measures and vocal features is ongoing. An example of zoo classification for 5 female MARP speakers across short-term aging is provided in Figure 1-5.

Multi-Session SID Ongoing/future work:

- Vocal feature correlation analysis: find relationships between speaker zoo migration metrics and vocal features. Use findings to augment front-end features and/or incorporate new information in score calibration.
- Within-session analysis: work to date has focused mainly on between-session aging analysis. From Fig.4 (right), and other experiments, the effect of position within a session has been observed to affect SID. Will track the change in vocal features within a conversation as part of this analysis. New higher-level and lexical features will be considered here, including word count and measures of engagement/communication.
- Between-session analysis: will evaluate the effect of incorporating longitudinal speaker data in PLDA training. Long-term longitudinal data from TCDSA [Kelly13] will be used for this purpose. We will also explore lightly supervised data harvesting from YouTube, where audio from bloggers, or other users with regular uploads across time, will be identified, downloaded and then processed automatically.
- Will consider conducting a listener test, in house and/or over Mechanical Turk, which could address how human speaker recognition performance is affected by:
    - Time elapsed between recording of samples
    - The gender of the speaker and of the listener

- The age of the speaker and of the listener
- The duration of the samples
- Text-dependence vs text-independence
- 'In-house' vs Mechanical Turk
- Additive noise of different types and SNRs
- The native language of listener

- Work on this project will be submitted for publication in top-tier journals.
  - Potential JASA (Journal of the Acoustical Society of America) paper: a presentation of characteristics of short-term vocal aging and the effect on human and machine speaker recognition
  - Potential IEEE Transactions on Audio, Speech and Language Processing or Speech Communication paper: Aging calibration and aging PLDA modeling for SID

Publications from Task 1.1: Multi-Session SID

- F. Kelly and J.H.L. Hansen, "The effect of short-term vocal aging on automatic speaker recognition performance," *International association of forensic phonetics and acoustics (IAFPA) conference*, Leiden, The Netherlands, July 2015 **[accepted]**
- F. Kelly and J.H.L. Hansen, "An overview of speaker variability as a source of error in forensic automatic speaker recognition," *NIST International Symposium on Forensic Science Error Management - Detection, Measurement, and Mitigation,* Washington,DC, July 2015 **[accepted]**
- F. Kelly and J.H.L. Hansen, "Evaluation and calibration of short-term aging effects in speaker verification," *ISCA INTERSPEECH-2015*, Dresden, Germany, September 2015 **[submitted]**

Multi-Session SID References: (cited in this section)

[Alexander14] A. Alexander et al. G. Doddington et al., "Zooplots for Speaker Recognition with Tall and Fat Animals." *IAFPA,* 2014

[Doddington98] G. Doddington et al., "SHEEP, GOATS, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," *ICSLP* 1998

[Godin10] K. W. Godin and J.H.L. Hansen, "Session variability contrasts in the MARP corpus," *ISCA INTERSPEECH-2010*, Makuhari, Japan

[Lawson09a] A. D. Lawson et al., " The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design and Initial Findings," *ISCA INTERSPEECH-2009,* Brighton, U.K.

[Lawson09b] A. D. Lawson et al. (2009). Long term examination of intra-session and inter-session speaker variability. *ISCA INTERSPEECH-2009*, Brighton, United Kingdom

[Leeuwen07] D. A. v. Leeuwen and N. Brummer. "An Introduction to Application-Independent Evaluation of Speaker Recognition Systems*," Speaker Classification, I,* 330-353, 2007.

[Leeuwen08] D. A. v. Leeuwen. "A note on performance metrics for speaker recognition using multiple conditions in an evaluation*," Technical report*, June 2008

[Kelly13] F. Kelly, A. Drygajlo, and N. Harte. "Speaker verification in score-ageing-quality classification space," *Computer Speech & Language*, 27:1068-1084, 2013

[Kinnunen10] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication* 52(1):12-40, 2010

[Mandasari13] M. Mandasari, R. Saeidi, M. McLaren, and D. van Leeuwen, "Quality Measure Functions for Calibration of Speaker Recognition System in Various Duration Conditions," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. PP, pp. 1-1, 2013.

[Sadjadi13a] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research," November 2013.

[Sadjadi13b] S.O.Sadjadi, J.H.L.Hansen, "Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux," *IEEE Signal Processing Letters,* vol. 20, pp. 197-200, 2013.

[Saeidi13] R. Saeidi et al., " I4U Submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification," *InterSpeech,* Lyon, 2013.

[Yager10] N. Yager and T. Dunstone, "The Biometric Menagerie.," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32(2):220-30, 2010

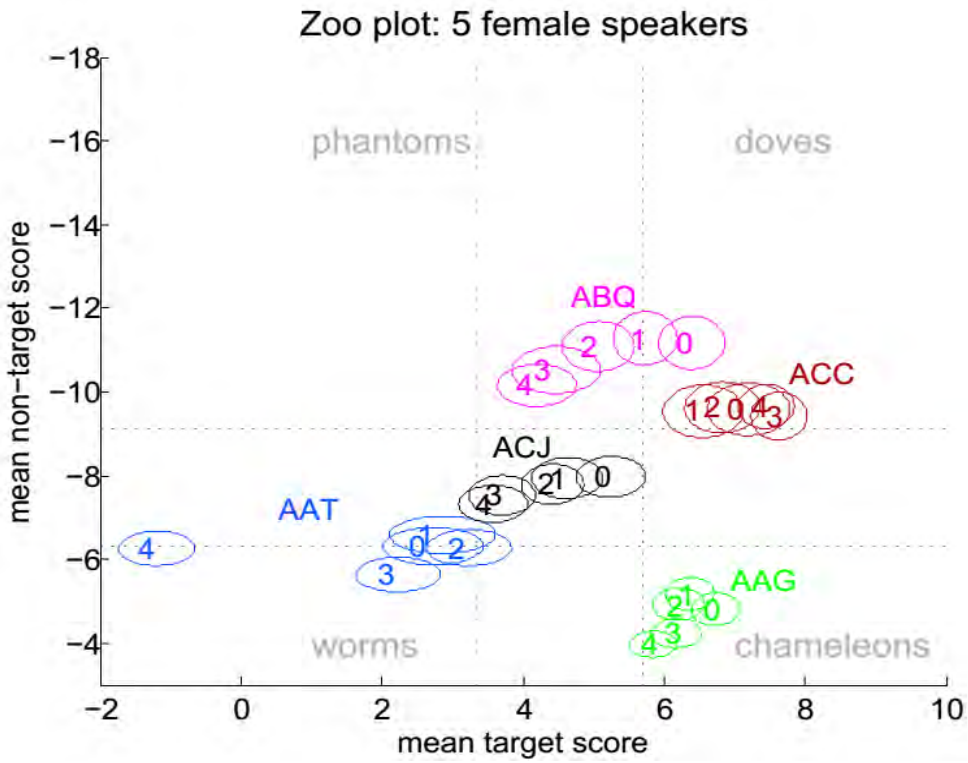


Figure 5: *Zoo plot for 5 female speakers (each identified by a different color and three-letter label) at 5 different absolute session difference (ASD) ranges: 1-3, 4-7, 8-11, 12-15, 16-19 (denoted by the labels 0-4 respectively). The center of each ellipse indicates the mean of the target and non-target scores for a particular speaker and ASD. The shape of the ellipse indicates the (scaled) standard deviation of these scores. The boundaries between classifications are determined from the scores of all females at ASD range 1-3. The absolute classification of speakers into categories is not of great importance; generally however, speakers in the top right (doves) are best performers, and speakers in the bottom left (worms) are the worst. An example of what can be interpreted from this plot: speaker ACC shows good recognition performance (dove classification). Performance is stable across all ASDs, in terms of mean and standard deviation. Speaker AAT shows poor recognition performance (worm classification). Performance is not stable across all ASDs (large negative movement between ASD 3 and 4), and the standard deviation of target scores is also variable.*

Task 1.2. Language Mismatch Compensation in SID:

Language mismatch has been one of the less explored areas of speaker verification, yet is a common challenge in speaker ID since many individuals are bi-lingual and may switch or code-switch between primary and secondary languages during conversations. Most prior work done to address acoustic mismatch between train and test conditions has been mainly focused on channel mismatch, believing that channel mismatch compensation techniques will also suppress language mismatch [1,2,3,4]. However, in our study, we observed that even in presence of state-of-the-art channel mismatch compensation techniques, language mismatch persists and degrades the performance of a speaker verification system. To conduct our study, we extracted bilingual speakers from NIST SRE -04-08 corpus and built a state-of-the-art i-vector-PLDA based speaker verification system. Using this system, we showed that language mismatch alone, causes a severe degradation, dropping the performance metric by a factor of 2.5. To improve performance, we considered adding small amounts of multi-lingual data to the PLDA formulation and observed a relative improvement of +62.18% [5]. Table 1-2 shows the results of our initial experiments.

Table 2: *Results in the language matched and mismatched conditions*

Condition	EER (%)
Matched	1.745
Mismatched	4.395
Mismatched (with multi-lingual PLDA)	1.662

The above proposed solution, however, requires significant amount of multi-lingual data to improve performance. In reality, such amount of data might not be easily available for low resource languages. Previous studies on the subject also mostly rely on the availability of multi-lingual data. In [6], speaker models trained on both train and test languages were used. In [7], the authors estimated a language dependent subspace using a Joint Factor Analysis (JFA) framework and later suppress it as a nuisance attribute. To compute the language subspace Oregon Graduate Institute (OGI) 22 language multi-lingual data-set was used.

To better understand the problem and come up with an algorithm which would utilize significantly less development data and could equally work for all the languages, we went forward and collected a small database, containing language as the only source of mismatch. Table 3 shows the statistics of our collected database (CRSS Bi-Ling SID).

Table 3: *CRSS Bi-Ling SID corpus statistics*

Language	Gender	Number of speakers
Mandarin	Male	13
Mandarin	Female	11
Hindi	Male	26
Hindi	Female	20
Total		**70**

Different languages have different phoneme structure, grammar and vocabulary. In [8], the authors observed that phonemes with the same amount of training data gave different errors, leading to the conclusion that the system uses discrimination based on phonemes for speaker recognition. Also, in [9] it was observed that in broad groupings nasals and vowels provide the best speaker recognition performance followed by fricatives, affricates and approximants, with stops providing the worst performance of all. Motivated by these studies, we considered reducing the phonetic language mismatch between train and test utterances through phonetic tagging which we have used in the past for speaker ID [5]. A GMM-UBM system was set up using English train utterances and Hindi test utterances of 2.5 minutes each from the CRSS Bi-Ling SID corpus. Data from Mandarin utterances was used to train a UBM. Next, a Hungarian phoneme recognizer was used to transcribe Hindi and English phonemes for two reasons: i) it contains the highest number of unique phonemes (61), and ii) it will not be biased towards Hindi or English speech. Next, phoneme histograms were plotted for each of the English and Hindi utterances. Fig 1-6 shows one such set of histograms for a single 2.5 minutes English and Hindi utterance. A novel mismatch compensation method was proposed called **Phoneme Histogram Normalization (PHN)** to reduce the difference between the phoneme language distribution of train and test utterances. In PHN, the test utterance phoneme histogram is normalized to match it with the train utterance phoneme histogram. This is accomplished by dynamically weighing each phoneme at the scoring stage [5].

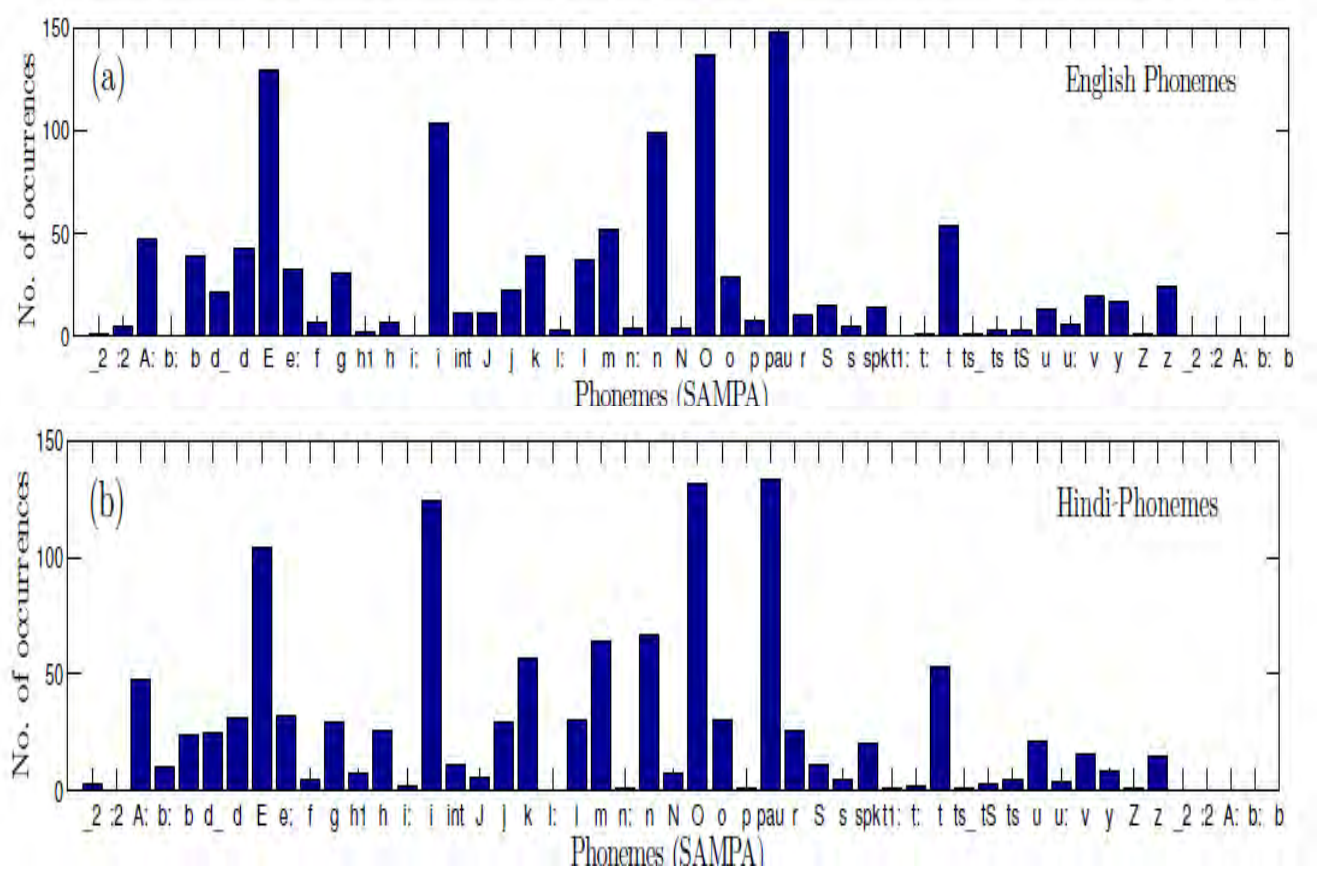


Figure 6: *Phoneme histograms for an English and Hindi spoken utterance.*

Using PHN, a relative improvement of **+16.6%** in speaker verification performance was observed on CRSS Bi-Ling SID database. This is encouraging, since language mismatch is an extremely challenging issue in SID with limited progress to date.

Another method to help address language mismatch in SID was formulated based on a bi-directional transform. Figure 7 shows the two transformational parameters T_{ET} (mapping enrollment language to test language) and T_{TE} (mapping test language to enrollment language).

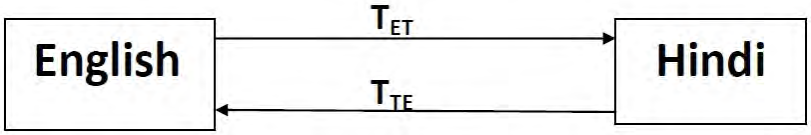


Figure 7: *Point set registration normalization for language mismatch in SID*

In essence, language mismatch is addressed by transforming different phonetic spaces of train and test utterances, thereby removing the need for phonetic tagging. The proposed method was entitled **Probabilistic Bi-Directional Feature Transformation (PBFT)** for the train-to-test features. The transformed features are aligned closer to each other, thereby suppressing language mismatch while simultaneously emphasizing speaker traits. A small amount of development data is used to learn two sets of transformation parameters that are later applied to train and test features. The basis of computing such parameters is to accomplish "rotation, translation and scaling" [10] of the features so as to align them closer to each other in the multidimensional feature space. Fig 8 explains the process of computing the parameters.

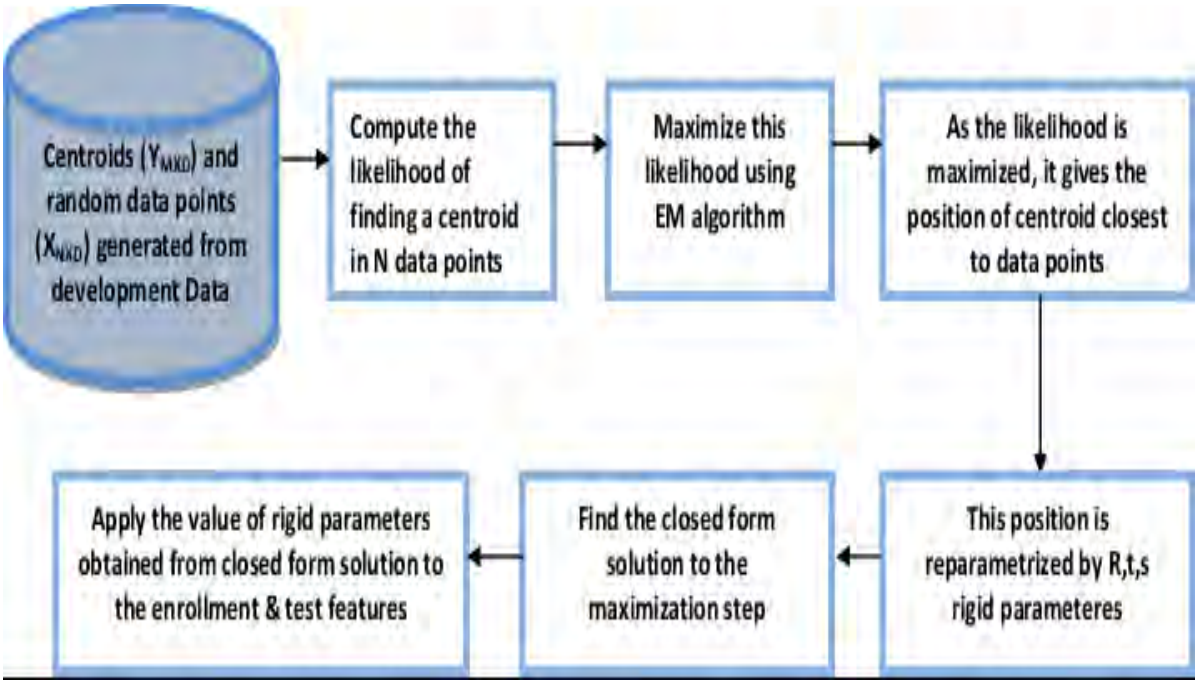


Figure 8: *Flow diagram of the proposed SID Language Mismatch algorithm*

With this approach, a similar GMM-UBM system to that used for PHN, was employed to evaluate PBFT. A relative **+12.7%** improvement in performance was observed using the proposed transformation. The methods here have therefore made effective steps towards helping to minimize the impact the of full language change between train and test spaces.

Specific Objectives/Next Steps: Language Mismatch Compensation in SID: Having shown the benefits of the proposed technique (PBFT), future work would consider leveraging the same concept in an i-Vector-PLDA based speaker verification system. Furthermore, it can be observed that though this technique has been developed primarily for language mismatch compensation, it is general purpose and can also be applied in case of channel or other types of mismatches in speaker verification.

Cross-Language SID Publications:

[1] A. Misra and J.H.L. Hansen, "Spoken language mismatch in speaker verification: An investigation with NIST-SRE and CRSS Bi-Ling corpus", in Spoken Language Technology Workshop, 2014. SLT 2014, Dec 2014.

[2] G. Liu, Che. Yu, N. Shokouhi, A. Misra, H. Xing, J.H. L. Hansen, "Utilization of Unlabeled Development Data for Speaker Verification," IEEE Spoken Language Technology Workshop, 2014. SLT 2014, Dec 2014.

[3] G. Liu, C. Yu, A. Misra, N. Shokouhi, J.H.L. Hansen, "Investigating State-of-the-Art Speaker Verification in the Case of Unlabeled Development Data," Proc. ISCA Odyssey Speaker and Language Recognition workshop, Joensuu, Finland, 2014.

Cross-Language SID References: (cited in this section)

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio Speech Lang. Process., May 2010

[2] D. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," EUROSPEECH-1997 Conference, Rhodes, Greece.

[3] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus Eigenchannels in speaker recognition," IEEE Trans. Audio Speech Lang. Process., May 2007.
[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," IEEE Trans. Audio Speech Lang. Process., July 2008
[5] A. Misra and J.H. Hansen, "Spoken language mismatch in speaker verification: An investigation with NIST-SRE and CRSS Bi-Ling corpus", in Spoken Language Technology Workshop, 2014. SLT 2014, Dec 2014.
[6] B. Ma and H. Meng, "English-Chinese bilingual text-independent speaker verification," in Proc. IEEE ICASSP, May 2004
[7] L. Lu, Y. Dong, X. Zhao, J. Liu, and H. Wang, "The effect of language factors for robust speaker recognition," in Proc. IEEE ICASSP, April 2009.
[8] L. Lu, Y. Dong, X. Zhao, J. Liu, and H. Wang, "The effect of language factors for robust speaker recognition," in Proc. IEEE ICASSP, April 2009
[9] R. Auckenthaler, E.S. Parris, and M.J. Carey, "Improving a gmm speaker verification system by phonetic weighting," IEEE ICASSP, 1999.
[10] A Myronenko and X. Song, "Point set registration:Coherent point drift," IEEE Transactions on Pattern Analysis and Machine Intelligence, Dec 2010.

Task 1.3. Singing vs. Speaking (applications for SID and LID):

During the 36 month project, a significant effort was undertaken to explore the impact of speaking versus singing for speaker ID and language ID. In general, acoustic analysis of singing has shown considerable deviation in spectral and prosodic characteristics of professional singing from speaking, especially for vowels. This study considered an acoustic analysis of untrained karaoke singing and its deviation from speaking for the Hindi language (two additional languages – Farsi and Mandarin were also collected). Spoken and sung vowels were analyzed in context and compared across speakers. The same speakers and same text were used for comparative analysis of speaking and singing. First, the dimensionality of vowel space is compared for speaking and singing based on subspace learning in perceptual linear prediction spectral feature space. It was shown that spoken vowels are more separable than singing vowels, and a higher number of dimensions is required for singing vowel identification compared to speaking. This result is specifically of value for acoustic modeling applications. Next, formant spaces of spoken and sung vowels are compared in terms of vowel space size, coefficient of variation, and vowel identification based on formant frequency features. Finally, phoneme duration and static and dynamic characteristics of fundamental frequency are studied. Fluctuations of fundamental frequency are analyzed for singing vowels, and speaker dependent differences for these fluctuations are considered.

1.3.1: UT-Sing Corpus Development for Speaker ID / Language ID: UT-Sing is a multilingual singing database that was collected for the purpose of singing speech analysis, and studying the effects of singing on various speech systems (Mehrabani and Hansen (2011, 2012, 2013a)). UT-Sing includes more than 23 hours of singing from 81 speakers and four languages: English, Farsi, Hindi, and Mandarin. Native speakers of each language selected 5 popular songs in their language. Each song was approximately 3-5 minutes in duration. Though a list of suggested songs was made available, each subject was allowed to select their songs even if it was not on the list, so they would be familiar with the songs they were singing, and therefore be able to sing more comfortably. The speakers' voice was recorded in a sound-booth using a close-talk microphone while singing as well as reading the lyrics of the same songs. Karaoke system prompts were used for singing data collection. While subjects were listening to the music through headphones, the lyrics were displayed, and only the subject's singing voice was recorded (i.e., no music was captured within the audio stream). UT-Sing subjects were not professional singers, and had a variety of singing skills.

1.3.2: Vowel Space Spectral Analysis: Analysis of singing vowel separation and vowel space dimensionality compared to neutral speaking is important for improving singing acoustic modeling applications such as speech recognition, speaker identification, and language identification for singing.

Spectral features used for this study were 12-dimensional Perceptual Linear Prediction (PLP) feature. PLP feature vectors were classified using a k-Nearest Neighbor (k-NN) classifier. For each speaker, three songs were used for training, and two songs for open test. There was no overlap between train and test songs. Fig. 9 shows frame level vowel classification accuracies for spoken and sung vowels when increasing k from 1 to 20.

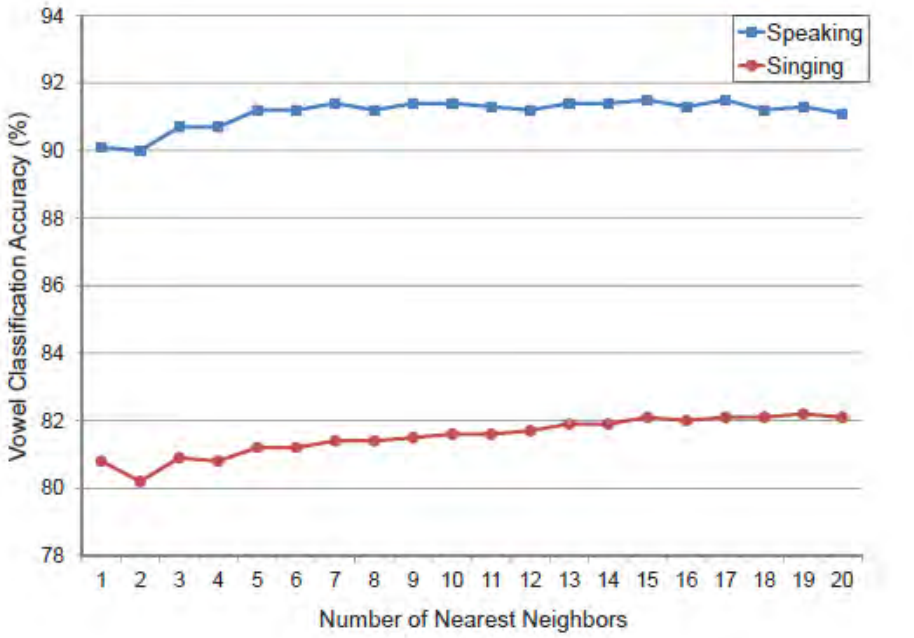


Figure 9: *Vowel classification accuracy for spoken & sung vowels using k-NN classifier (increasing k from 1 to 20).*

Classification accuracy for spoken vowels in Fig. 9 is on the average 9.6% more than singing vowel classification, which can be explained as speaking vowels being more separable compared to singing with 12-dimensional PLP features. For the remaining of this study, the value of k was fixed at k = 15 for the k-NN classifier, since it produced the best accuracies for both speaking and singing vowel classification. Table II shows vowel classification accuracy when train and test data are either speaking or singing, and when train data is speaking and test is singing and vice versa. It is shown that when the k-NN vowel classification system is trained with singing vowels, classification accuracy decreases only by 3% when tested with speaking compared to singing vowels. However, when the system is trained with speaking vowels, there is 22.8% classification accuracy loss when tested with singing compared to speaking. This can be interpreted as more vowel production variability for singing compared to speaking. Next, we analyze dimensionality of the vowel spaces for singing and speaking, and will show that the minimum production space dimensionality is greater for singing versus speaking.

A dimensionality analysis was also performed to uncover the intrinsic speech production dimensional space between speaking and singing. An analysis was conducted which compared the k-NN vowel classification performance for singing and speaking in LPP subspaces of PLP feature vectors. Here, LPP refers to Locality Preserving projections (He and Niyogi (2003)). Fig. 10 shows classification accuracy for spoken and sung vowels with LPP dimension sequentially reduced by one dimensional steps: l = 12, 11, ..., 2, 1. The average singing vowel classification accuracy for all dimensions is 11% lower than speaking. From Fig. 10, from dimension 2 to 5 for speaking, singing requires at least one additional dimension to produce similar vowel classification results to speaking.

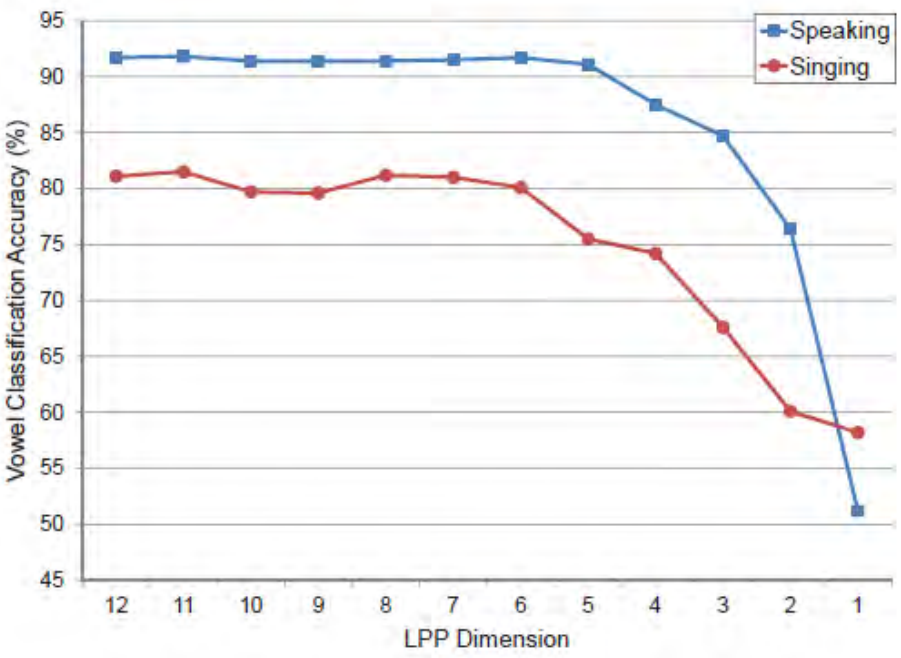


Figure 10: *Vowel classification accuracy for spoken and sung vowels when reducing LPP dimension from 12 to 1.*

Fig. 11 shows the mean F1 and F2 space of five Hindi vowels for speaking and singing. The formant frequency for each vowel is averaged over all utterances spoken/sung by three male speakers. As shown, the size of the acoustic vowel space is altered and reduced from speaking to singing. This can be interpreted as having less separable vowels for singing compared to speaking, which explains the vowel classification results.

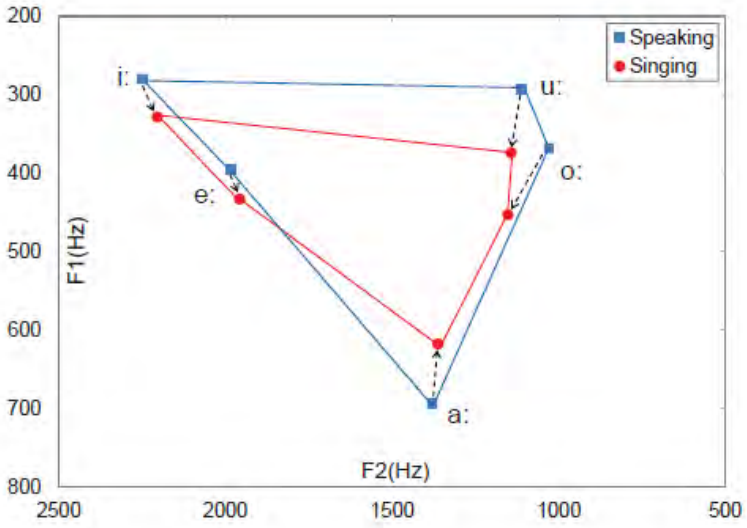


Figure 11: *F2/F1 vowel configuration change from speaking to singing for male Hindi speakers.*

1.3.2: Speech Production Analysis for Speaking vs. Singing: English, Hindi, Farsi

Section 1.2.1 highlighted initial work in the area of singing vs. speaking speech production analysis for Hindi subjects. Here, the study is expanded to consider speakers under English and Farsi as well.

In the area of Speaker ID, this sub-task focused on addressing the mismatch due to speaking vs. singing. In general, speech production changes from speaking to singing due to articulatory modifications that happen while singing. This difference is mostly expressed in the vowel production. Singing also introduces acoustic differences in the time-frequency structure of a speaker's voice, which causes a reduction in performance of speech recognition and speaker identification algorithms designed for neutral speech. In order to better understand the complicated biomechanics of the singing voice and in order to be able to model it better, the focus was placed on only the differences that happen when the speaking style moves from speaking to singing. In this task, we studied this change by analyzing the prosodic differences such as phoneme class duration, mean fundamental frequency (pitch) and the formant frequency vowel space of untrained karaoke singers of English, Hindi and Farsi. We also looked at the Kullback-Leibler distance between the Gaussian probability distribution models trained in singing and speaking and finally used the same type of models in a GMM-UBM speaker identification system. Lastly, we performed vowel classification based on different formant frequency feature dimensions. We obtained a thorough analysis using a multilingual database and explored the cross-language dependencies in the changes that happen during speaking and singing in the three languages.

For the database, we used speech data selected from the UT-Sing corpus that contains 81 speakers from four different languages: English, Farsi, Hindi and Mandarin. We have chosen 12 speakers for the English and the Farsi language, with 6 speakers from each language. Three out of the 6 speakers for each language are male, and 3 are female speakers. Each speaker sings about 5 songs and reads the lyrics of the same songs. There are 16 songs total, with some overlap between the songs. At this time, we started with the phonetic transcription of the Farsi and the English song files. We also classified different phoneme classes for the Farsi language and translated them into the IPA and Arpabet system for easier transcription. During the time that the transcription was taking place, we were trying to develop a baseline of different scripts that will allow us to conduct the different analysis.

During this most recent effort, we phonetically transcribed 10 songs which were read/spoken and produced in singing mode for an English male subject, 8 sung and spoken songs by an English female subject, and 4 sung and spoken songs by another English male subject. We have developed Matlab and Perl routines for analyzing speaking and singing vowels of English speaking subjects. The types of analysis conducted pertain to the production space, such as phoneme duration and spectral acoustic analysis, such as formant analysis. Figure 12(a,b,c) shows the changes in the percentage of the word duration from each phonetic group from speaking to singing for English, Hindi and Farsi. As we can see from the figure, vowels and diphthongs occupy the largest percentage of the word duration in both English and Hindi for both speaking and singing. However, in Farsi the largest percentage in word duration during speaking is occupied by affricates and fricatives. This result for Farsi changes as the speaking style is modified to singing, with vowels leading on the largest percentage of word duration and fricatives with the second largest. The lead of fricatives and affricates in the overall percentage of word duration during speaking in Farsi could be due to the amount affricates and fricatives that are found in the Farsi language compared to vowels. However, during singing, vowels still occupy the largest percentage of word duration.

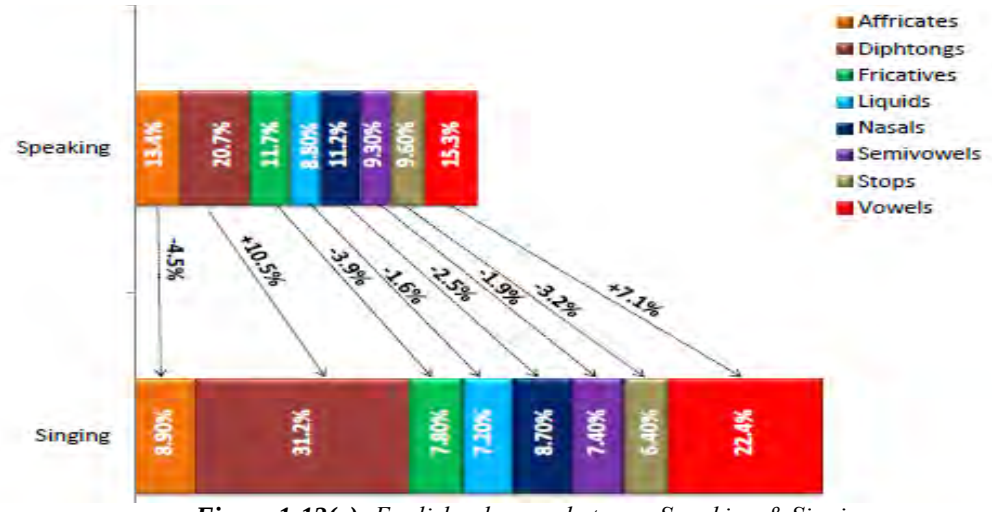


Figure 1-12(a): *English: changes between Speaking & Singing*

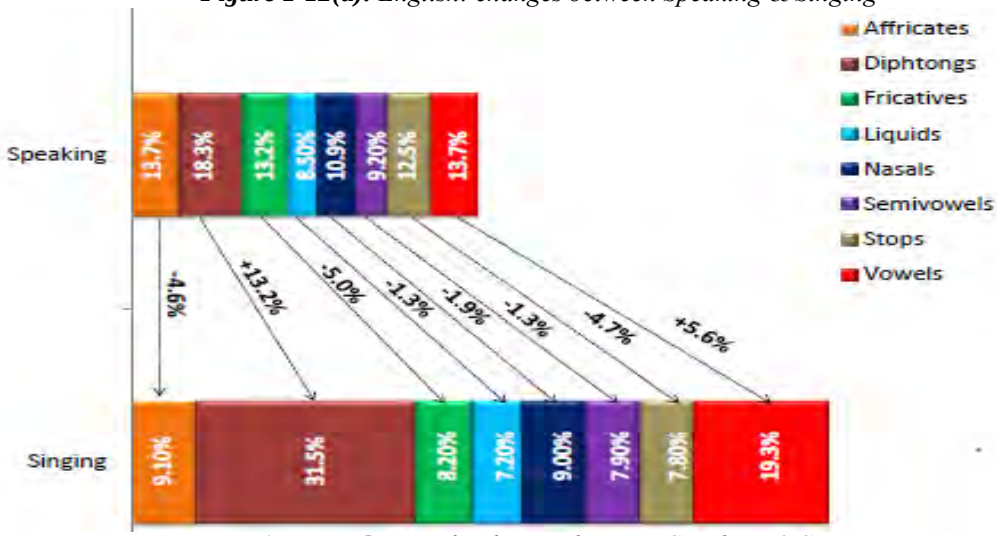


Figure 12(b): *Hindi: changes between Speaking & Singing*

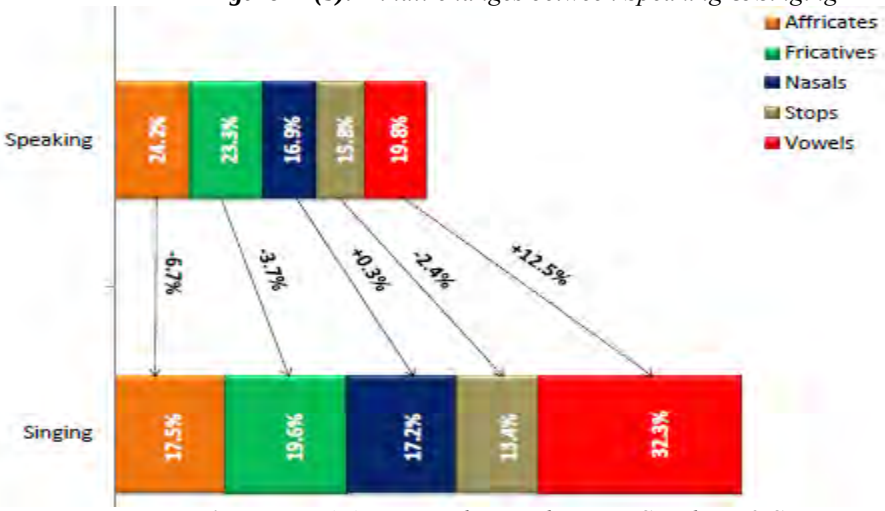


Figure 1-12(c:) *Farsi : changes between Speaking & Singing*

Figure 12. *Percentage of each phonetic class in word duration for singing and speaking for (a) English, (b) Hindi and (c) Farsi subjects*

In spring 2014, we also conducted experiements on changes in the corresponding acoustic space by calculating the average first and second formants of different vowels. In Figure 13, the formant spaces are plotted for vowels ‘a:’, ‘e:’, and ‘u:’ for singing and speaking for English, Hindi and Farsi speakers. The results shown in Figure 13 confirm that all 8 speakers (excluding the Farsi male speaker) have smaller/contracted vowel spaces during singing than while speaking. Even though the formant space ratio of sing/speak is less than 1 in almost all cases for the speakers in this figure, dependencies between speakers still exist. The first two English female speakers (Speaker 1 and Speaker 2) have lower sing/speak formant space ratio than the English male speaker (Speaker 3). Except for one English male speaker, the rest of the male speakers show significantly higher sing/speak ratios than the English female speakers, meaning that the formant space area of English male speakers is larger than in speaking. This result introduces a gender dependency of vowel spaces in the formant frequency plane, which is perfectly reasonable considering the difference in the formant frequencies between the two genders in speaking. Moreover, this dependency carries on when we consider the male and female speakers of Hindi and Farsi, as well.

During the fall 2014, we observed the change in mean fundamental frequency that happens from speaking to singing, for vowels: /a:/, /e:/, /i:/, /o:/, /u:/. Besides the mean fundamental frequency, we also calculated the coefficient of variation (CV) across all 6 subjects in each language separately. The CV was included in order to show the dispersion of our variables. In Figure 14, speaking vs. singing results are plotted for the five English, Hindi and Farsi vowels. Our results show that the mean F0 for spoken vowels is lower than that of the sung vowels for all three languages. The coefficient of variation for spoken vowels (excluding English vowel /i:/) is also lower than the sung vowels.

Also during the fall 2014, vowel classification tasks were also performed with formant frequency features. Results are shown in Figure 15. Generally, it was observed that for homogeneous train/test styles, the K-NN classifier gives better accuracy for the Hindi vowel classification compared to English and Farsi in the higher dimensional formant feature spaces (F1F2, F1F2F3, F1F2F3F4). When the train/test style is sing/read, the English vowel classification accuracy is higher in the higher dimensional formant feature space. In this case, Farsi is second highest with slightly lower classification accuracy. When the train/test style is read/sing, the Farsi vowel classification accuracy is the highest. According to our results, vowel classification performance of English speakers is generally higher than Hindi and Farsi speakers in the 1-D formant feature space for formants F1, F3 and F4. However, the Hindi vowels are more distinguishable than the English vowels in the 1-D space when the features contain formant F2.

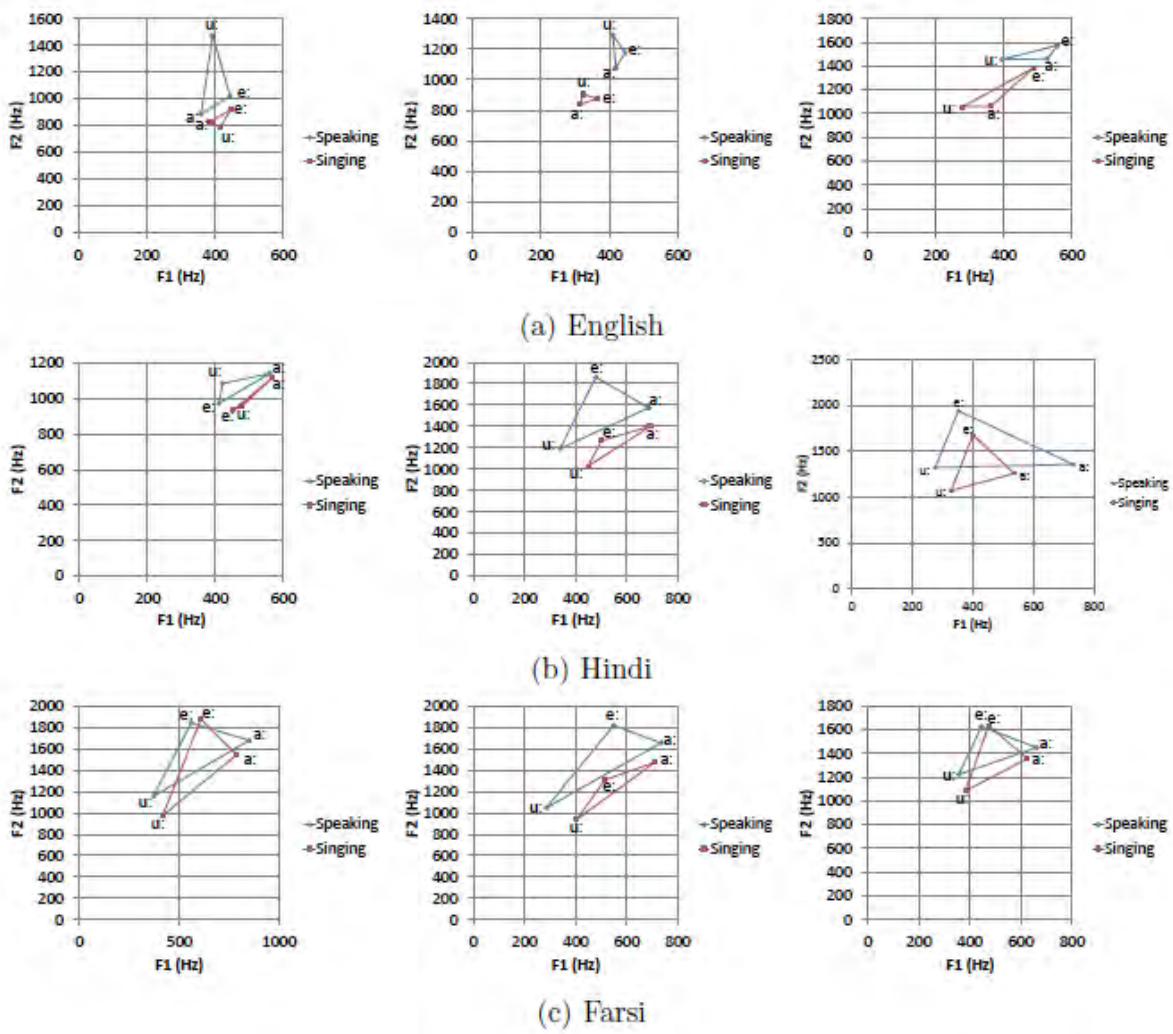


Figure 13. *Vowel Space configuration of vowels for (a) 2 female (left) and one male (right) English speakers, (b) 2 female (left) and one male (right) Hindi and (c) 2 female (left) and one male (right) Farsi speakers.*

Further experiments were also performed on speaker model comparison for speakers of three languages (English, Farsi and Hindi). For this purpose, we trained a separate GMM model for each speaker in each language in the two speech styles: reading and singing. We then used a KL divergence metric to find the difference in the probability distributions of the two GMM's. The results are plotted in Figures 16 and 17. Figure 16 shows the box plot with the mean and distribution of KL divergences for each speaker during reading and singing: Fig. 17 shows KL divergence for GMM models of each speaker during reading and 7-12 shows the KL divergence for the same speakers, but during singing.

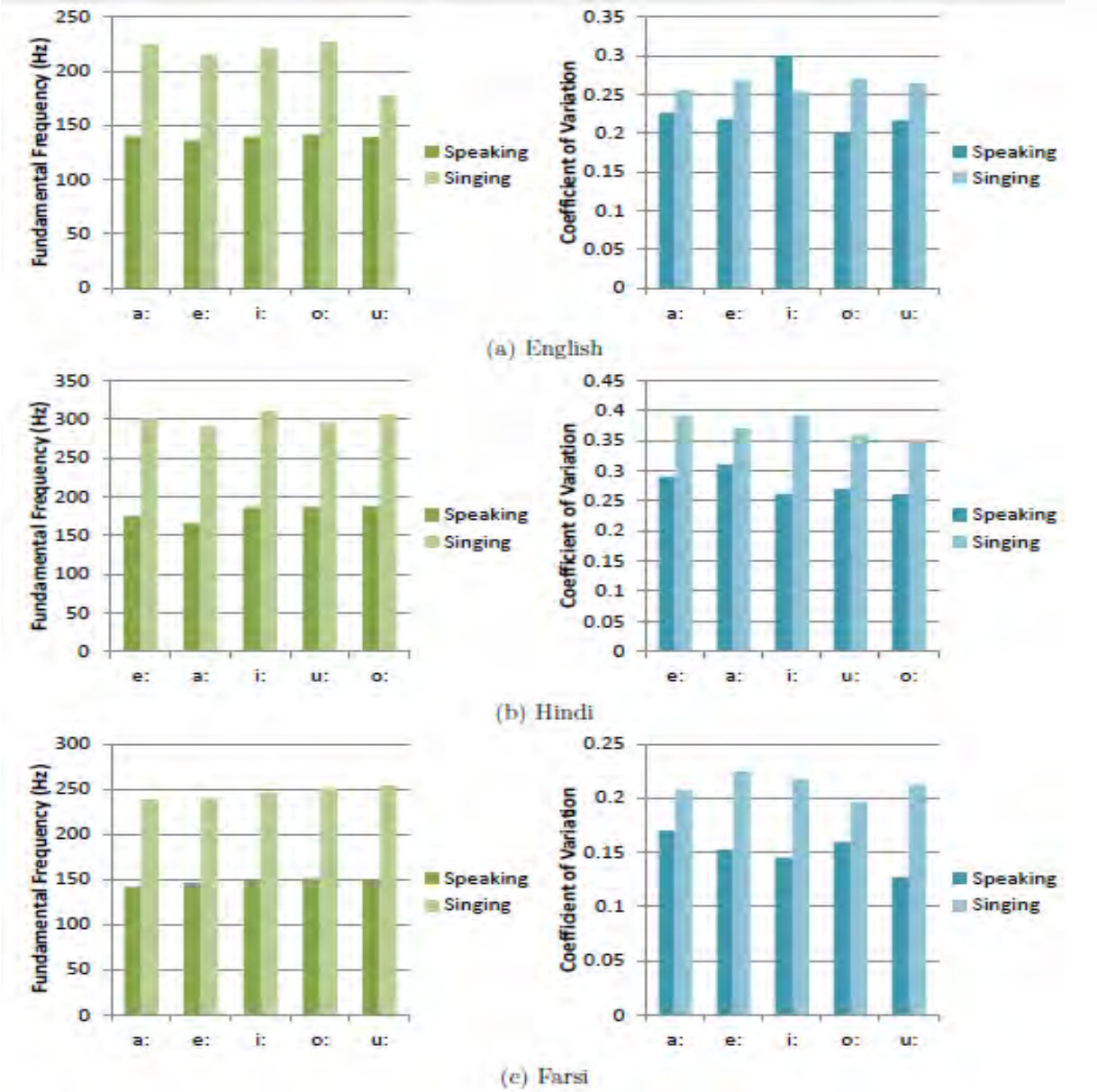


Figure 14: *Fundamental Frequency and Coefficient of Variation for (a) English, (b) Hindi and (c) Farsi speakers.*

As we can see from Figure 17, the mean KL divergence of each speaker is significantly higher when the regular speech model is compared to singing. Generally, the Hindi language produces the highest KL divergence in cross model comparison of singing and reading (mean KL divergence is 15.27). The Farsi language ranks second highest with a mean cross-model KL divergence of 14.91. The mean cross-model KL divergence of English speakers is only 7.47.

We have also plotted the KL divergence results in a 3-D plot (Figure 17) in order to visually express the change in KL divergence when changing the speaking style from regular speech to singing. We can also see the shift in KL divergences across different languages.

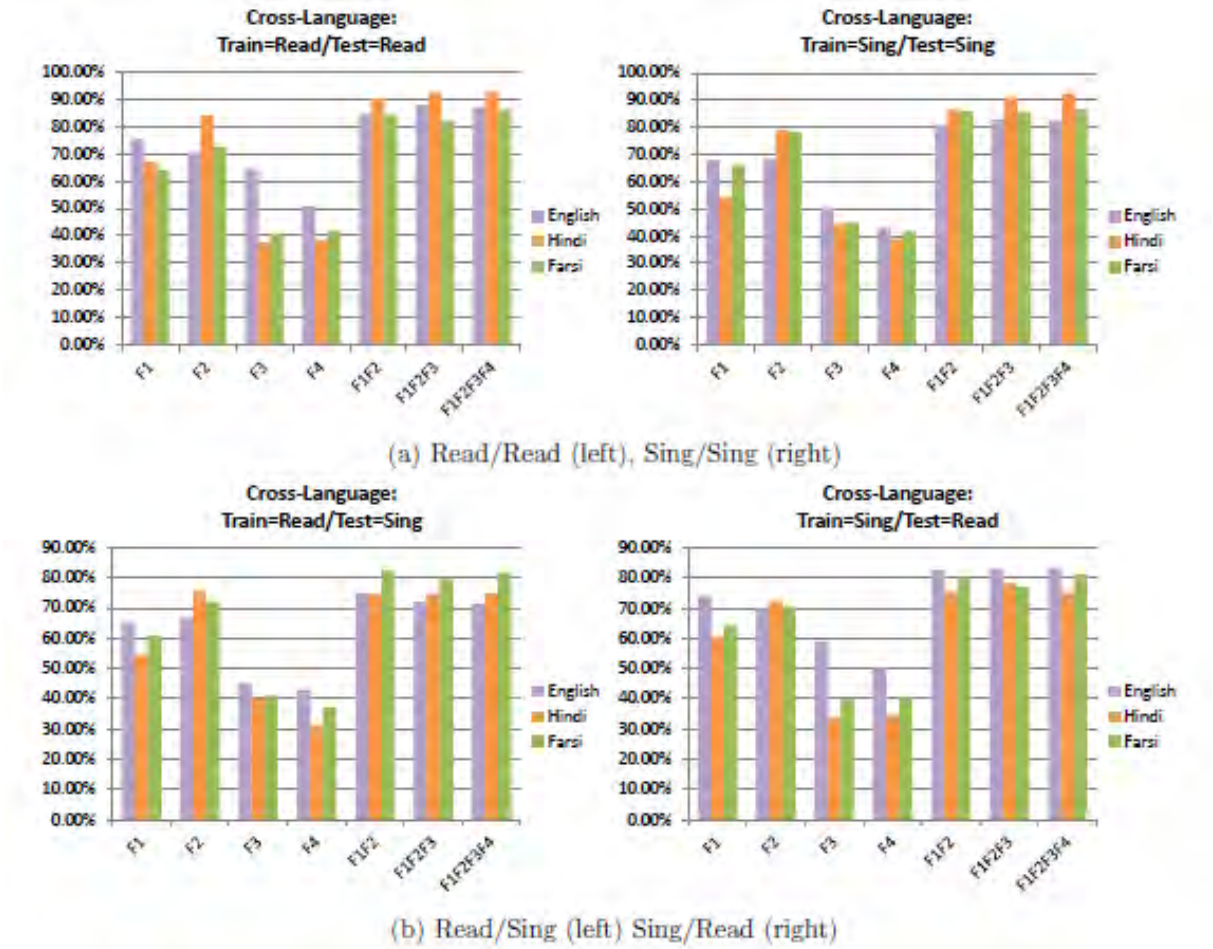


Figure 15. *English, Hindi and Farsi K-NN (k=10) accuracy for different train/test styles: (a) Read/Read (left), Sing/Sing (right) and (b) Read/Sing (left) Sing/Read (right)*

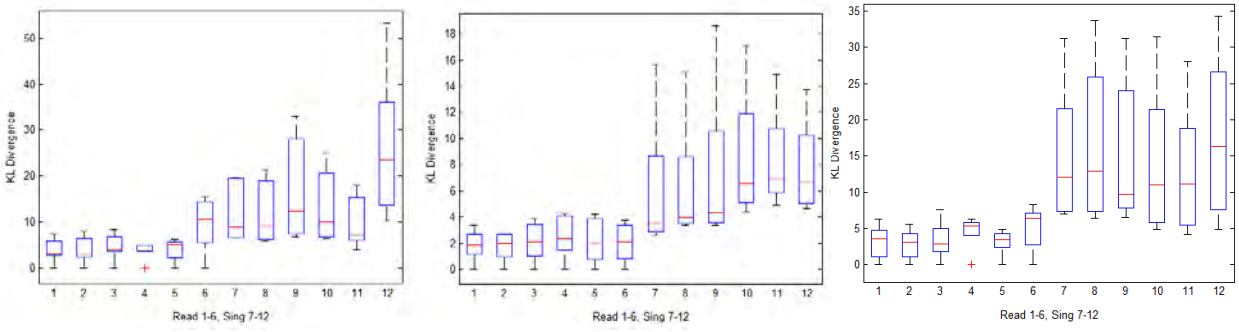


Figure 16. *KL Divergence of Reading and Singing for speakers of (a) Farsi, (b) English, and (c) Hindi*

More recently (spring 2015), a speaker identification based open set system was developed and performance benchmarked when singing and spoken speech applied. For this case, a GMM-UBM based system was employed. A universal background model (UBM) was constructed using TIMIT data with 438 male speakers and 192 female speakers. A total of 1024 mixtures were used to train the UBM. Next, a Maximum A Priori (MAP) adapted Gaussian mixture model was obtained for each of the English, Hindi and Farsi speakers. For MAP adaptation, a total of 17 speakers of English, 17 speakers of Hindi, and 13 speakers of Farsi were used.

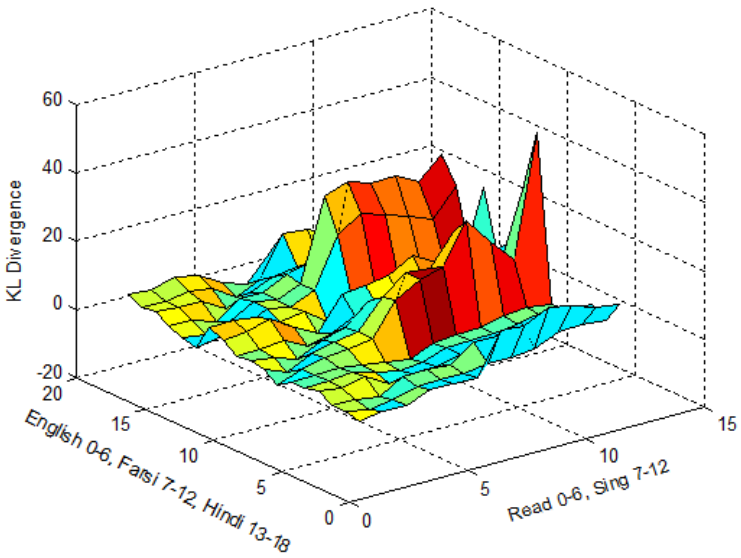


Figure 17. *KL Divergence of speakers of all 3 languages of Reading and Singing*

Results are shown in Table 4. It is observed that the Hindi speaker ID system performs significantly better than the English system, and slightly better than the Farsi system when trained and tested with neutral speech. However, Hindi shows the highest degradation in performance (difference in EER = 30.54%) when singing speech is introduced. Farsi ranks second with a degradation of EER = 23.81%, and English with the lowest degradation of EER = 20.17%. Our previous results on KL Divergence indicated that Hindi and Farsi show the highest dissimilarity between singing and speaking. If we compare these results to our speaker identification results, it can be seen that system performance degrades mostly in these two languages. The English language showed the lowest dissimilarity between singing and speaking (lowest KL divergence out of the three languages), and these results carry over to speaker identification performance as well. We can see that English shows the lowest performance degradation when the system is tested with singing versus neutral speech. Our results in the speaker identification are strongly tied to the similarity measure (KL Divergence) between singing and speaking. The more regular speech deviates from singing, the higher will be the degradation in performance of speaker identification system.

Table 4. *Equal Error Rate (EER) of Speaker ID system trained on neutral read/spoken speech and tested on spoken/read or singing for English, Hindi and Farsi subjects.*

SID	English	Farsi	Hindi
Train = Read Test = Read	18.1818	12.8553	11.8464
Train = Read Test = Sing	38.3578	36.6720	42.3935

Task 1.4. Speaker ID based on Non-Speech Sounds

In this area, the effort has concentrated on speaker recognition of acoustic material from speakers who are not producing speech, but instead producing non-speech acoustic sounds. The specific non-speech sounds that were the focus include: (i) screams, (ii) whistles, and (iii) coughs. This effort took place over the past year (not the entire 36 month period), and will continue as potential follow-on research. In the current completed effort, a new framework has been developed for classification of scream and human whistles [1,2,4]. The motivation was based on the issue that individuals in the field were attempting to employ speaker recognition technology to screams for speaker ID, when in fact the technology had never been formally evaluated for that form of vocalization.

1.4.1: UT-NonSpeech Corpus Collection:

At present, there is no corpus available for research on non-speech vocalizations, and specifically for exploring speaker recognition for such audio. In order to advance research for non-speech vocalizations, two corpora were developed at the Center for Robust Speech Systems (CRSS), UT Dallas.

(i) *UT-NonSpeech-I*: This corpus contains material from 6 male speakers and one non-speech vocalization (human scream). The focus was probe experiment for Scream Analysis.

(ii) *UT-NonSpeech-II*: This corpus contains a total of 56 speakers (33 males and 23 females). Other than human scream vocalizations, speaker specific whistle and cough sounds were also recorded. (UTD IRB file #14-09)

Both corpora also have read speech and spontaneous speech for each speaker.

1.4.2: Analysis of Human Scream: A detailed analysis of human screams was performed to identify discriminating features from neutral speech. Analysis was performed in terms of four different acoustic parameter and following differences were observed:

(i) *Fundamental Frequency*: A drastic increase in F0 was observed. Mean F0 was increased up to two times in case of some speakers. Standard deviation across F0 distribution for speakers were greater in case of scream versus neutral speech.

(ii) *Frame Energy Distribution*: In scream vocalizations, the number of high energy frames are much greater compared to speech. Also the variance of frame energy distribution is more for speech compared to scream.

(iii) *Spectral Peak/ Formant Shift*: Scream is produced with a greater oral/lip mouth opening and therefore is expected to have a shift in spectral peaks, particularly the first formant like resonance F1. Also, lower jaw position during scream resulted in an increase in F1 location. The vowel space designated by F2 and F3 also shifts. It was also seen that as higher spectral peaks are considered, the separation between scream and speech formant locations is reduced.

(iv) *Spectral Slope:* Spectral slope was steeper for neutral speech compared to scream. The change in slope for scream suggests that there are more regular shape glottal pulses in scream vocalization compared to speech, and that there is more balance between low and high frequency energy.

1.4.3: Impact on Speaker Verification Systems:

It was observed that for the case of speaker verification from speech and scream, PMVDR front-end performed better compared to the MFCC front-end. In the case of scream trials, system performance decreases drastically for both feature types. Because of the difference between excitation structure for speech versus scream portions, speaker dependent information is greatly suppressed in human screams based on EER values [1].

1.4.4: Detection of Human Screams in Noisy Environments:

The T2-BIC-SAD based solution was used for detection of human screams in noisy acoustic environments. The proposed solution is unsupervised in nature.

Also, five different noise environments were considered from NOISEX-92 database as well as five different noise levels. The first task was to assess performance of vocal activity detection followed by speech/scream detection.

- SAD: Vocal activity detection (TO-Combo-SAD) results show that for low levels of noise, the SAD EER is below or near 10%, but as noise levels increase, the EER degrades rapidly (i.e. EER approaches 50%). It should be noted that T2-BIC is more suited for unsupervised wide duration audio streams.
- Evaluation of T2-BIC based detection for scream detection works well for clean +20, +10 db SNR levels, with performance declining as SNR decreases to -20 dB across a number of noise sources considered.

Fig. 18 illustrates the TO-Combo-SAD decision and detected break points for scream in a continuous audio stream.

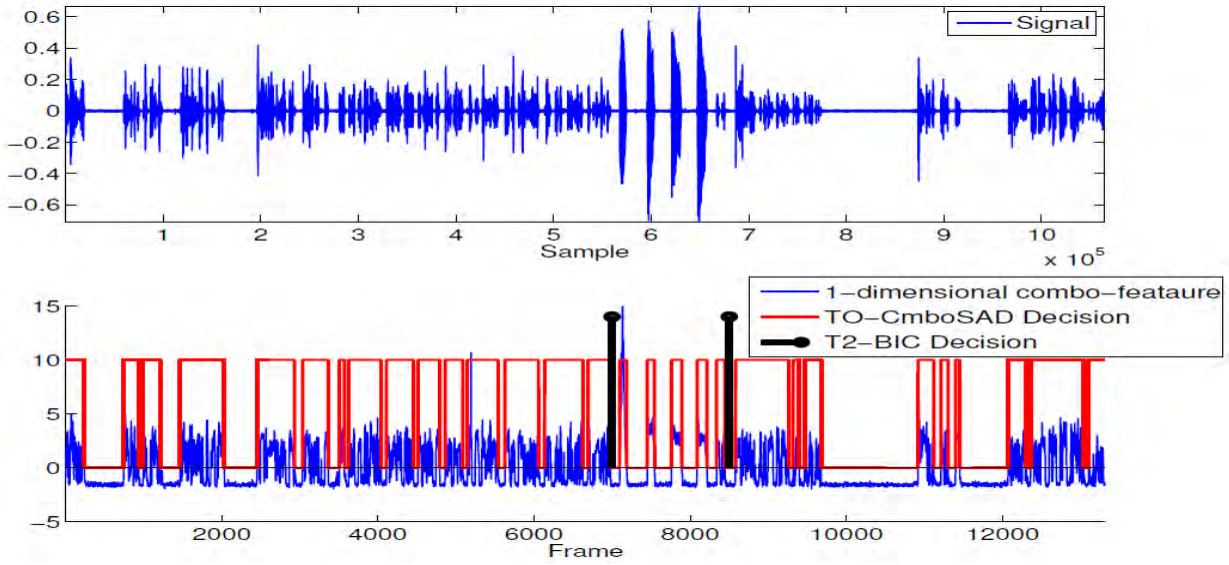


Figure 18: *TO-Combo-SAD decisions and scream detection using CompSeg across an audio stream.*

1.4.5: Human Whistle Analysis and Classification:

In addition to scream, human whistle has also been explored as another class of non-speech vocalizations. Human whistle, a single tone sound, is produced by controlling the stream of air flow generated via lungs, from the oral cavity. Here, the oral cavity works as a resonant chamber.

For this work, since there are a variety of forms of whistle, the focus has been on pucker whistles. Pucker whistles are produced by curving the tongue inside the oral cavity such that the top of the tongue touches the roof of the oral cavity, where the tip of the tongue should be downwards to create turbulence followed by blowing out, or sucking air into the mouth. Alternative resonants can be produced by changing the shape of the tongue and position of the jaw.

Human speech and whistle samples were analyzed in terms of their spectral and pitch properties. The Spectral center of gravity (SCG), representing the `center of mass' of the power spectrum, and spectral energy spread (SES), which represents the standard deviation of the spectral energy distribution from SCG, were extracted from speech and whistle samples.

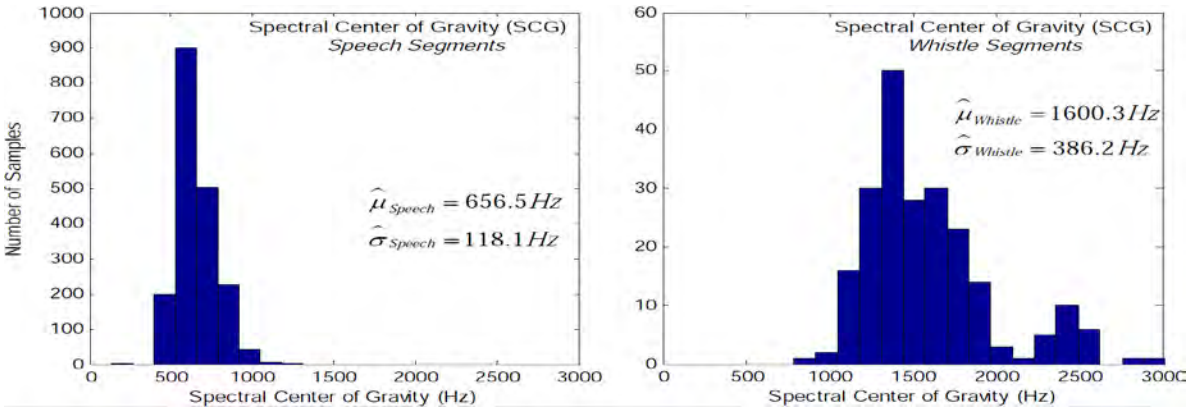


Figure 19: *Distribution of spectral center of gravity (SCG) in speech and whistle samples.*

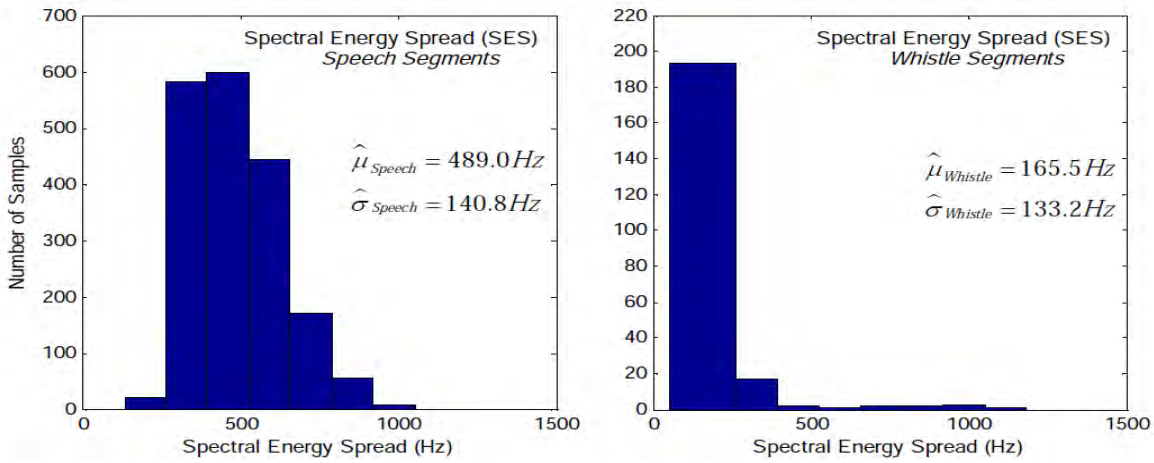


Figure 20: *Distribution of spectral energy spread (SES) in speech and whistle samples.*

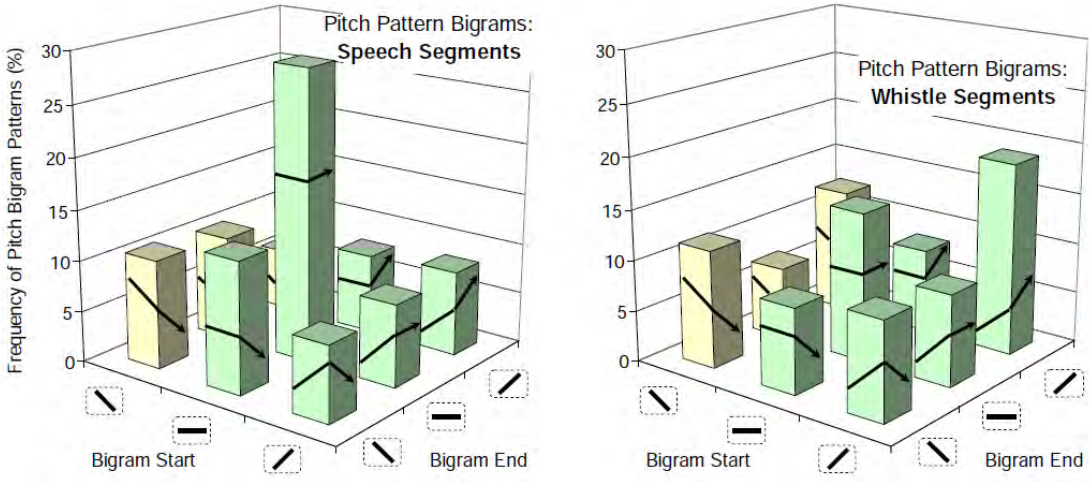


Figure 21: *Frequency of bi-gram pitch patterns in speech and whistle.*

- It can be seen that the SCG distribution of speech is sharper and centered at a significantly lower frequency compared to whistle. Due to the variety in the choice of a whistled melody and pitch, the

overall SCG distribution extracted across all subjects displays larger variation compared to speech. (Fig. 19)

- As expected, SES distributions in Fig. 20 show the opposite trend compared to SCG – the energy spread around SCG within individual samples is much wider for the 'broad' speech spectrum compared to sharp whistle spectra.
- Frequencies of pitch pattern bigrams were analyzed. It can be seen from Fig. 21 that *flat-flat* bigram dominates speech pitch contours and *up-up* is dominating in whistle. The overall distribution of pitch pattern bigrams is more uniform in whistle than in speech, suggesting higher variability of the whistled pitch contours.

1.4.6: Human Whistle Classification:

Based on the fundamental differences in speech and whistle spectral and pitch, we proposed a combination of time-domain and spectral-based parameter for speech/whistle classification. The feature vector includes: zero-crossing rate (ZCR), spectral center of gravity (further denoted as spectral centroid – SC), spectral energy spread (further denoted as 'SS'), spectral crest factor (SCF), spectral decrease (SD), spectral kurtosis (SK), and spectral skewness (SSk).

Front-End	#Dim	No Filter		3*pt* Median Filter		5*pt* Median Filter		7*pt* Median Filter	
		SVM	GMM	SVM	GMM	SVM	GMM	SVM	GMM
MFCC	12	75.4	90.0	75.2	91.6	75.2	94.2	75.1	**94.6**
Proposed	7	84.4	91.4	87.1	93.3	88.8	**94.6**	89.9	94.2
Fusion	19	84.7	94.8	87.4	97.0	89.2	96.9	90.1	**97.7**

Table 5: *Frame-level classification results for different front-end.*

All the seven dimensions were combined and normalized for classification. As can be seen in Table 5, with one exception, the classifier utilizing the proposed feature vector outperforms the MFCC baseline both for SVM and GMM based classifiers. Median filtering of the frame-level decisions further improves the classification performance.

1.4.7 Impact of Whistle on SID and Compensation:

In this study, the impact of human whistle on speaker verification was also demonstrated. To observe the effect of whistle on speaker verification, audio files consisting of sequences of speech segments alternating with whistle segments were constructed. A GMM-UBM based SID system was used.

- EER in case of testing with neutral speech is 10.80% whereas when we test it against speech-whistle audio streams it increases to 21.60%.

For the compensation of this train/test mismatch we use our proposed front-end for speech/whistle classification. Entire scheme is depicted in Fig. 1-22.

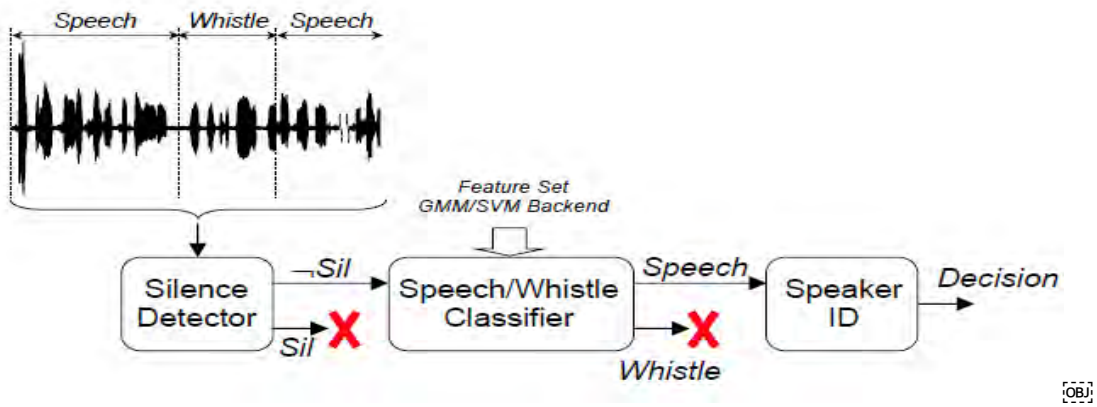


Figure 22: *Proposed compensation scheme for speech/whistle mismatch in speaker verification.*

The results for SID after applying the whistle based compensation scheme are reported in Table 6. Further details are presented in the reference for this sub-section [4], however the performance in Table 6 shows that there is benefit for both SVM and GMM based speaker ID systems with the proposed strategy. From Table 1-6, it is clear that the proposed scheme results in a lower EER by a larger margin.

Table 6: *Speaker verification compensation results.*

EER (%)	feature	7pt Median filter	
Test on		SVM	GMM
speech+whistle	proposed	16.33	18.00
speech+whistle	proposed+MFCC	15.38	15.11

Publications:

1. Nandwana M.K., Hansen J.H.L.,"Analysis and Identification of Human Scream: Implications for Speaker Recognition," *Interspeech-2014*, 14-18 Sept. 2014, Singapore.
2. Nandwana M.K., Ziaei A., Hansen J.H.L.,"Robust Unsupervised Detection of Human Screams in Noisy Acoustic Environments," *IEEE ICASSP-2015,* 19-24 April 2015, Brisbane, Australia.
3. Nandwana M.K.,"Analysis and Classification of Non-Speech Vocalization with Applications to Speaker Recognition," M.S. thesis, University of Texas Dallas, TX.
4. Nandwana M.K., Boril H., Hansen J.H.L.,"A New Front-End for Classification of Non-Speech Sounds: A Study on Human Whistle," *Interspeech-2015,* 6-10 Sept. 2015, Dresden, Germany. (**submitted)**

Task 1.5: Classification of Physical Task Stress with Application to SID:

- Stress is an external aspect that impacts physical speech production when people produce speech while performing secondary tasks. Addressing noise is not sufficient to overcome performance loss in actual noisy stressful scenarios for robust speech systems, even if noise is eliminated completely [1]. Speech production variability introduced by stress or emotion can severely degrade speech/speaker recognition accuracy [2-4]. Detection of paralinguistic information, such as physical task load, gender and cognitive load can guide human computer interaction systems to automatically understand and adapt to different users states and environments. Thus, this technique can be directly applied to stress level classification [5], as well as emotion surveillance. At the same time, it can also be employed as a front-end for spoken dialog systems, speaker diarization, speaker identification and automatic speech recognition (ASR) systems.

Table 7 shows an overview of factors which impact physical task stress detection. Physical status (e.g. heart rate) is changing with exertion, which can be reflected in the corresponding speech [6]. The acoustic environment or noise level varies with different physical task scenarios, which could be sustained background noise in a typical 24-hour operating workplace or random noise in a gym. Even if we remove all external environmental factors, physical task stress still shows differences within speaker (e.g. the same speaker, different exercise durations give different stress load) and across speaker (e.g. the same task, different speakers show different stress load levels). These factors taken together make physical stress detection a challenging research task.

Table 7: *Influential factors for speech under physical task*

Physical Changes of Speaker	Noise type / Environment	Speaker Variability
● Heart rate ● Breathing ● Fatigue ● Muscle control	● Workplace ● Gym ● Constant ● Random	●Within speaker ●Across speaker

Research in this sub-area has been performed over the past 12 months of the 36 month research effort. Advancements have focused on analysis, acoustic features, i-Vector framework and physical task stress detection.

1.5.1: UT-Scope-Physical: Speech corpus under physical task stress condition:

This study employs the UT-Scope Physical task stress Corpus for system development and evaluation. From the corpus details described in [7], physical task stress is introduced into speech by having subjects exercise on a Stamina Conversion II Elliptical/Stepper machine in the elliptical mode. There are 66 speakers in the UT-Scope Physical Corpus. For this portion of the study, we employ 50 female speakers, each producing 35 sentences under neutral and physical conditions. We consider the stress classification experiments in a speaker independent scenario. In each experiment, 40 speakers are used in the training set, and 10 subjects in the test set. Next, we rotate the training/test set, resulting in 5 speaker independent physical stress detection experiments where all speech and speakers are open test. For more details, please see Table 8.

- ### 1.5.2: Acoustic Features for Physical Task Stress Detection:

MFCCs: 39-dimention feature vector (13 MFCC+Δ+ΔΔ).

TEO-CB-Auto-Env: The TEO profile obtained from the critical band based Gabor bandpass filter output is segmented on a short-term basis, Auto-correlation is applied after framing. Once the auto-correlation response is found, the area under the autocorrelation envelope is obtained and normalized. One area coefficient is obtained for each filter bank. This area coefficient is intended to determine the regularity of speech production, it has been shown to be large for neutral speech and low for speech produced under stressed conditions [8,9]. In this study, we employ an 18 dimensional Gabor filterbank. Thus, 18 dimensional TEO-CB-Auto-Env features are extracted from each frame.

Table 8: *Statistics of UT-Scope Corpus.*

	Set1	Set2	Set3	Set4	Set5
SNR/dB	33.28	36.87	36.07	35.84	37.78
Duration/s	2.56	2.86	3.06	2.97	3.26
Speaker Count	10	10	10	10	10

- **1.5.3: I-Vector Framework:** Our proposed system for physical task stress detection utilizes the concept of i-Vector modeling, which is proposed in [10,11]. By constraining the total variability into a lower dimensional total variability space, the i-Vector is capable of effectively representing the variability factors within each speech utterance [11,12]. In this work, we attempt to model the speaker-independent physical task stress using i-Vectors. Fig. 23 shows the i-Vector framework used in our study. For each speaker independent experiment, a Universal Back-ground Model (UBM) with 256 Gaussian mixtures is trained using the training dataset outlined in corpus description.
- **1.5.4: I-Vector based Physical Task stress Detection with Fusion Strategies.** Two fusion strategies are used to further improve system performance.

  I-Vector level fusion: Using the i-Vector extraction described in Fig. 23, two kinds of i-Vector are derived from each utterance, (i.e., MFCC-based and TEO-CB-Auto-Env based i-vectors). The new i-Vector integrating both MFCC and TEO-CB-Auto-Env acoustic information is obtained by concatenating the two i-Vectors together. The dimensionality is reduced to the original length using linear discriminative analysis (LDA) [13].
  Score level fusion: We apply Adaboost algorithm to do score level fusion [14]. Fig. 1-24 is the flow diagram of score fusion we proposed. When applying AdaBoost to our physical task stress detection system, we assume the MFCCs score S_1 and TEO-CB-Auto-Env score S_2 are presented by a weaker classifier respectively (Actually, we can claim additional weaker classifiers to present a two-dimension feature, since third or high number classifiers are just the linear combination of first two classifiers).

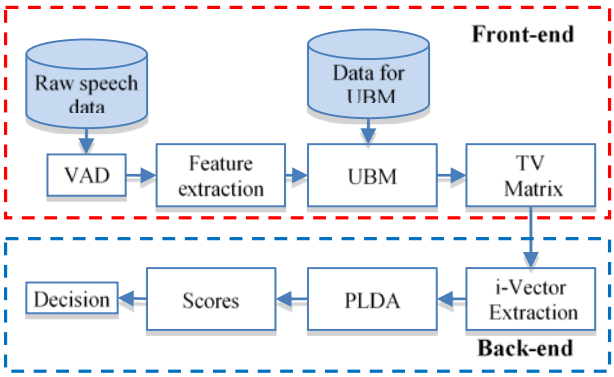


Figure 23: *Flow diagram of i-Vector framework.*

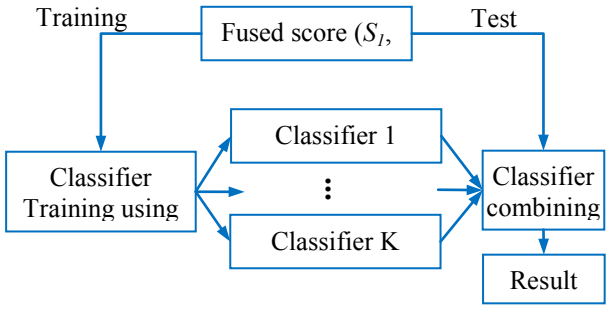


Figure 24: *Flow diagram of score fusion using AdaBoost.*

- **1.5.5: Experimental Results:**

From the speaker independent physical stress detection experiments given in Fig. 25, we can see: a) both i-Vector fusion and score fusion achieve reasonable performance, which

show the effectiveness of MFCCs and TEO-CB-Auto-Env features and their complementary effects; b) compared to i-Vector fusion, score fusion always performs better than single feature based systems, which shows the stability of our proposed approach; c) there is a greater than 10% percent accuracy difference between Set1 and Set5 indicates the variability across speakers; and d) although Set4 has lower SNR and shorter duration compared to Set2 and Set3, the relative better detection performance further shows the across speaker variability of physical stress, which reflects a challenge in formulating a robust physical stress model.

To examine the physical exertion level reflected by speech within each speaker, we split each speaker set into 3 parts over the exercise fine frame (e.g., begin, middle, end) and repeat the physical stress detection experiments. We assume the physical stress load follows in this order: Phy3>Phy2>Phy1, since the entire exercise time period follows in the same way. The results from Fig. 1-26 show: a) physical stress load increases with the exercise time period, which indicates the variability introduced by physical exertion level within each speaker with exception of Set1, others generally show increasing physical stress level over time; b) the results here show that effective physical stress detection is possible, and increasing levels of stress are seen across the exercise period; c) it should also be noted that corresponding heart-rate monitoring during speech production for the UT-Scope Physical Corpus collection confirm the increased levels of physical task stress [6].

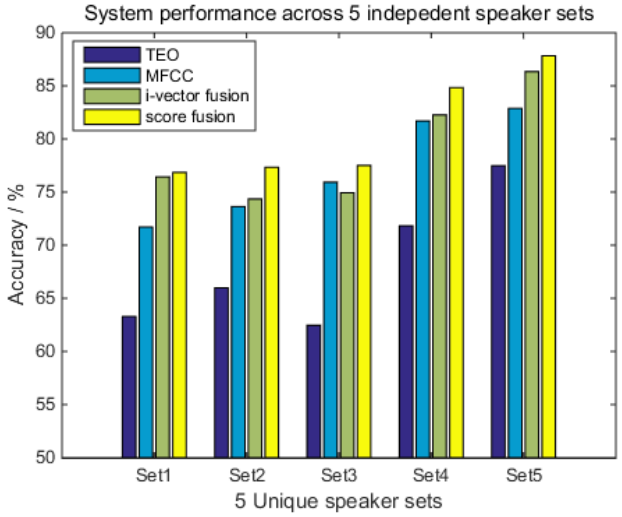


Figure 25: *System performance across 5 unique speaker set. We use scores from MFCC system and i-Vector fusion system to perform score-level fusion.*

- ## 1.5.6: Conclusions – Physical Task Stress and Speaker ID/Traits

In this study, an i-Vector based physical task stress detection system was proposed. MFCCs and TEO-CB-Auto-Env based features were investigated in an i-Vector framework for stress detection tasks. Using i-Vector fusion, a relative accuracy gain of +2.68% is obtained; by score fusion using the AdaBoost algorithm, a further relative +6.52% performance gain is achieved (both compared to best single feature system used in our study, e.g. MFCC based i-Vector system). It also notes that our proposed system outperforms human listener testing described in [6]. The i-Vector dimensionality for our specific physical task stress detection is determined by parameter tuning. Variability across and within speakers was investigated. From the experiments presented in Fig. 26, it has been shown that approximate physical exertion level differences are represented in the speech signal. Future work will focus on physical stress level classification, especially over speakers given heart rate ground truth. Also, other variations such as gender, channel or age will be explored.

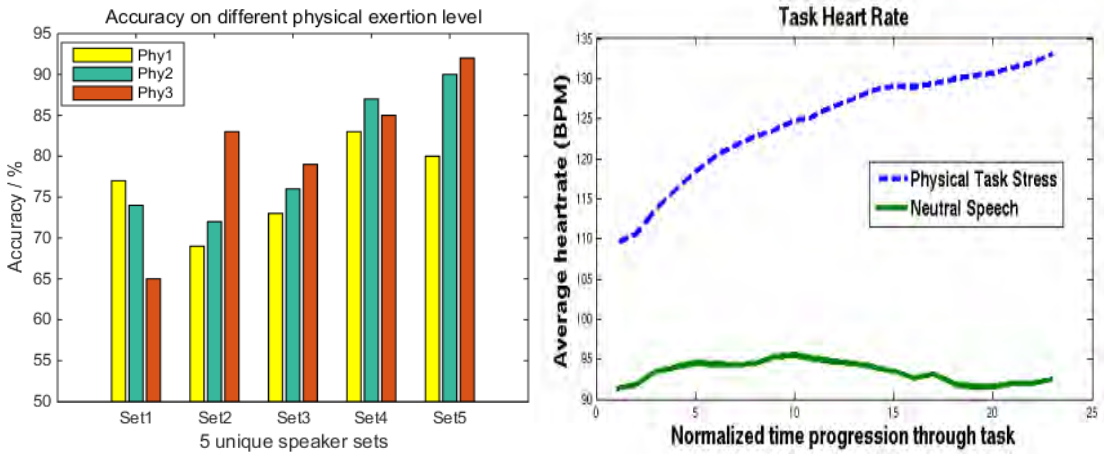


***Figure 26**: Left) Detection performance on physical stressed speech employing MFCC based i-vector system. Each speaker produces 35 utterances under physical task condition. We split them into three categories (i.e. Phy1 for 1-12, Phy2 for 13-24 and Phy3 for 25-35); Right) Physical task duration VS. heartrate [6].*

o 1.5.7: Next Phase/Steps – Physical Task Stress and Speaker ID/Traits

The **long term vision** is that the technique could be applied directly to physical task stress detection/classification (for example, physical load monitoring for over fit people or astronauts committing a mission in out space) as well as speech systems (e.g. speaker identification and speech recognition)

- **Physical stress level classification**

Presently, automatic stress detection methods for speech employ a binary decision approach, deciding whether the speaker is or is not under stress. Since the amount of stress a speaker is under varies and can change gradually, a reliable stress level detection scheme becomes necessary to accurately assess the condition of the speaker. The experimental results from Fig. 1 give intuition that physical stress classification is possible.

- **Unsupervised clustering techniques for stress level classification**

In this previous study, the stressed speech recordings were classified into three categories manually (i.e. begin, middle and end of exercise). In practice, it is not expect that it will be possible to always know the target speech segment corresponding to which exercise stage. Also, the variability across speakers show that it is better to classify physical load level by actual stress information from speech not by exercise duration. Thus, unsupervised clustering can be employed to label the stress load level [15].

- **Model adaption using stressed speech for SID or ASR**

If it is possible to correctly classify physical stress, then it would be possible to use stressed speech to adapt acoustic models for SID or ASR systems. Thus, performance gain is expected by employing stress detection as a front-end of these systems.

- **Standard physical task Corpus**

In the next research period, more physical task stressed speech should be collected for further development. A new corpus should enroll more male speakers to balance the male/female ratio in the current UT-Scope Physical Corpus. Also, a greater variety of physical tasks is needed to build a more generally represented physical stress corpora.

1.5.8: Reference for Physical Task Stress and SID

[1] J.H.L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication,* vol. 20, no. 1, pp: 151-173, 1996.

[2] J.H.L. Hansen, S. Patil, "Speech under stress: Analysis, modeling and recognition," in *Speaker Classification I*. Springer, 2007. pp. 108–137.
[3] J.H.L. Hansen, A. Sangwan, W. Kim, "Speech under stress and lombard effect: impact and solutions for forensic speaker recognition," in *Forensic Speaker Recognition*. Springer, 2012, pp.103–123.
[4] C. Baber, B. Mellor, R. Graham, J.M. Noyes, C. Tunley, "Workload and the use of automatic speech recognition: The effects of time and resource demands," *Speech Communication*, vol. 20, no. 1, pp. 37–53, 1996.
[5] J.H.L. Hansen, E. Ruzanski, H. Bořil, J. Meyerhoff. "TEO-based speaker stress assessment using hybrid classification and tracking schemes." *International Journal of Speech Technology,* vol. 15, no. 3, pp: 295-311, 2012.
[6] K. W. Godin, J.H.L Hansen. "Analysis and perception of speech under physical task stress." ISCA *INTERSPEECH*, 2008.
[7] A. Ikeno, V. Varadarajan, S. Patil, J.H.L. Hansen, "UT-Scope: Speech under Lombard effect and cognitive stress," in Aerospace Conference, 2007 IEEE, 2007, pp. 1–7.
[8] G. Zhou, J.H.L. Hansen, J. F. Kaiser, "Nonlinear feature based classification of speech under stress," IEEE Transactions on Speech and Audio Processing, vol. 9, no. 3, pp. 201–216, 2001.
[9] K. Wooil, J.H.L. Hansen. "Robust angry speech detection employing a TEO-based discriminative classifier combination." In *Proc. INTERSPEECH*, 2009, pp. 2019-2022.
[10] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speech, Language Processing, vol. 19, no. 4, pp. 788–798, 2011.
[11] P. Kenny, G. Boulianne, P. Dumouchel, "Eigenvoice modeling with sparse training data," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 3, pp. 345–354, 2005.
[12] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, R. Dehak, "Language recognition via i-vectors and dimensionality reduction." in Proc. INTERSPEECH. Citeseer, 2011, pp. 857–860.
[13] G. Liu, J.H.L. Hansen, "An investigation into back-end advancements for speaker recognition in multi-session and noisy enrollment scenarios," IEEE Trans. Audio, Speech and Language Processing, vol. 22, no. 12, pp. 1978–1992, 2014.
[14] Y. Freund, R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[15] B. Clarkson, A. Pentland. Unsupervised clustering of ambulatory audio and video. IEEE ICASSP-1999, vol. 6, pp. 3037-3040. 1999.

Task 1.6: Robust Processing for Speaker ID under Noise & Reverberation:

In this area, the focus has been on developing robust front-end solutions to combat mismatch between training and test conditions for automatic speaker identification (SID) systems. The mismatch could be due to background noise, room acoustics (reverberation), speaker variability in different recording sessions, and/or variability in communication channels. In our previous reports, we presented our progress on developing features and noise/reverberation compensation strategies and showed SID results for simulated noisy and reverberant conditions [1], [2].

Table 9. *Speaker verification results on the MultiRoom8 corpus with BSW, LTLSS, and NMF as pre-processing stage in the MFCC feature extraction*

Set	**EER (%)**				
	MFCC	**MFCC+BSW**	**MFCC+LTLSS**	**MFCC+NMF**	**MHEC+BSW**
1	13.16	10.53	13.16	13.87	8.18
2	11.17	7.89	10.53	15.86	7.46
3	21.15	16.28	20.93	20.93	13.95
4	13.46	9.30	11.63	16.28	6.74
5	11.67	10.26	14.91	12.82	9.84
6	19.44	16.67	22.22	25.00	19.44
7	10.93	7.96	10.66	10.66	7.22

1.6.1: Blind Spectral Weighting (BSW) for Reverberation Mitigation in Speaker Identification: Over the past three year project, a blind spectral weighting (BSW) technique has been proposed in the short-time Fourier transform (STFT) domain [3] for reverberation mitigation in SID. The proposed technique is blind in the sense that prior knowledge of the room impulse response (RIR) or the reverberation time is not required. Our technique can be easily integrated into the extraction process of popular features such as MFCCs. We have shown that the proposed technique, when incorporated as a pre-processing stage in the MFCC feature extraction framework, results in significant SID performance improvements under actual reverberant mismatched conditions (i.e., MultiRoom8). We have benchmarked the proposed technique against two other blind reverberation mitigation techniques, namely long-term log-spectra subtraction (LTLSS) [4], and Gammatone sub-band based non-negative matrix factorization (NMF) [5]. Actual reverberant data from the MultiRoom8 corpus were utilized for evaluations. Results are presented in Table 9 in terms of equal-error-rate (EER). Our proposed technique has two major advantages: 1) it is more effective, and 2) there is no need for signal reconstruction with our technique, as required by the other two methods.

Also given in Table 9 are results obtained with MHEC features in combination with the proposed BSW technique. It is clear that this combination (i.e., MHEC+BSW) provides the most robust solution for SID under actual reverberant mismatched conditions.

In addition to the SID under reverberation and noisy conditions, the MHEC features have also been adopted for language identification (LID) [6], as well as large vocabulary continuous speech recognition (LVCSR) [7] tasks. Very promising results have been obtained for both applications; in fact, better LID and LVCSR performances are achievable when the MHEC replaces the MFCC in the front-end.

1.6.2: Robust Unsupervised Speech Activity Detector for Harsh Acoustic Noise Environments

The goal here has been to develop a robust unsupervised SAD system for long audio recordings with small a priori speech presence probability, in order to help human listeners avoid listening to long noisy non-speech intervals. However, our system has the potential to be adapted for automatic speech application such as speaker and language identification, with only small modifications. Significant improvements have been make to our robust and unsupervised SAD system [12] which works based on several voicing measures and the perceptual spectral flux. The SAD was evaluated and compared against other commonly adopted SAD schemes, namely ITU G729B [8], SOLRT [9], SOLRT paired with an HMM-based hangover smoothing scheme [10], MOLRT [11], and a modified version of MOLRT that has been recently proposed [12]. The evaluation was performed using speech material from SPINE2 evaluation set and RATS dry-run data. SPINE2 evaluation set consists of 64 talker-pair conversations in stereo format (128 mono waveforms) recorded in simulated military background noise environments. On average, each of the mono waveforms is 180 seconds long and only contains 78 seconds of speech activity. Background types include quiet, office, Humvee, aircraft carrier, E3A, MCE field shelter. RATS dry-run data consists of a total of 111x 900-second long conversational telephone speech (CTS) waveforms that were retransmitted and recorded over 8 degraded communication channels with distinct noise characteristics and qualities.

Receiver operating characteristic (ROC) curves obtained from the evaluations are shown below. These curves indicate that: 1) as expected, RATS dry-run data is much more challenging for the SAD task. All the SAD schemes perform reasonably well on the SPINE2 evaluation set, however except for our proposed system and the modified MOLRT (hrmfrq_MOLRT), we observe a significant drop in speech detection performance on the RATS dry-run set. 2) Our multi-feature unsupervised system performs equally well on both high and low SNR data, which points to its robustness against environmental noise and channel conditions.

1.6.3: Comprehensive Evaluation of MHECs for SID using NIST-SRE 2010 extended Trials

It has previously been shown that MHECs are an effective alternative to MFCCs for robust speaker identification under noisy and reverberant conditions in relatively small tasks. Here, the effectiveness of these acoustic features has also been investigated in the context of a state-of-the-art speaker recognition system. The i-vectors were used to represent the acoustic space of speakers, while modeling and scoring

was performed via probabilistic linear discriminant analysis (PLDA). For evaluations, we utilized the NIST SRE-2010 extended telephone and microphone trials for both female and male genders. Results are presented in Table 10 in terms of EER. These results confirm consistent superiority of MHECs to traditional MFCCs within i-vector speaker verification, particularly under microphone and telephone training-test mismatch conditions (e.g., condition 3). In addition, fusion of subsystems trained with the individual front-ends proves that the two acoustic features (i.e., MHEC and MFCC) provide complimentary information for recognizing speakers.

For these experiments, we used our unsupervised SAD framework to remove the non-speech segments from conversations before feature extraction. Improvements seen in Table 10 are significant given the scale of experiments which is reflected in the number of trials used in evaluations (more than 6 million trials!). Detailed results of these experiments, which represent the most comprehensive NIST-SRE scale evaluation performed in our center, have been submitted to IS-2012 [13].

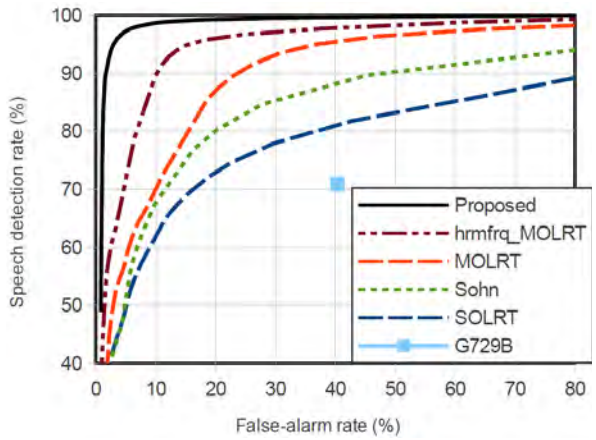


Figure 27. *ROC curve performance comparison on RATS dry-run data.*

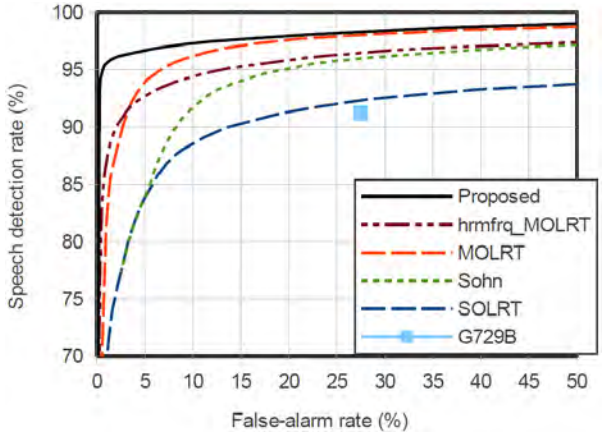


Figure 28. *ROC curve performance comparison on RATS SPINE2 data.*

Table 10. *MHEC vs MFCC performance on 2010 NIST-SRE extended male and female trials*

Gender	Condition	EER (%)		
		MFCC	MHEC	Fusion
Female	1	2.49	2.49	2.14
	2	4.47	3.99	3.50
	3	4.13	3.42	3.01
	4	2.97	2.68	2.37
	5	3.73	3.48	3.37
Male	1	1.04	0.71	0.61
	2	1.83	1.47	1.21
	3	2.78	1.97	1.92
	4	1.72	1.64	1.38
	5	2.53	2.10	1.85

1.6.4: References for Robust Processing for SID in Noise and Reverberation:

[1] S.O. Sadjadi and J.H.L. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. INTERSPEECH*, Sept. 2010, pp. 2138-3141.

[2] S.O. Sadjadi and J.H.L. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," *IEEE ICASSP*, May 2011, pp. 5448-5451.

[3] S.O. Sadjadi, H. Boril, J.H.L. Hansen, "A Comparison of Front-End Compensation Strategies for Robust LVCSR under Room Reverberation and Increased Vocal Effort," IEEE ICASSP-2012, pp. 4701-4704, (paper #3395), Kyoto, Japan, March 25-30, 2012

[4] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition," in *Proc. INTERSPEECH*, Sept. 2002, pp. 2185-2188.

[5] K. Kumar, R. Singh, B. Raj, and R. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in Proc. *IEEE ICASSP*, May 2011, pp. 4604-4607.

[6] G. Liu, S.O. Sadjadi, T. Hasan, J.-W. Suh, C. Zhang, M. Mehrabani, H. Boril, A. Sangwan, J.H.L. Hansen, "UTD-CRSS systems for NIST language recognition evaluation 2011," in *Proc. LRE Workshop*, Dec. 2011.

[7] S.O.Sadjadi, H. Boril, J.H.L. Hansen, "A comparison of front-end compensation strategies for robust LVCSR under room reverberation and increased vocal effort", in *Proc. IEEE ICASSP*, Mar. 2012, pp. 4701-4704.

[8] ITU, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70," in ITU-T Recommendation G.729-Annex B, 1996.

[9] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1-3, Jan. 1999.

[10] J. Ramirez, J.C. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, pp. 689-692, Oct. 2005.

[11] L.N. Tan, B.J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," *IEEE ICASSP*, Mar. 2010, pp. 4466-4469.

[12] S.O. Sadjadi, J.H.L. Hansen, "Unsupervised Speech Activity Detection using Voicing Measures and Perceptual Spectral Flux," IEEE Signal Processing Letters, vol. 20, no. 3, pp. 197-200, March 2013

[13] S.O. Sadjadi, J.H.L. Hansen, "Mean Hilbert Envelope Coefficients (MHEC) for Robust Speaker Recognition," ISCA Interspeech-2012, Wed-O7b-05, pg. 1-4, Portland, OR, Sept. 9-13, 2012

[14] S.O. Sadjadi, H. Boril, J.H.L. Hansen, "A Comparison of Front-End Compensation Strategies for Robust LVCSR under Room Reverberation and Increased Vocal Effort," IEEE ICASSP-2012, pp. 4701-4704, Kyoto, Japan, March 25-30, 2012

[15] T. Hasan, O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, J.H.L. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," IEEE ICASSP-2013, pp. 6783-6787, (paper #5335), Vancouver, Canada, May 26-31, 2013

Task 1.7: Speaker ID using DNNs, GPUs: application to Lombard Effect SID:

Deep Neural Networks (DNN) training using Graphics Processing Unit (GPU): for Speech Processing and Recognition. In this area, a limited effort was undertaken to explore the prospects of comparing GPUs + DBBs for speech based tasks. Goals identified so are as follows:

- Motivation: It was proposed to train and use a Deep Belief Network (DBN) consisting of stacked Restricted Boltzmann Machines (RBM) for UBM-like acoustic space division. The division based labels are then used to train a DNN with pre-initialized parameters rolled over from the DBN.

- Method: The NIST SRE corpora were used so that it's easier to benchmark the results and reproduce them. It also presents a more rigorous SID challenge for the system. The new system was tested on the NIST SRE 2010 dataset. A novel approach was used to cluster the development data which tries to replace the GMM-UBM based method. The approach is unsupervised as well since it comprises of training multiple layers of RBM stacked on top of each other to make a DBN. A new output layer is added for the DNN which consists of Softmax function nodes, as in classification tasks. The outputs of these nodes are then used as posterior probabilities for the i-vector statistics:

$$N_s^{C=c} = \sum_t P(c|X_t, \theta_{UBM})$$

Using these statistics, the i-vectors are extracted using:

$$w_s^* = (T' N_s \Sigma^{-1} T + I)^{-1} T \Sigma^{-1} F_s$$

- Experiments where performed for SID assisted by DNN-generated Baum-Welch statistics for i-Vectors generation. The RBM based UBM is used for supervised training of DNN models to provide more discriminative prediction for enrollment and test data in NIST SRE 2010 datasets.

  The experiments are focused on one central core train-test condition for male speakers out of the nine Common Evaluation Conditions in the NIST SRE 2010 Speaker Recognition Evaluation Plan, (i.e., *All different number trials involving normal vocal effort conversational telephone speech in training and test).* The core condition was based on an average of 5 minutes of speech for training and test. More development data is gathered to explore whether it assists the new DBN-DNN based models.

- A corresponding investigation using the same infrastructure was undertaken to explore Lombard effect "flavor" detection for improved speaker ID. The study

  The presence of Lombard Effect in speech is proven to have severe effects on the performance of speech systems, especially speaker recognition. Varying kinds of Lombard speech are produced by speakers under influence of varying noise types [1]. This study proposes a high-accuracy classifier using deep neural networks for detecting various kinds of Lombard speech against neutral speech, independent of the noise levels causing the Lombard Effect. Lombard Effect detection accuracies as high as 95.7% are achieved using this novel model. The deep neural network based classification is further exploited by validation based weighted training of robust i-Vector based speaker identification systems. The proposed weighted training achieves a relative EER improvement of 28.4% over an i-Vector baseline system, confirming the effectiveness of deep neural networks in modeling Lombard Effect.

Table 11 (a): *Classification Accuracies for Neutral and Lombard speech types; Unweighted means raw accuracy on test data, while Balanced means adjusted/weighted accuracy per class;* ***(b):*** *Confusion matrix for 4-way classification between neutral and Lombard speech (Classification rates are in %, figures in bold refer to matched train/test conditions).*

(a)

Classification Type	Neutral/ Lombard	Neutral, Noise-type	Neutral, Noise -type/level
Classes	2	4	10
Unweighted	95.7	69.1	60.0
Balanced	94.9	66.0	49.4

(b)

Test Condition	NEU	LCR	HWY	PNK
NEU	**94.2**	0.7	1.4	3.7
LCR	5.2	**43.3**	21.7	29.8
HWY	4.6	17.3	**62.8**	15.3
PNK	6.2	18.1	12.5	**63.2**

Further details from this probe investigation can be found in the following references:

- M.M. Saleem, G. Liu, J.H.L. Hansen, "Weighted Training for Speech Under Lombard Effect for Speaker Recognition," IEEE ICASSP-2015, Paper#3979, pp. 4350-4354, Brisbane, Australia, April 19-24, 2015.
- M.M. Saleem, "Deep Learning for Speech Classification and Speaker Recognition," CRSS: Center for Robust Speech Systems, Univ. of Texas at Dallas, completed Dec. 2014.

Task 1.8: Speaker ID and Tracking for Apollo Audio Streams.

Since speaker recognition performance can differ between controlled and unrestricted conditions, an effort was undertaken to pursue advancements for speaker ID and tracking for extended Apollo archive audio streams. Early in the analysis, it was observed that the astronaut voice fundamental frequency and speaker acoustic model differs significantly when subjects were in space. The study to considered fundamental frequency analysis, as well as formant location variation and vocal track spectrum displacements. Based on the study on four vowels /AA/,/AE/,/EH/ and /UW/, it was observed that the first formant of all vowels

increased a substantial amount. The experiments on vocal tract spectrum indicates that the sub-band of astronaut voice between 200-1000Hz has been stretched toward higher frequencies while the other sub-band that ranges between 800-2500Hz was mostly compressed toward lower frequencies.

Formant Locations: The mean of first and second formant of vowel /AA/ of all conditions across Apollo 11 mission are listed in Table 12. The mean of first formant of other four vowels are listed in Table 12. From Table 13, we could see that both frequency location of F1 and F2 of vowel /AA/ are consistently higher during space mission compared to the condition on Earth. The frequency location of F1 and F2 of astronauts are slightly lower when they were on the Moon compared to other times during space traveling. These variations of F1 and F2 are consistent for all three astronauts. In a similar way, Table 13 indicates that the first formant of other four vowels are also consistently higher during space mission. As the result of vowel /AA/, the F1 location of astronauts are slightly lower when they were on the Moon compared to others stages of space traveling except on vowel /AE/ and /UW/ of Aldrin's speech.

Vocal Tract Spectrum Shift: It is difficult to achieve reliable estimation of second (or higher) formant locations from Apollo 11 data due to its relatively low quality. Therefore, we propose a novel maximum likelihood frequency warping (MLFW) based analysis method to understand the vocal tract spectrum characteristic of astronaut in space. Maximum likelihood frequency warping based techniques, such as vocal tract length normalization (VTLN), have been widely applied in the community of speech recognition to mitigate the mismatch brought by vocal tract length differences between speakers. The motivation of MLFW based analysis is to detect the vocal tract spectrum shift caused by vocal tract variations in a maximum likelihood fashion. Compared to formant estimation which relies on the peaks of the spectrum, the MLFW based analysis are focus on the overall shape of vocal tract spectrum.

MLFW analysis focused on four vowels /AA/, /AE/, /EH/ and /UW/. The entire spectrum of each vowel was separated into two sub-bands: 300-1000Hz and 1000-2500Hz, in accordance with the frequency locations of F1 and F2-F3. The spectrum above 2500Hz was not considered in these experiments.

Table 12: *The mean of first formant (F1) and second formant (F2) location of the /AA/ vowel from three astronauts voices during their mission for Apollo 11.*

	/AA/					
	Armstrong		Aldrin		Collins	
	F1	F2	F1	F2	F1	F2
Earth	641.7	1192.4	619.0	1231.3	632.2	1309.0
Travel	694.8	1254.8	746.3	1318.7	707.5	1333.3
Moon	691.2	1220.6	747.0	1308.1	N/A	N/A

Table 13: *The mean of first formant (F1) and second formant (F2) location of three vowels (/AE/, /EH/ and /UW/) for three astronauts voices during their mission for Apollo 11.*

	AE			EH			UW		
	Armstrong	Aldrin	Collins	Armstrong	Aldrin	Collin	Armstrong	Aldrin	Collins
Earth	530.7	559.9	553.1	513.2	500.3	521.3	398.6	379.5	402.3
Travel	656.1	726.0	675.0	633.2	657.1	684.5	534.6	522.2	567.4
Moon	642.5	728.9	N/A	597.8	631.0	N/A	502.3	532.7	N/A

Results shown in Fig. 29 represent the optimal frequency warping factor ά for four vowels examined in this experiment. A frequency warping factor ά which is less than one indicates that the spectrum in that sub-band was stretched toward higher frequencies, while greater than one indicates the compression of the spectrum towards lower frequencies. It can be seen that the first sub-band has been stretched toward higher frequencies in all four vowels examined. This result is also consistent with the results in Table 12 where the frequency locations of the same vowels are shifted toward upper frequencies. In terms of second sub-band, the vowel /AA/ showed optimal frequency warping factor less than one indicating the spread of spectrum toward upper frequencies. However, in other three vowels /AE/, /EH/ and /UW/, the optimal frequency warping factor are higher than one indicating the compression of spectrum toward lower frequencies.

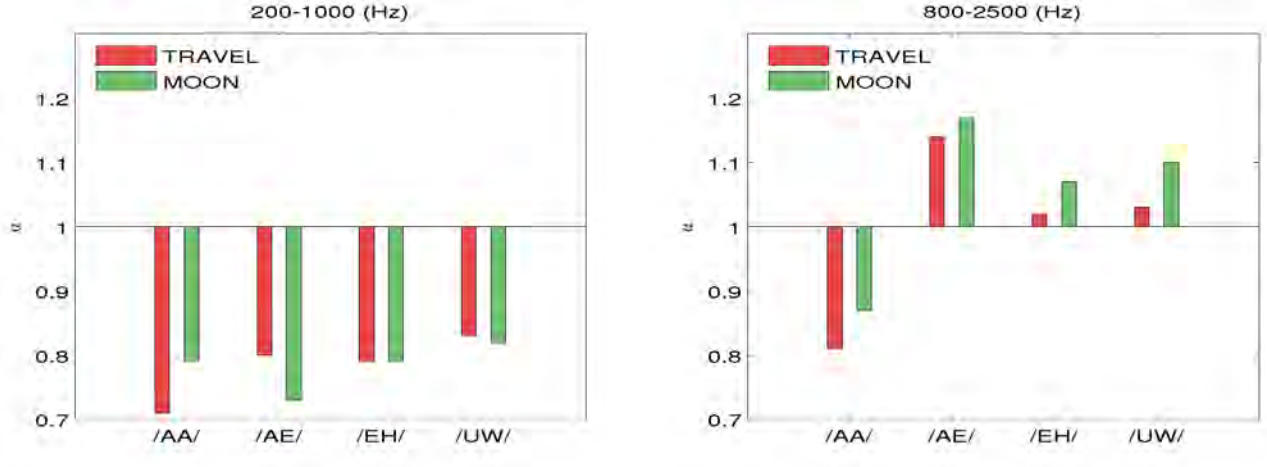


Figure 29: *Optimal frequency warping factor of four vowels /AA/, /AE/, /EH/ and /UW/.*

Additional Evaluation Results for Apollo Speaker Analysis and Tracking:

This effort also included developing an alternative distant metric for speaker diarization/tracking using the Fisher Linear Distance. In this new algorithm, instead of using a Bayesian information criterion (BIC), a Fisher Linear Discriminant was adopted as a distance metric for deciding whether two segments belong to the same speaker. The reason behind the use of Linear Discriminant as distance metric in diarization task is based on the assumption that when two segments comes from the same speaker they are less discriminative with each other. On the other hand, when two segments come from different speakers they are much easier to discriminate against each other. To use the Fisher Linear Discriminant as a distance metric between two segments, the dimension has to decrease to one as the limitation within all Linear Discriminant method. We evaluate the performance of this distance metric on Apollo 11 audio corpus. For initial evaluation of development distant metric, we did a simple cross comparison between segments belong to three astronauts. Each segment is 10ms. As the result is cross comparison distance, the three block along diagonal should have high emery (high similarity) if the distance works well. The result showed in Figure 30 is based on the BIC criterion, while the result showed in Figure 31 is based on proposed Fisher linear discriminant. By comparing the result visually, it can be seen that proposed Fisher linear discriminant performs reasonably. However, traditional BIC metric performs slightly better than proposed algorithm.

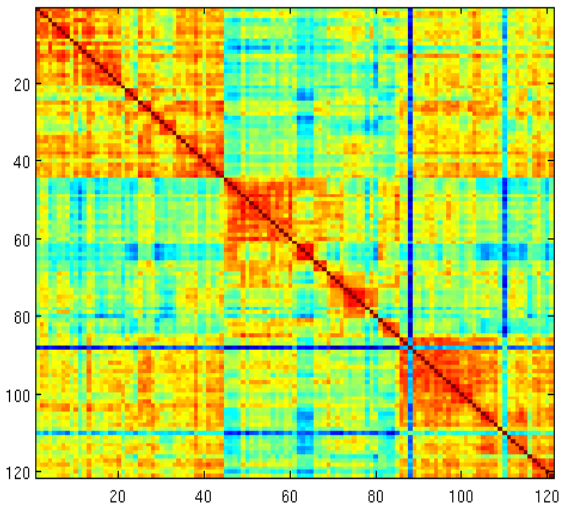


Figure 30: *Cross comparison between segments belongs to three astronaut using Bayesian information criterion (BIC). The segment index from 1-41 belongs to Armstrong, 42-82 belongs to Aldrin and 82-120 belongs to Collins. Higher energy indicates high similarity.*

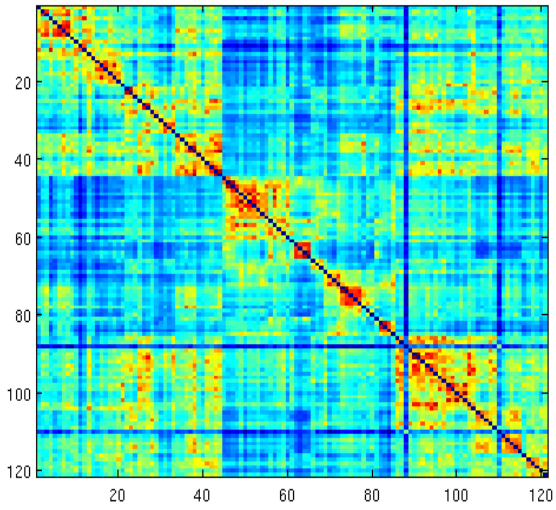


Figure 31: *Cross comparison between segments belongs to three astronaut using Fisher linear discriminant. The segment index from 1-41 belongs to Armstrong, 42-82 belongs to Aldrin and 82-120 belongs to Collins. Higher energy indicates high similarity.*

Task 1.9: Speaker ID for NIST SRE:

Front-End Processing for CRSS NIST-SRE: Based on previous CRSS-UTDallas ASR studies on increased vocal effort, noise, and reverberation, a series of front-end feature extraction techniques were proposed that provides increased robustness in the area of speaker recognition. A rectangular filter bank cepstral coefficients front-end, RFCC (see Figure 32), which utilizes a bank of rectangular non-overlapping subbands was found to be particularly successful using the NIST-SRE-2012 corpus evaluation, together with other subsystems utilizing QCN-RASTALP normalizations (see Table 14). These sub-system solutions have been delivered to USAF during bi-yearly progress site visits.

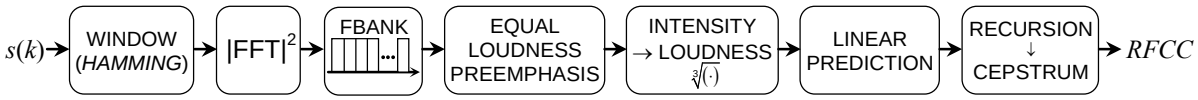


Figure 32. *Rectangular filter bank cepstral coefficients (RFCC) front-end.*

Table 14. *CRSS final NIST-SRE2012 submissions utilizing RFCC and QCN-RASTALP.*

#	Feature/VAD/Norm	Back-end	Male						Female					
			Dev			Eval			Dev			Eval		
			%EER	$C_{primary}$	$minC_{primary}$	%EER	$minC_{primary}$	$C_{primary}$	%EER	$minC_{primary}$	$C_{primary}$	%EER	$minC_{primary}$	$C_{primary}$
1	MHEC-VAD1-CMVN*	PLDA-1	1.358	0.160	0.203	1.934	0.199	0.265	2.496	0.241	0.240	2.448	0.260	0.294
2		GCDS	1.845	0.194	0.378	1.460	0.180	0.359	2.156	0.283	0.473	1.663	0.238	0.440
3		L2LR	2.761	0.248	1.377	2.206	0.223	1.163	3.884	0.338	1.366	2.590	0.255	1.032
4		UBS-SVM	1.905	0.173	0.381	1.460	0.161	0.356	2.293	0.261	0.477	1.719	0.215	0.440
5		PLDA-2	1.139	0.161	0.399	1.382	0.198	0.508	2.163	0.248	0.420	1.709	0.257	0.497
6	RFCC-VAD2-Warp	PLDA-1	1.325	0.183	0.193	1.883	0.218	0.243	2.353	0.249	0.235	2.283	0.259	0.267
7		GCDS	1.690	0.204	0.350	1.354	0.190	0.328	2.066	0.259	0.427	1.379	0.218	0.394
8		L2LR	2.365	0.238	1.274	1.810	0.207	1.288	3.286	0.317	1.367	1.955	0.229	1.189
9		UBS-SVM	1.753	0.188	0.360	1.423	0.188	0.336	2.200	0.239	0.430	1.442	0.200	0.399
10		PLDA-2	0.990	0.180	0.347	1.271	0.216	0.441	1.815	0.249	0.370	1.383	0.254	0.430
11	MFCC-VAD2-QCN-RASTA$_{LP}$	GCDS	1.684	0.210	0.396	1.422	0.190	0.376	2.225	0.296	0.496	1.869	0.249	0.466
12		UBS-SVM	1.749	0.193	0.395	1.428	0.173	0.372	2.424	0.265	0.499	1.929	0.229	0.469
13		PLDA-2	1.048	0.180	0.367	1.221	0.207	0.479	2.132	0.263	0.407	1.777	0.275	0.487
14	PMVDR-VAD1-CMVN	GCDS	1.842	0.199	0.380	1.394	0.179	0.356	2.076	0.258	0.449	1.480	0.216	0.416
15		PLDA-2	1.162	0.183	0.388	1.307	0.206	0.488	2.012	0.250	0.399	1.523	0.248	0.458

Task 2 – Open-Set Language ID (LID) / Dialect ID (DID):

In this task over the past three years, the language identification (LID) task has been addressed from two perspectives, close-set task and open-set task. As is known, robust language identification is typically hindered by factors such as the presence of background noise, channel mismatch, and speech data or duration mismatches. Today, acoustic and phonotactic models have been widely used for LID with some success. Phonotactic approaches usually are based on various phone recognizers and phoneme n-gram statistical analysis to extract discriminative information of each language of interest. The most popular phonotactic modeling techniques are Parallel Phone Recognition with Language Modeling (P-PRLM) and Phone Recognition-SVM (PR-SVM). However, phonotactic models usually perform well on relatively clean speech. In contrast, acoustic systems are usually based on some spectral features, which are followed by an effective analysis model. The state-of-the-art i-Vector approach, which has become a popular technique used for different verification and recognition tasks, can represent each conversation in parallel with a set of low-dimensional total variability factors and demonstrates session variation robustness. The study here focuses on both the close-set task and open-set task by means of phonotactic system and acoustic system.

This task is partitioned into sub-tasks, all dealing with both robustness issues as well as out-of-set LID rejection for language/dialect recognition/identification (LID/DID). Here, highlights of the research advancements over the past 36month period are highlighted. Further details can be found the following publications:

[1] Q. Zhang, J.H.L. Hansen, "Training Candidate Selection for Effective Out-of-Set Rejection in Open-Set Language Identification," submitted to *IEEE Trans. on Audio, Speech and Language Processing*, Sept. 2014

[2] Q. Zhang, J.H.L. Hansen, "Training Candidate Selection for Effective Rejection in Open-Set Language Identification," IEEE SLT-2014: Spoken Language Technology Workshop, paper PT3.4, Lake Tahoe, Dec. 7-10, 2014

[3] S. Amuda, H. Boril, A. Sangwan, J.H.L. Hansen, T.S. Ibiyemi, "Engineering analysis and recognition of Nigerian English: An insight into a low resource languages," Transactions on Machine Learning and Artificial Intelligence, 2(4), Aug. 2014, 115-126.

[4] Q. Zheng, G. Liu, J.H.L. Hansen, "Robust Language Recognition Based on Diverse Features," ISCA Odyssey-2014 Workshop on Speaker and Language Recognition, Joensuu, Finland, June 16-29, 2014

[5] D. Wang, J. Kates, J.H.L. Hansen, "Investigation of perceptual importance for temporal envelop and temporal fine structure between tonal and non-tonal languages," ISCA Interspeech-2014, Singapore, Sept. 14-18, 2014

[6] F. William, A. Sangwan, J.H.L. Hansen, "Automatic Accent Assessment Using Phonetic Mismatch and Human Perception," IEEE Trans. Audio, Speech & Lang. Proc., vol. 21(9), pp. 1818-1828, Sept. 2013

[7] H. Boril, Q. Zhang, P. Angkititrakul, J.H.L. Hansen, D. Xu, J. Gilkerson, J.A. Richards, "A Preliminary Study of Child Vocalization on a Parallel Corpus of US and Shanghainese Toddlers," ISCA INTERSPEECH-2013, pp. 2405-2409, Lyon, France, August 25-29, 2013

[8] Q. Zhang, H. Bořil, J.H.L. Hansen, "Supervector Pre-Processing for PRVM-Based Chinese and Arabic Dialect Identification," IEEE ICASSP-2013, Vancouver, Canada, May 26-31, 2013

[9] A. Sangwan, J.H.L. Hansen, "Automatic Analysis of Mandarin Accented English using Phonological Features," Speech Communication, vol. 54, no. 1, pp. 40-54, Jan. 2012

[10] H. Boril, A. Sangwan, J.H.L. Hansen, "Arabic Dialect Identification - 'Is the Secret in the Silence?' and Other Observations," ISCA Interspeech-2012, Mon-O1b-01, pg. 1-4, Portland, OR, Sept. 9-13, 2012

[11] Y. Lei, J.H.L. Hansen, "Dialect Classification via Text-independent Training and Testing for Arabic, Spanish and Chinese," IEEE Trans. Audio, Speech and Language Processing, vol. 19, no. 1, pp. 85-95, Jan. 2011

[12] M. Mehrabani, J.H.L. Hansen, "Language Identification: Analysis of Singing Speakers," IEEE ICASSP-2011, Prague, Czech Republic, May 22-27, 2011

2.1: Phonotactic system addressing closed-set task

2.1.1: Pre-processing variations based on SVM back-end: For closed-set language/dialect identification, variations to supervector pre-processing for phonotactic models (PRSVM) are explored in [Zhang13], which includes: (i) normalization of supervector dimensions in the pre-squashing stage, (ii) impact of alternative squashing functions, and (iii) N-gram selection for supervector dimensionality reduction. Also, a newly proposed dialect salience measure is applied in supervector dimension selection and compared to a common N-gram frequency based selection.

2.1.2: Exploration on NN/DBN backend: Since Deep Neural Networks (DNN) have become quite popular in current speech recognition research, it was decided to consider if such a backend could benefit a specific LID task. In this stage, system work concentrated only on a small LID dataset (Chinese dialect corpus). As is known, phonotactic dialect modeling utilizes phone recognizers and support vector machines (PRSVM) as an effective way of addressing the SID/LID task. Also, a previous study has already demonstrated that NN/DBN as the back-end, instead of an SVM, highlighted clear benefits in system performance according to phone recognizers developed by Brnu Univ., Czech.. As a further step, the combination of bottleneck feature followed by an SVM, as well as bottleneck feature followed by a Gaussian backend, was also investigated. Instead of only using bottleneck features, all of the hidden nodes were also used as a new feature for the final classifier. In addition, the system fusion of 9 different phone recognizers is also considered at this stage. Table 15 summarizes the core results from this study, with further details presented in [Ziang14b]. While the performance varied by about 4% absolute, there were clear benefits in setting aside the SVM solution, and employing bottleneck features with a Gaussian backend classifier.

Table 15: *Phonotactic LID system performance comparison based various back-ends*

Classifier	Fusion system performance (Accuracy)
SVM	85.3%
NN	87.5%
DBN	87.2%
Bottleneck feature + SVM	88.5%
Bottleneck feature + GB	**89.7%**
All hidden nodes + SVM	88.4%
All hidden nodes + GB	89.4%

Table 16 *USAF Chinese Language/Dialect corpus*

	Chinese			
Dialect	CMN	HSN	WU	YU
Train(Hrs)	6.3	8.9	5.1	7.7
Test(Hrs)	2.2	2.9	1.7	2.6
Avg. Dur.	10sec			

2.1.3: Acoustic system addressing closed-set task

Next, a continued effort also explored alternative features and feature processing. Therefore, based on a consistent acoustic model, this part of the study investigated a set of diverse acoustic features, partitioned into three broad classes: (i) classical, (ii) innovative, and (iii) extensional features as shown in Table 17, using a range of back-end classifiers derived from the CRSS-UTDallas MS-AcID toolkit [Liu14]. These included Gaussian backend, Gaussianized cosine distance scoring (GCDS), and system fusion for close-set LID. In addition, the proposed strategy was tested under both highly noisy communication channel conditions (DARPA RATS), and large-scale dataset (NIST LRE09), which is shown in Table 18.

Table 17: *LID Feature Configuration*
('Default' implies original settings for that feature by their authors).

Feature	Special Configuration	Coefficients (Dim)
Mel-frequency cepstral coefficients (MFCC)	# of Channels = 26	12
Perceptual linear predictive (PLP)	# of Channels = 20	13
Linear frequency cepstral coefficients (LFCC)	# of filter banks = 32	20
Gammatone frequency cepstral coefficients (GFCC)	Default	12
Power normalized cepstral coefficients (PNCC)	Default	12
Perceptual minimum variance distortionless response (PMVDR)	Default	12
RASTA-PLP	Default	9
RASTA-LFCC	Default	20
Multi-peak MFCC	K=8	13
Thomson MFCC	K=8	13
Sine-weighted cepstrum estimator (SWCE) MFCC	K=8	13

Table 18: *Corpus statistics for the DARPA RATS and NIST LRE09.*

Corpus	DARPA RATS		NIST LRE09(Inset)	
Data set	TRAIN	TEST	TRAIN	TEST
Count	12035	877	11158	31178
Avg. Duration(*sec./file*)	58.3	18.0	39.3	12.4
SNR(dB)	5.9	8.0	23.30	23.8

System fusion usually significant benefits the overall system performance because of the complementary effect among each individual session. Therefore we proposed to leverage the impact of both the front-end and back-end fusion. The front-end fusion is implemented by concatenating the i-Vectors from different raw features. The back-end fusion is implemented by utilizing the FoCal multi-class toolkit. By measuring the goodness of each recognizer and assigning a proper weight based on the supervised development data, FoCal (linear logistic regression fusion) provides a calibrated fusion of the scores of multiple recognizers. All the performance are shown in Table 19.

Table 19: *Performance on RATS and LRE09 database (C_{avg}*100).*

Feature category	Feature type	RATS			LRE09		
		GB	GCDS	Backend fusion	GB	GCDS	Backend Fusion
Classical features	MFCC	15.6	14.9	12.8	15.8	14.0	11.5
	LFCC	16.5	16.0	15.0	16.6	15.1	12.2
	PLP	18.7	17.6	16.5	18.7	16.0	13.9
	GFCC	14.9	14.5	13.1	16.6	14.8	12.0
Innovative features	PNCC	14.4	14.0	11.6	16.0	15.1	12.2
	PMVDR	19.1	19.0	16.7	17.8	15.8	13.5
Extensional features	RASTA-LFCC	15.2	13.7	11.3	16.7	14.1	11.3
	RASTA-PLP	14.1	14.2	12.7	15.6	13.3	10.5
	Multi-peak MFCC	15.5	14.0	12.8	15.5	13.7	10.8
	Thomson MFCC	16.0	13.7	12.1	15.7	13.7	11.1
	SWCE MFCC	15.3	13.5	11.5	15.5	13.7	10.9
Feature concatenation		13.1	12.4	9.6	11.7	11.3	8.5

In this study, a series of front-ends and back-ends were systematically investigated, which demonstrated that by properly fusing various types of acoustic features and back-end classifiers, performance can be improved significantly. In addition, the latest proposed GCDS back-end outperforms a generative Gaussian back-end. To be more specific, for the DAPRA RATS scenario, hybrid fusion benefits average cost

function C_{avg} with a relative +38.5% improvement. For the NIST LRE09 relatively clean scenario, the performance of whole utterance achieved a +46.2% relative improvement. Through the feature relative significant factor analysis, the interesting phenomenon is PASTA_PLP are the most significant feature across two corpora. These observations offer useful practices for other practitioners in the LID field. More detail could be found in [Zhang14a].

2.2: Acoustic System addressing Open-set task: candidate selection

In the real scenario, open-set task is more general, where training data might not cover all possible test languages. Abundant data collection is one of the most direct and effective ways to address problem with more unknown/out-of-set (OOS) language characteristic exploration. Our study also focuses on effective candidate selection methods for universal OOS language coverage in [Zhang14b] through a language relationship analysis. The state-of-the-art i-vector system followed by a generative Gaussian back-end achieves effective performance for LID. The reason is that i-vector contains sufficient language dependent information, and the following generative classifier could model the feature distribution precisely with abundant data. According to the generative modeling theory, we continue proposed three effective and flexible candidate selection methods.

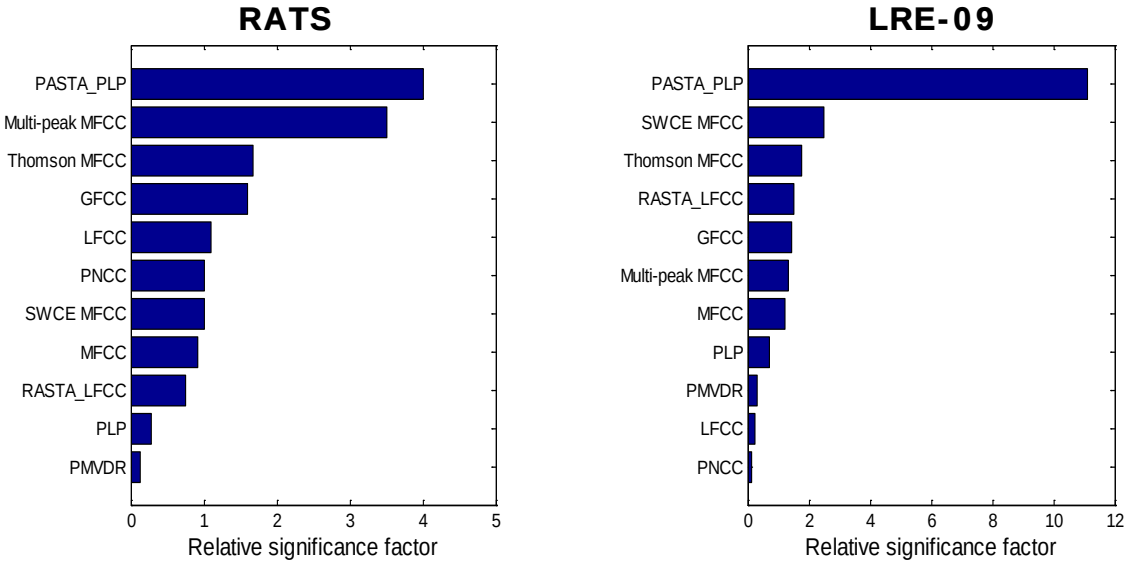


Figure 33: *Relative significance factor analysis across two corpora.*

2.2.1: Candidates: best k cluster representatives

K-means clustering is a popular strategy in data mining which aims at partitioning n observations into *k* clusters so that each observation belongs to the cluster with the nearest mean. In addition, these centroids are assumed to be the best representatives of the entire category if only using *k* points for distribution modeling. To be more specific, we calculate the Euclidean distance between the mean feature vector of each language and the centroid of each cluster. Then the closest languages according to each centroid are selected as our proposed candidates for OOS modeling. Fig 1 shows the distribution of all 40 languages in a 2-D feature space by means of the mean vectors, where the red crosses represent inset languages, the unselected OOS languages are black circles, and the rest colorful asterisks refer to the selected OOS candidates based on the k-means clustering algorithm (k=5).

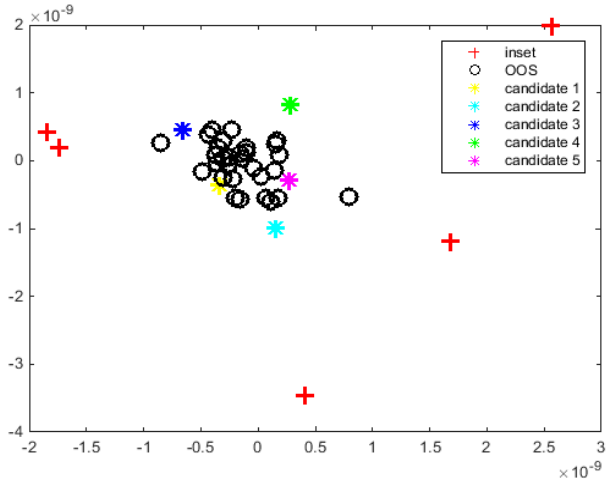


***Figure 34**: K-means clustering based OOS candidates (k=5)*

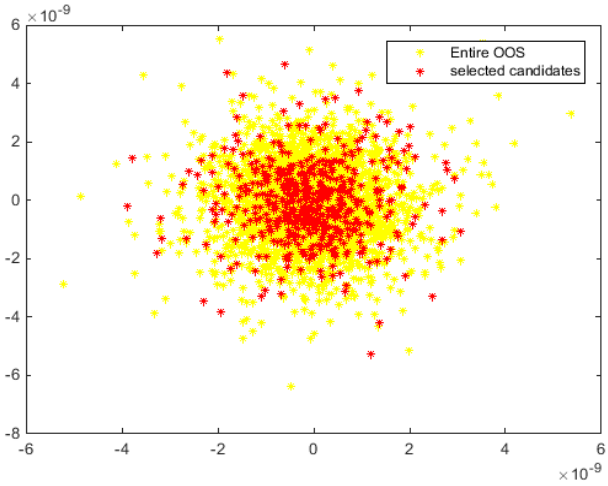


***Figure 35**: Best k complementary OOS candidate selection (k=5).*

2.2.2: Candidates: best k complementary representatives

Gaussian back-end as a generative classifier is used for capture the distribution characteristic of each category. Since the goal is find a subset of training samples which could model the entire OOS with minimum lost, the combination of effective selected candidates should share similar distribution characteristics with entire OOS. In probability theory and information theory, the Kullback–Leibler divergence is a non-symmetric measure of the difference between two probability distributions P and Q. Therefore, a combination of k complementary candidates can be selected with a small KL-divergence from entire OOS. Fig. 35 shows the feature scatters or distribution comparison between k complementary candidates (proposed candidates) and all 40 OOS languages (baseline) in a 2-D feature space.

2.2.3: Candidates: best k general representatives

Alternatively, the OOS language generality in terms of in-set languages can be investigated on back-end level. More specifically, through generative Gaussian back-end processing, each in-set language possesses a corresponding individual model, and one general model is used for evaluating all OOS languages. Therefore, every test utterance is assigned 6 scores (score vector) according to each back-end model. For simplicity, only the average score vector is employed as the new feature for each language. Ideally, the feature would be a base vector, because there is a high probability which is only derived from that corresponding model. However, any overlap between different languages attenuates the importance of this absolute value. Therefore, instead of a single score, the score patterns are adopted for distance calibration. A engineering language tree is generated by hierarchical clustering based on the cosine distance across these score vectors. However, the hierarchical tree can only provide a general view of the relationship between the languages. Quantizing the distance between each OOS language to the entire in-set group of languages would provide more precise information for the distance based OOS candidate selection. Finally, a back-end score vector based OOS confusability rank is generated according to the total distance values. The most k general representatives can be selected with the descending order of OOS confusability rank, because the candidate with higher rank represents less confusability versus all the in-set languages.

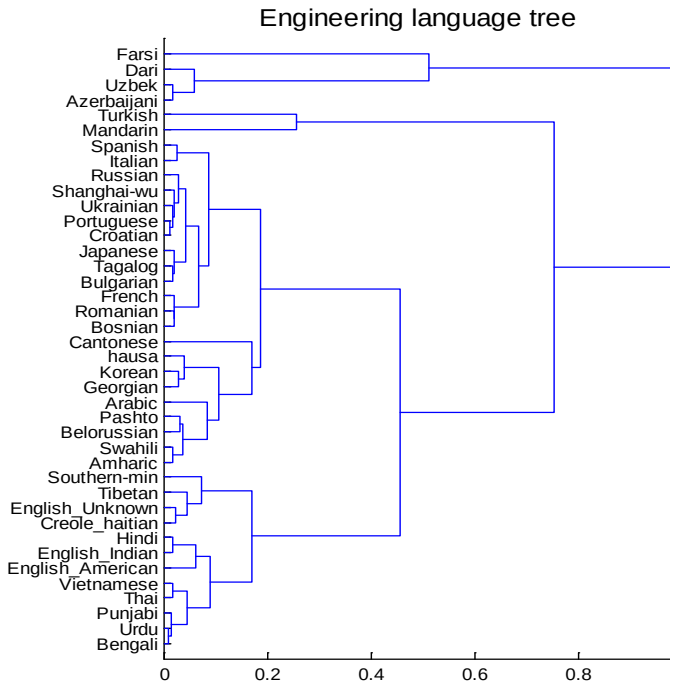


Figure 36: *Acoustic feature based language tree.*

As noted earlier, three effective OOS candidate selection methods are proposed to address open-set LID. Fig. 37 shows us the overall classification performance comparison of three proposed selection methods along with different value for the parameter *k* (i.e., number of choosing OOS candidates). In addition, the performance of OOS rejection and overall classification share a similar trend. A baseline system was realized using all available OOS languages for modeling. Base on different proposed selection method, an optimal minimum number of OOS candidate was found out. More specific, it can be noted that the performance using only 4 candidates for OOS modeling is better than that with more candidates (i.e., 5 or even 10 candidates) based on k-means clustering. While using complementary method and along with 91% data reduction compared with close-set baseline, the performance only dropped by 14.6% and 2% according to EER for OOS rejection and average cost function for overall classification, respectively. In addition, compared with a dozen of random selections with the same amount of OOS languages, the proposed method outperforms random selection consistently.

2.2.4: References for this Sub-Task

[Zhang14a] Q. Zhang, J.H.L. Hansen, "Training candidate selection for effective rejection in open-set language identification," Spoken Language Technology Workshop (SLT), Dec. 7-10, South Lake Tahoe, NV, USA, 2014.

[Zhang14b] Q. Zhang, G. Liu, J.H.L. Hansen, "Robust language recognition based on diverse features," Odyssey 2014: The Speaker and Language Recognition Workshop, 2014, Joensuu, Finland, pp. 152-157, 2014.

[Zhang13] Zhang, Q., Boril, H., Hansen, J. H. L. (2013). "Supervector Pre-Processing for PRSVM-based Chinese and Arabic Dialect Identification," IEEE ICASSP'13, 7363-7367, Vancouver, Canada, May 2013.

[Liu14a] G. Liu, J.H.L. Hansen, "An Investigation into Back-End Advancements for Speaker Recognition in Multi-Session and Noisy Enrollment Scenarios," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1978-1992, Dec. 2014

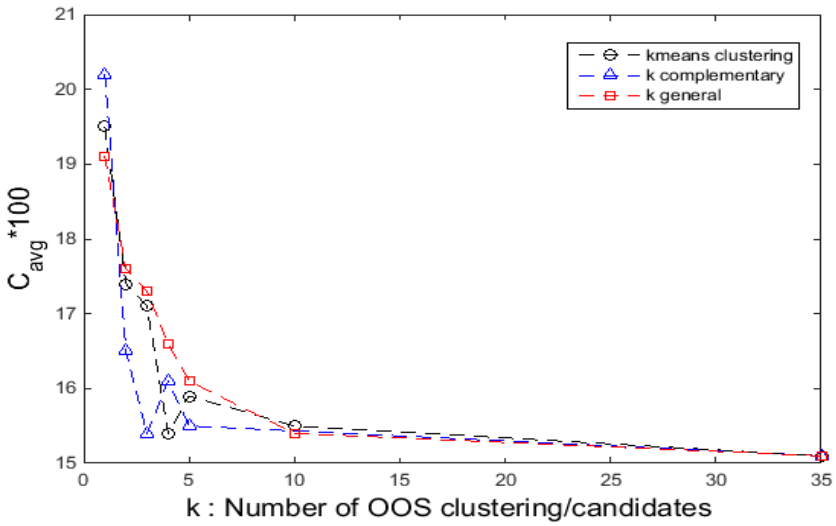


Figure 37: *Overall classification performance (Cavg) comparison based on three proposed candidate selection methods.*

2.3: UTDallas-CRSS MS-AcID Toolkit Development: Over the project period, there was a significant effort to develop an integrated toolkit which would address mismatch issues relating to Speaker ID (SID) as well as Language ID (LID). The resulting effort has produced the MS-AcID toolkit, which has been delivered to USAF. This includes (i) features, (ii) feature processing, (iii) backend systems, and (iv) score fusion methods. With respect to LID, the sub-task tools include a Data Purification Selection (i.e., screening potential input language ID training data for acoustic mismatch, music, noise, distortion, overlap speech, program re-transmit, etc.).

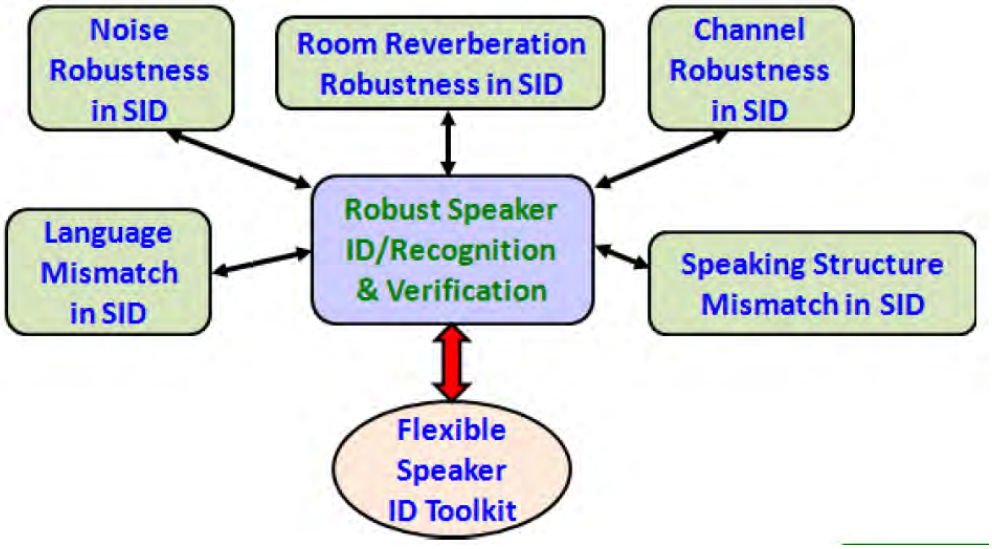


Figure 38: *Robustness in SID: as part of TASK#1, a flexible took for SID would be developed. This framework was generalized to support SID, LID, and DID.*

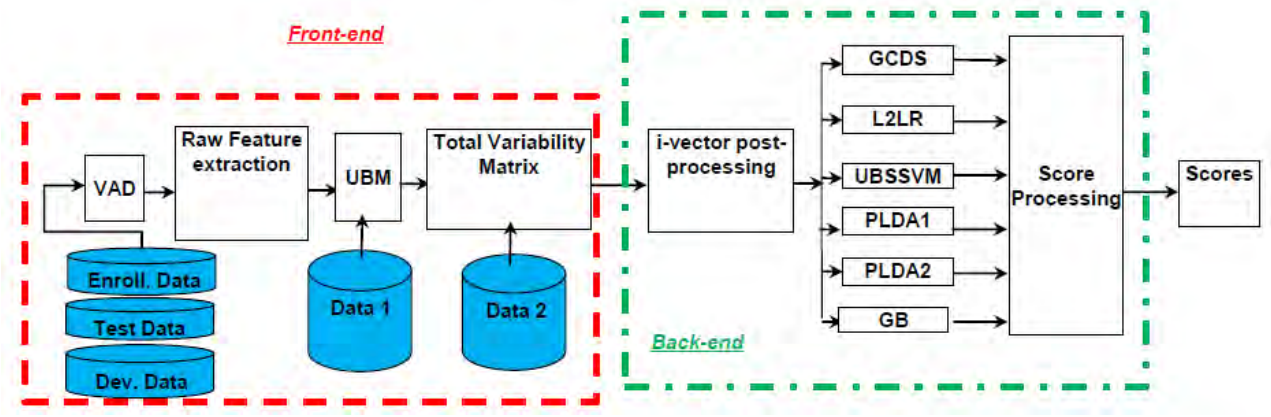


Figure 39: *MS-AcID Toolkit: Multi-Session Acoustic Identification Toolkit – for speaker ID and language/dialect ID.*

With respect to the specific advancements for LID/DID using MS-AcID, the following has been accomplished:

2.3.1: LRE data purification:

A research investigation dealing with data purification for LID training/testing using the NIST LRE-2013 corpus was completed. The following manuscript was submitted which details the algorithm formulation, evaluations, and impact of the data purification strategy. Table 20 summarizes the key results from that study. A Gaussian backend produced LID EER of 13.28%, while the corresponding data purification solution with SVM and Gaussian backend was able to reduce the EER to 9.94%. Further details can be found in this manuscript.

- G. Liu, J.H.L. Hansen, "Big Data Purification for Acoustic Model Advancements in Language Recognition," submitted to *IEEE Trans. on Audio, Speech and Language Processing,* Nov. 2014

Table 20: *System performance improvement flow from baseline system to the final grand system.*

System	GB baseline	GB CRSS	SVM + GB CRSS	SVM + GB CRSS
Cavg(%)	13.28	11.61	10.24	9.94

2.3.2: Audio+Text based LID:

A previous probe study which leveraged text and audio content for dialect ID of English family-tree dialects (i.e., American English:AE, United Kingdom:UK, and Australian: AU) was completed. This study used audio content captured from local podcast websites in these three countries. The podcast corpus was collected over a year, with corresponding transcripts from website materials used for text based dialect ID processing. An initial journal submission of this audio plus text DID classification effort focused on the UT-Podcast corpus with GMM-UBM based DID solution for the audio portion. Over the past 36 month period, a parallel solution using an i-Vector framework along with additional evaluations using the NIST LRE-2013 corpus was completed. Also, a number of new metrics were formed which leveraged the differences between audio/text processing for DID were introduced. A high level summary of results are shown in Table 21. A more complete discussion is included in the following submitted manuscript.

- J.H.L. Hansen, G. Liu, "Unsupervised Accent/Dialect Classification for Deep Data Fusion of Acoustic and Language Information," submitted to Speech Communication, Nov. 2014.

Table 21: *Performance comparison of different Dialect Identification approaches on UT-Podcast. AU, UK and US correspond to Australian English, United Kingdom English, and American English, respectively. 'UAR' is the unweighted average recall.*

System #	Classification System	Recall(%)			UAR(%)	Cavg(%)
		AU	UK	US		
0	Audio System (Baseline: GMM)	85.5	32.6	62.9	60.3	29.7
1	Audio System (i-Vector)	78.0	61.8	83.8	74.5	19.1
2	Text System(N-Gram Perplexity)	22.6	53.9	30.4	35.6	54.0
3	Text System(LSA Cosine Distance Scoring)	55.4	48.3	51.7	51.8	36.2
4	Text System(TF-IDF Logistic Regression)	83.1	32.6	60.4	58.7	31.0
5	Proposed Audio-Text system (Fusion of Sys 1&4)	86.1	60.7	82.1	76.3	17.8

References for this section:

[1] G. Liu, J.H.L. Hansen, "Big Data Purification for Acoustic Model Advancements in Language Recognition," submitted to IEEE Trans. on Audio, Speech and Language Processing, Nov. 2014

[2] J.H.L. Hansen, G. Liu, "Unsupervised Accent Classification for Deep Data Fusion of Acoustic and Language Information" submitted to Speech Communication, Nov. 2014.

2.4: Language Identification using the South Indian Language Corpus:

In this area, an additional effort was undertaken to consider again the LID for the South Indian Corpus. The motivation was to establish an effective LID framework for which to formulate more robust noisy LID performance. This study focused on the USAF South Indian Corpus.

- The initial experiment was to first perform an estimation of SNR for the conversational data (lapel microphone), for all five languages, using the NIST STNR algorithm. The SNR measurements, and the corresponding standard deviations are indicated in Table 22. The results here show that for the 5-way language space, there is a background noise dependency, which would imply some dependency on the background noise acoustics between the corresponding languages.
- Next, a series of three-way cross validation Language Identification experiments using i-Vectors were performed on the lapel-microphone conversational data. In each of the three splits, two portions of the data were used for training and development, while the third portion was used for testing. The results are shown in Table 23, for GCDS and PLDA back-ends. The Gaussian Cosine Distance Scoring (GCDS) backend performed worse than the Probabilistic Linear Discriminative Analysis (PLDA) backend. There is also some degree of variability in the 5-way cross validation scheme – suggesting that the data sets have some acoustic variability across the data. The results show that prior acoustic analysis of the training data is needed to ensure consistency in the background acoustics, so that the resulting classifiers are in fact focused exclusively on the language differences. Second, that leveraging more than one back-end would potentially offer benefits when the classifier strategies are different but complementary.

Table 22: *SNR, and the corresponding standard-deviation (in SNR) of the 5 languages from the South Indian Languages Corpora, evaluated using the NIST STNR algorithm.*

Language	**SNR (dB)**	**STD-SNR (dB)**
Kannada	32.1566	4.5266
Malayalam	38.7598	5.6137
Marathi	38.2854	3.6355
Tamil	30.1450	5.2551
Telugu	30.9747	6.4472

Table 23: *Results of 3-way cross validation experiments on the South Indian Languages Corpora test-utterances using GCDS and PLDA back-ends.*

Split	EER (GCDS)	EER (PLDA)	GCDS Target-Class Accuracy (%) (Kan, Mal, Mar, Tam, Tel)	PLDA Target-Class Accuracy (%) (Kan, Mal, Mar, Tam, Tel)
1	1.35	1.30	100, 94, 100, 100, 97	100, 97, 100, 100, 97
2	4.31	3.36	97, 88, 100,100, 82	97, 88, 100, 97, 82
3	2.67	1.97	97, 91, 100, 100, 88	91, 94, 100, 100, 88
Avg.	2.77	2.21	98, 91, 100 ,100, 89	96, 93, 100, 99, 89

2.5: Language Identification: Is the Secret In the Silence?

From previous research in the domain of dialect ID using the USAF Pan-Arabic dialect corpus, CRSS wanted to explore the trade-offs in performance between the system developed for DID [1,2]. Studies published during the same time CRSS-UTDallas was exploring dialect ID on Arabic dialects reported unrealistically high classification accuracies. The study by Biadsy, Hirschberg, and Habash [5] in 2009 reported a 4-way dialect ID performance of 81.60% using 30s test files from four Arabic corpora from LDC. A 5-way detection performance for the PAN-Arabic corpus by Lei and Hansen [2] in 2008 achieved 71.7% accuracy. Subsequently, Lei and Hansen extended that work [3] in 2009 and achieved 14.09% EER with a 100 factor baseline IIFA (their proposed Information Integration Factor Analysis) solution, which they improved to 12.77% using additional gender and speaker information knowledge. However, the fact remained that the study [5] achieved a 10% greater classification accuracy – where they were also using additional suprasegmental knowledge. Since that study used four independently collected Arabic corpora from LDC in four different dialects, CRSS-UTDallas wanted to explore if there were other factors that lead to the greater classification performance. The focus here were on the following four dialects: Gulf Arabic, Iraqi Arabic, Egyptian Arabic, and Levantine Arabic.

2.5.1: Dialect ID for PAN-Arabic Corpus: Conversational telephone speech (CTS) collections of Arabic dialects distributed through the Linguistic Data Consortium (LDC) provide an invaluable resource for the development of robust speech systems including speaker and speech recognition, translation, spoken dialogue modeling, and information summarization. They are frequently relied on also in language (LID) and dialect identification (DID) evaluations. The first part of this study attempts to identify the source of the relatively high DID performance on LDC's Arabic CTS corpora seen in recent literature. It is found that recordings of each dialect exhibit unique channel and noise characteristics and that silence regions are sufficient for performing reasonably accurate DID. The second part focuses on phonotactic dialect modeling that utilizes phone recognizers and support vector machines (PRSVM). A simple *N*-gram normalization of PRSVM input supervectors utilizing hard limiting is introduced and shown to outperform the standard approach used in current LID and DID systems.

2.5.1: Is the Secret in the Silence?: In the first part of this study, channel and noise characteristics of selected LDC Arabic dialect CTS corpora are analyzed and found to be unique and fairly distinctive for each dialect corpus. As a consequence, silence segments are found to carry sufficient information to perform a reasonably accurate DID. It is also demonstrated that performing channel normalization may, to a large extent, help equalize channel differences between the dialect databases, but this is not sufficient in addressing other, most likely noise-related non-speech 'dialect cues' present in the recordings. In the second part of this study, a normalization of PRSVM input supervectors is proposed and evaluated alongside with the standard normalization on the LDC corpora as well as on an in-house Pan-Arabic corpus.

In this study, two sets of Arabic dialect data are employed. The first is represented by LDC's Arabic conversational telephone speech corpora: Iraqi Arabic CTS (IRQ), Gulf Arabic CTS (GLF), Arabic CTS

Levantine Fisher Training Data Set 3 (LEV), and CALLHOME and CALLFRIEND Egyptian Arabic Speech (EGY). It is noted that the studies [5-7] used Levantine Arabic CTS (LDC2007S01) instead of the Fisher corpus. While this represents a difference between our study and the past literature, it is assumed that the observations made do apply also to the mentioned studies as 3 out of 4 dialect sets are overlapping.

2.5.2: Is the Secret in the Silence? Classifier Performance: In the preliminary experiments involving a naive maximum likelihood GMM classifier, separate 32-mixture GMMs were trained for each of the 4 dialects captured in the LDC data set. The number of training chunks was as follows: Iraqi (5075), Gulf (10526), Levantine (3771), and Egyptian (10628). The confusion matrix for an in-set dialect identification (pick 1-out-of-4) on the level of individual speech chunks is shown in Table 24(i). It can be seen that in spite of the simplicity of the system, the initial performance on short speech chunks is relatively high compared to the chance performance (25%). In order to verify to what extent can the DID performance be attributed to the linguistic content present in the recordings, another experiment, where pure silence chunks were used both for GMM training and evaluation, was conducted. As can be seen in Table 24(ii), the average dialect classification accuracy increased from 82.0% seen for speech chunks, to 83.3% using silence chunks. This suggests that the silence regions themselves carry sufficient information for identifying the database origin and, in the case of the simplistic GMM classifier, presence of speech is actually not helpful to the task.

Table 24: *GMM-based DID on (i) speech chunks, and (ii) silence chunks*

	Assigned Dialect (Speech Chunks)				Acc (%) Avg 82.0
Ground Truth	Gulf	Iraqi	Levantine	Egyptian	
Gulf	**510**	120	4	1	80.3
Iraqi	184	**527**	1	2	73.8
Levantine	120	10	**370**	0	74.0
Egyptian	8	0	0	**3174**	99.7

	Assigned Dialect (Silence Chunks)				Acc (%) Avg 83.3
Ground Truth	Gulf	Iraqi	Levantine	Egyptian	
Gulf	**260**	78	0	0	76.9
Iraqi	96	**228**	0	0	70.4
Levantine	24	1	**158**	1	85.9
Egyptian	0	0	0	**1973**	100

2.5.3: Is the Secret in the Silence? Long-term Transfer Functions & Compensation: For each database, the long-term channel transfer function was estimated by averaging short-term log-amplitude spectra of silence segments in the recordings. Average transfer function estimates for the four dialect recordings are shown in Fig. 40(i). It can be seen that the channel characteristics are very consistent across each dialect recordings and fairly distinctive between dialects. In order to equalize database channel differences, a normalization procedure was implemented. While the normalization could be conveniently applied directly on the cepstral coefficients, the goal here was to reconstruct the normalized time-domain speech samples that could be later processed by any DID scheme of choice. The estimated average transfer functions of the normalized silence segments are shown in Fig. 40(ii). Since the long-term spectra are now closer, it is expected that performance should improve and be more dependent on the dialects. However, it is noted that while the long term average is the same, there are a variety of other factors such as variability of the channel/noise/microphone/recording conditions in addition to the long-term average which can introduce recording dependency in the corpus (which was actually noted in the evaluations).

For a comparison, an identical GMM-based classification procedure was repeated also for silence segments from the in-house Pan-Arabic corpus. A classification accuracy in a four-way task on PS, SY, IRQ, and PS silence chunks (AE was omitted to mimic the complexity of the LDC task) yielded a slightly below chance (24.7%) accuracy. This suggests that the acoustic characteristics of the silence segments here are much more consistent across the dialects.

Table 25 details EERs per each dialect for English- and Hindi-based PRSVMs, repsectively. The *hard limit* normalization is found to reduce EERs for all dialects compared to the original 'GBF+Log' setup. It can be seen that Egyptian samples are classified in the LDC task with a significantly lower error rate than the other dialects. In the case of the in-house Pan-Arabic task, EERs are well balanced, with the Iraqi dialect yielding slightly lower errors compared to the rest three dialects. Finally, it can be seen that the absence of non-linguistic cues in the in-house Pan-Arabic corpus results in considerable reduction of the classifier performance compared to the LDC task.

The results here suggest that care must be exercised in exploring speech and language classification tasks such as Dialect ID, especially when corpora are collected under different acoustic environments. Further research in this domain is not only needed, but critical for real-world advancements in the speech and language processing domains. Further details are presented in [8].

2.5.4: Is the Secret in the Silence? References:

[1] Y. Lei, J.H.L. Hansen, "Dialect Classification via Text-independent Training and Testing for Arabic, Spanish and Chinese," IEEE Trans. Audio, Speech and Language Processing, vol. 19, no. 1, pp. 85-95, Jan. 2011.

[2] Q. Zhang, H. Boril, J.H.L. Hansen, "Supervector Pre-Processing for PRVM-Based Chinese and Arabic Dialect Identification," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 7363-7367, (paper #5548), Vancouver, Canada, May 26-31, 2013

[3] Y. Lei, J.H.L. Hansen, "Factor Analysis-Based Information Integration for Arabic Dialect Identification," IEEE ICASSP-2009: Inter. Conf. Acoustics, Speech, and Signal Processing, pp. xxx-xxx, Taipai, Taiwan, 2009

[4] Y. Lei, J.H.L. Hansen, "Dialect Classification via Discriminative Training," ISCA INTERSPEECH-2008, pp. 735-738, Brisbane, Australia, Sept. 2008

[5] F. Biadsy, J. Hirschberg, N. Habash, "Spoken Arabic Dialect Identification Using Phonotactic Modeling," EACL 2009 Workshop on Computational Approaches to Semitic Languages, pages 53–61, Athens, Greece, 31 March, 2009.

[6] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," ISCA INTERSPEECH-11, Florence, Italy, 2011, pp. 745–748.

[7] M. Akbacak, D. Vergyri, A. Stolcke, N. Scheffer, , and A. Mandal, "Effective Arabic dialect classification using diverse phonotactic models," INTERSPEECH'11, Florence, Italy, 2011.

[8] H. Boril, A. Sangwan, J.H.L. Hansen, "Arabic Dialect Identification - 'Is the Secret in the Silence?' and Other Observations," ISCA Interspeech-2012, Mon-O1b-01, pg. 1-4,, Portland, OR, Sept. 9-13, 2012

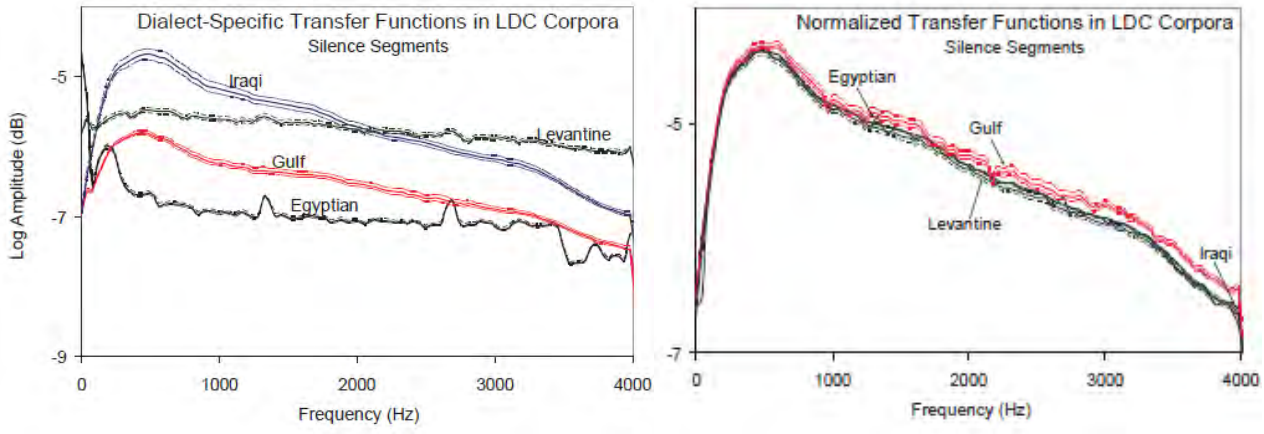


Figure 40: (i) *Arabic dialect-specific channel characteristics in LDC corpora estimated as long-term averages of log amplitude spectra in silence segments. Dashed lines – intervals of ±5σ.*
(ii) *Normalized channel characteristics estimated from silence segments. Dashed lines – intervals of ±5σ (depicted for clarity only for Gulf and Levantine).*

Table 25: *Detailed dialect EERs (1-vs-3 task) of English- and Hindi based PRSVM DID systems*

Local Bigram Frequency (*LBF*) Normalization	1-vs-3; EER (%)							
	LDC Corpora (Eng. Recognizer)				In-House Pan-Arabic (Hindi Recognizer)			
	GLF	IRQ	LEV	EGY	PS	IRQ	SY	EGY
GBF + Log	10.3	8.2	9.5	2.8	31.8	28.5	33.3	33.0
GBF + Sigmoid	9.5	7.8	8.5	2.6	29.9	26.2	30.5	31.1
Sigmoid	9.6	7.9	8.1	2.5	31.3	27.0	31.7	31.0
Hardlimit	**9.5**	**7.5**	**8.0**	**2.3**	**30.1**	**26.3**	**30.3**	**30.7**

2.6: Further Advancements in Normalization for LID/DID

2.6.1: Proposed Pre-Squashing Normalizations for PRSVM LID/DID

- **Within-dimension mean/variance norm (WD MVN)**: for each N-gram, mean and variance of its relative frequency (no squashing) is estimated from training tokens across all classes. During the supervector extraction for SVM modeling and classification, the stored `train' means and variances are used for dimension-wise MVN. This can be viewed as alternative or complementary norm to the traditional global frequency normalization (GFN) by the inverse square root of N-gram prior probabilities estimated from the training data, only here rather the means and variance of across-class priors are equalized.
- **Across-dimension mean/variance norm (AD MVN)**: mean and variance are estimated across the elements of an individual supervector, and are applied in MVN of the supervector elements. AD MVN normalizes the frequency profile seen across the supervector dimensions. In AD MVN, the variance normalization is paired with a multiplicative constant α to control dynamics of the normalized frequencies. When combined with a nonlinear squashing function, α can be effectively used to shape contours of the frequency distributions.
  - $NNF_{t,m}^{ADMVN} = \left(NNF_{t,m} - \overline{NNF_t}\right) \Big/ \sqrt{\alpha \operatorname{var}\left(NNF_t\right)}$
- **GFN with uniform priors**: instead of global N-gram frequencies extracted from the training set (*train priors*), uniform N-gram priors are substituted in the square root term. This uniform amplitude norm is introduced to allow for `switching-off' the traditional GFN and yet numerically accommodate the log squashing function.

2.6.2: Proposed Distribution Tail Compression with AD MVN and Sigmoid Squashing Function

- Sigmoid squashing function is introduced $g(x) = \frac{1}{1+e^{-x}}$
- Combined with AD MVN, α can be used to control the variance of relative frequencies and together with sigmoid also their distributions (see Figure 41). Expanding the variance will push the distribution tails into the saturated regions in the sigmoid, resulting in the compression of the tails. This may help equalize the impact of N-gram outliers (extremely frequent or extremely rare) on the subsequent SVM modeling.

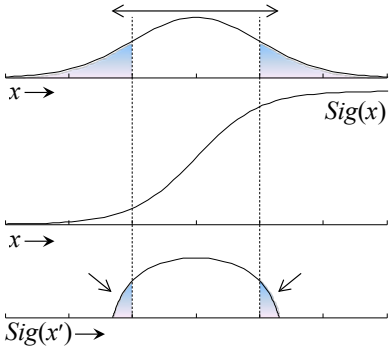


Figure 41: *Transformation of N-gram frequency distribution by sigmoid squashing function. The rate of expansion or compression of distribution tails can be controlled through α in AD MVN.*

- N-gram Frequency- vs. Dialect Salience-Based Supervector Reduction:
  - Past studies deal with the high dimensionality of supervectors in phonotactic-based PRSVM DID by preserving only the most frequent N-grams as the supervector dimensions and discarding the rest. This approach inherently assumes that the most frequent N-grams carry the cue to a successful DID. While frequent N-grams are a favorable choice, as they are more likely to appear in the test utterances than rare N-grams, it is not clear whether we could not select a another subset of N-grams (perhaps still very frequent ones) that would be more dialect-salient. This lead to the proposal of the dialect salience dimension reduction [Zhang13] where dimensions strongly populated by one language/dialect but sparsely represented in other languages/dialects are selected.
  - Performance of both the conventional frequency based supervector reduction and dialect salience based reduction are compared in Figure 42.
  - The effort in this area was to focus on the question concerning how big is the overlap between the most frequent and the most dialect-salient N-grams.
  - Both dimensionality reduction approaches operate on N-gram frequencies. The frequency based approach ranks N-grams by their overall prior probabilities extracted *across all dialects*. The salience based approach utilizes N-gram priors found *within individual dialects* and ranks N-grams whose occurrence is most non-uniform across dialects (i.e., N-grams that occur frequently in some dialects and are very rare in others). Clearly, the criteria chosen by the two methods are not mutually exclusive and may lead to the same or similar N-gram subsets.

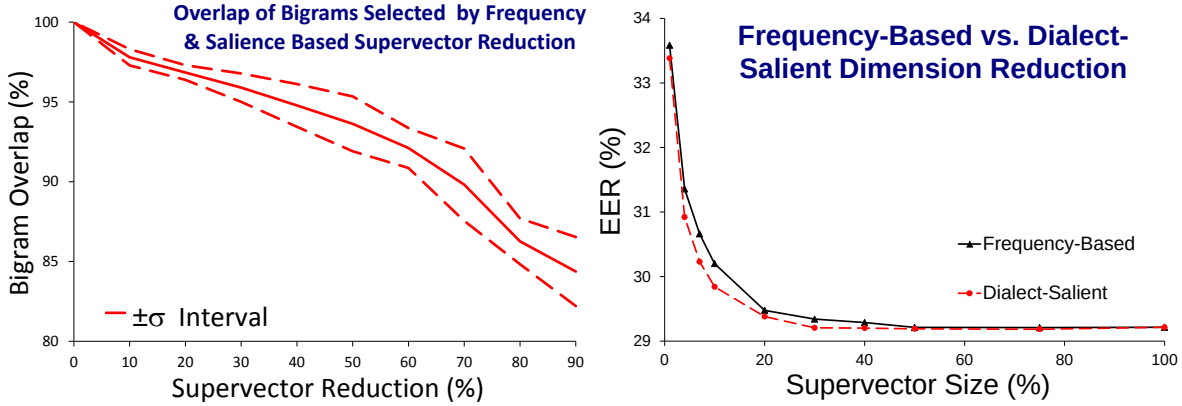


Figure 42: *Left - Overlap between N-grams selected by frequency and dialect-salience based methods; Chinese corpus; right - Impact of dimension reduction on DID performance. Results averaged across 9 phone recognizer systems.*

- Pitch Pattern Supervectors for SVM-Based LID/DID:
  - A pitch pattern modeling scheme has been proposed that is inspired by conventional PRSVM framework (see Fig. 43). In PRSVM, a phone recognizer is used to tokenize speech samples into sequences of symbols - estimated phones. These sequences are then processed by statistical language modeling to calculate the phone N-grams. Finally, the N-gram statistics are stored as dimensions of a supervector, which is processed by SVM models/classifier.
  - In this approach, a sequence of symbols representing pitch contour variations (up, flat, down) with a processing window is generated. These symbols are then grouped into longer sequences resembling words in terms of number of symbols (characters). Finally, LM statistics are extracted for these sequences of words and stored in supervectors. The proposed technique was successfully applied to very short utterances (~3 sec. long) for language background identification in toddlers [Boril14].

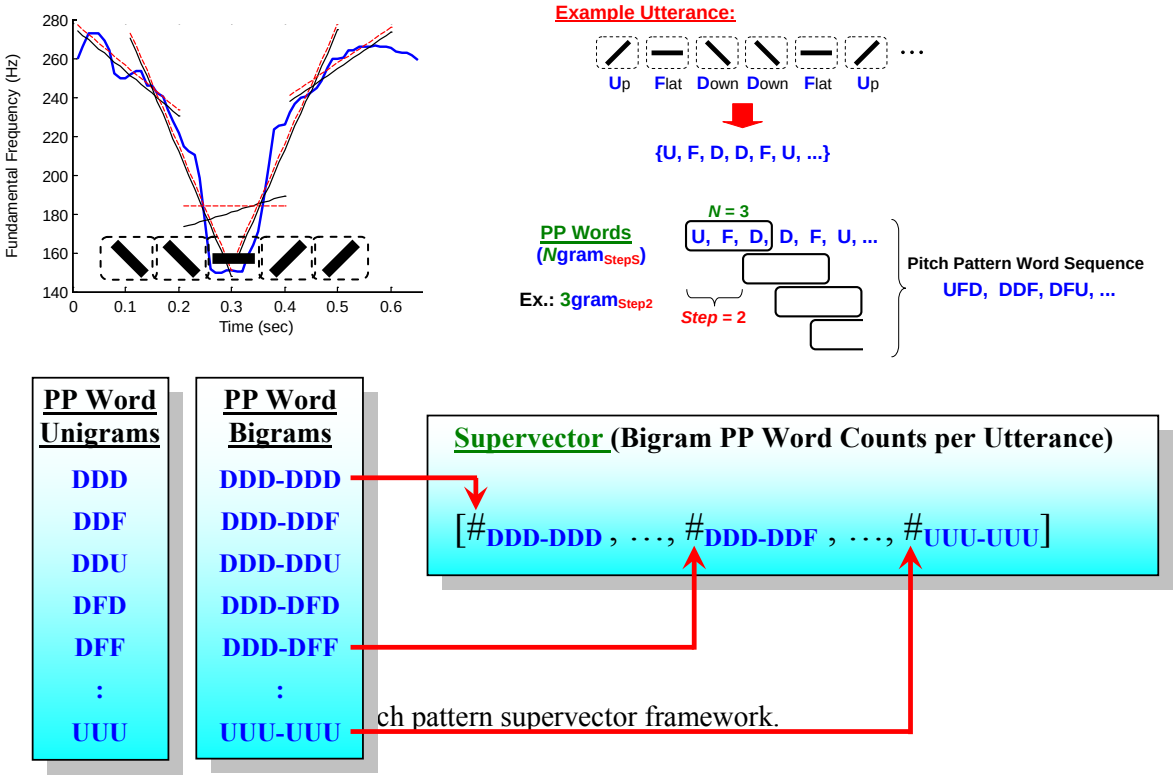


Figure 43: Fundamental frequency F0/pitch modeling with bigram pitch-pattern word models for dialect model development and classification.

2.6.3: References for this Section:

[Boril14] Boril, H., Zhang, Q., Ziaei, A., Hansen, J. H. L., Xu, D., Gilkerson, J., Richards, J. A., Zhang, Y., Xu, X., Mao, H., Xiao, L., Jiang, F. (2014). "Automatic Assessment of Language Background in Toddlers Through Phonotactic and Pitch Pattern Modeling of Short Vocalizations," Workshop on Child Computer Interaction (WOCCI), September 19 (Singapore).

[Zhang13] Zhang, Q., Boril, H., Hansen, H.L., Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification', IEEE ICASSP 2013.

[Boril12] Boril, H., Sangwan, A., Hansen, J. H. L. (2012). "Arabic Dialect Identification – 'Is the Secret in the Silence?' and Other Observations," ISCA Interspeech'12, September 9-13 (Portland, Oregon).

Task 3 – Co-Speaker Diarization/Environment (CoSpkrD):

This task area has focused on two related aspects: (i) first, to detect and address overlap speech in audio streams, and (ii) to advance speech/language technology for diarization of continuous audio streams. Fig. 44 illustrates the high level structure for the first sub-task which is overlap speech detection.

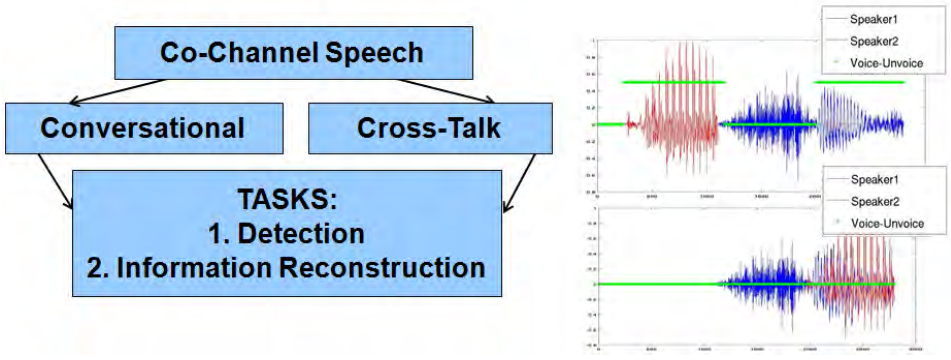


Figure 44: *High level structure for overlap speech detection and processing.*

Publications produced during this 36 month project period from CRSS-UTDallas for Task#3 are summarized here.

[1] N. Shokouhi, J.H.L. Hansen, "Overlapped Speech Detection with Applications to Speaker Identification," submitted to IEEE Trans. on Audio, Speech and Language Processing, April 2015.

[2] A. Ziaei, A. Sangwan, J.H.L. Hansen, "Effective Word Count Estimation for Long Duration Daily Naturalistic Audio Recordings," submitted to IEEE Trans. on Audio, Speech and Language Processing, Mar. 2015

[3] A. Ziaei, A. Sangwan, J.H.L. Hansen, "Prof-Life-Log: Environment Tracking Framework for Natural Audio Streams," submitted to IEEE Trans. on Audio, Speech and Language Processing, June 2014. Revised March 2015.

[4] N. Shokouhi, S.O. Sadjadi, J.H.L. Hansen, "Co-channel Speech Detection via Spectral Analysis of Frequency Modulated Sub-bands," ISCA Interspeech-2014, Paper #437, Singapore, Sept. 14-18, 2014.

[5] S.M. Mirsamadi, J.H.L. Hansen, "Multichannel Speech Dereverberation Based on Convolutive Nonnegative Tensor Factorization for ASR Applications," ISCA Interspeech-2014, Paper #414, Singapore, Sept. 14-18, 2014.

[6] A. Ziaei, L. Kaushik, A. Sangwan, J.H.L. Hansen, "Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions," ISCA Interspeech-2014, Paper #994, Sept. 14-18, 2014.

[7] A. Ziaei, A. Sangwan, J.H.L. Hansen, "A Speech System for Estimating Daily Word Counts," ISCA Interspeech-2014, Paper #1028, Singapore, Sept. 14-18, 2014.

[8] N. Krishnamurthy, J.H.L. Hansen, "Car Noise Verification and Applications," Inter. Journal of Speech Technology, Volume 17, Issue 2 (2014), Page 167-181 vol. 17, Issue 2, pp. 167--181, 2014

[9] N. Shokouhi, S.O. Sadjadi, A. Sathyanarayana, J.H.L. Hansen, "Analysis of In-Vehicle Speech Activity towards Driver Safety Assessment," Biennial Workshop on DSP for In-Vehicle Systems & Safety, Seoul, Korea, Sept. 29-Oct.2, 2013.

[10] N. Shokouhi, A. Sathyanarayana, O. Sadjadi, J.H.L. Hansen, "Overlapped-Speech Detection With Applications for Driver Assessment for In-Vehicle Active Safety Systems," IEEE ICASSP-2013, Vancouver, Canada, May 26-31, 2013

[11] A. Sathyanarayana, N. Shokouhi, S. O. Sadjadi, J. H. L. Hansen, "Belt Up: Investigating the Impact of In-Vehicular Conversation on Driving Performance," Intelligent Vehicle Symposium, Australia, June 2013.

Citations in this sub-section refer to the list of references at the end of this sub-section.

3.1: Overlapped Speech Detection for Speaker Recognition

3.1.1: Overlapped Speech Detection

Our past efforts in overlapped speech detection have been directed towards identifying the elements that make overlap detection challenging, including effective features, detection strategies, and speech system performance. A number of publications from this effort focus on the analysis of overlapped speech with a comparison of single and double-speaker feature spaces [1, 2]. In Shokouhi et al, 2013 [1], a phone-level comparison was presented in which more confusable phone-pairs in overlapped regions were ranked and given lower priority in the detection hierarchy. This analysis lead to a supervised overlap detection method in which confusable phone-pairs (for example pairs including nasals on one side, as shown in Fig. 45) were omitted from the training data in speech task such as speaker recognition.

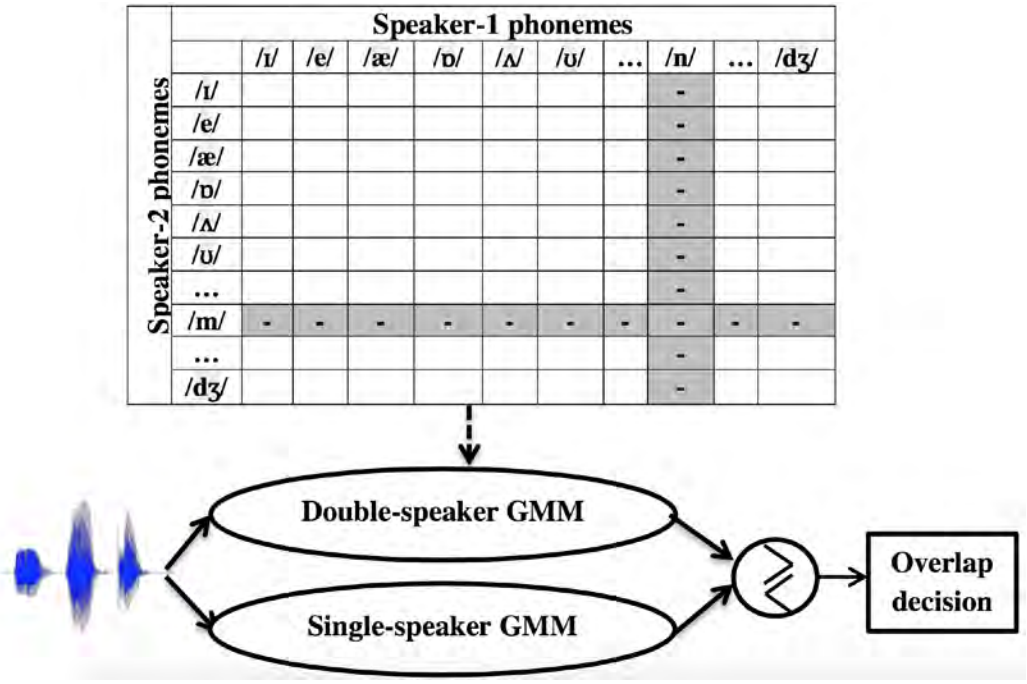


Figure 45: *Phone-based data purification for overlap detection.*

In later studies, the focus was on developing effective features for overlap detection [3,4,1]. Here, the Gammatone subband frequency modulation (GSFM) features were derived based on the idea of adding a nonlinearity component in the time-frequency analysis of the spectrogram in order to magnify the presence of multiple harmonics. Speech subbands were used as instantaneous frequency stimuli to sinusoidal carriers. The spectral characteristics of the frequency modulated sinusoids take on the form of Bessel functions, which become significantly distorted in the presence of multiple harmonics due to more than one speaker being present.

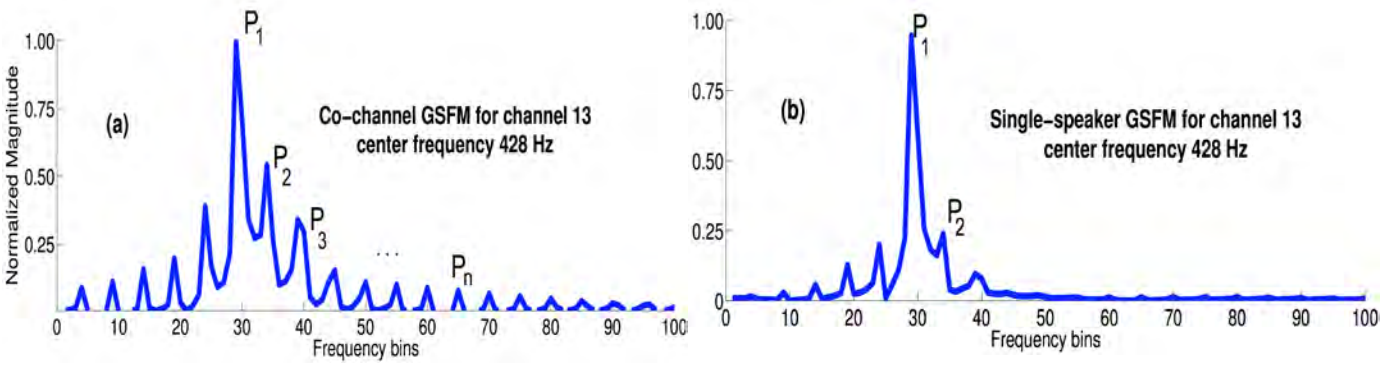


Figure 46: *GSFM spectra for a given time-frequency unit in the spectrogram. The number of significant peaks are more in the case of overlapped segments. This forms the basis of GSFM feature extraction.*

One of the drawbacks of the GSFM feature extraction algorithm is that it becomes un-reliable in real noise conditions. In an effort to achieve more robust performance in the presence of noise, an alternative overlap detection algorithm was proposed based on enhanced spectrograms, called pyknograms. In [4], pyknograms are extracted from speech spectrograms to remove all non-harmonic components from the signal, which include stationary ambient noise. The resulting time-frequency representation is a 2-dimensional pattern that contains only signal resonances/harmonics. This allows for the detection of the presence of interfering speakers by tracking coherent harmonic patterns that belong to the foreground speaker. The notion of connected the harmonic tracks in the time-frequency space during speech production can be viewed as connecting speech harmonic patterns with corresponding tracks in the snow of a mountain top from skiing. When tracks from different skiers intersect, it is easy to tell them apart due to the parallel structure that is expected from a single skier (as shown in Fig. 47). If the signal is truly a voiced speech section from a speaker, there must be a degree of connectivity between the harmonic tracks due to the resonant physiology in the speech production process.

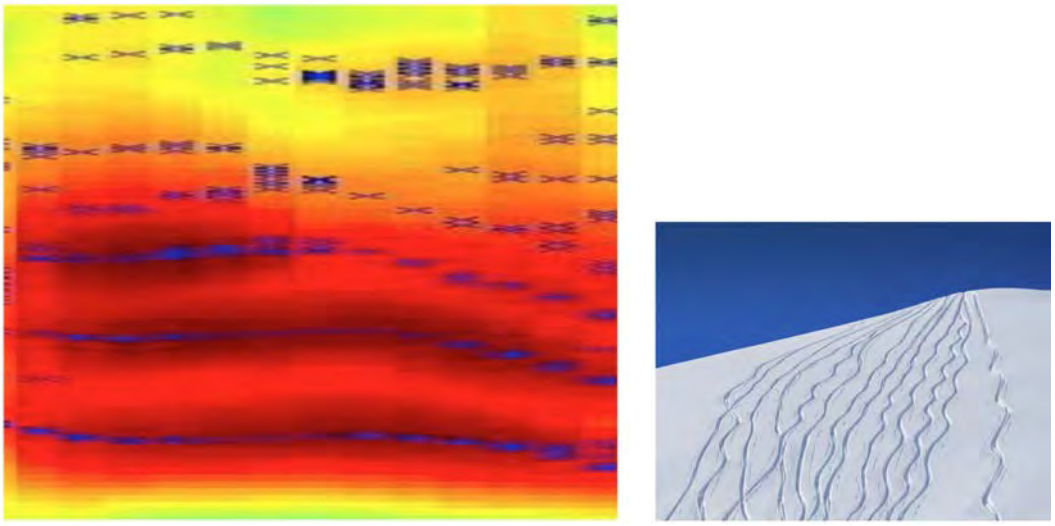


Figure 47: *A comparison of (i) pyknograms (foreground on the left) with corresponding skiing tracks.*

3.1.2: Separation of Co-Channel Speech

A significant portion of the work performed in overlap detection has been themed around speaker recognition applications. However, both from our own experience and also from existing studies, it can be concluded that overlapped speech on its own does not highly affect speaker identification in real conversations (as opposed to artificially generated overlap data), mainly due to its rare and short-term occurrence in conversational speech. The number of frames which contain overlap speech in general 2-way conversation is in general quite low. Co-channel speech (which we define as the presence of two speakers on a *single* recording channel regardless of whether they speak exactly at the same time), on the other hand, significantly drops speaker recognition performance. Therefore, the focus over the more recent period has been on improving speaker recognition in co-channel signals. More accurately, the goal is to detect the identity of the foreground speaker despite the fact that the signal at hand is a single channel recording distorted with speech from a secondary speaker.

The traditional approach with respect to co-channel speech has been to perform speaker diarization on the co-channel signal. However, performing a task as time-consuming and sophisticated as diarization is intangible in a large scale speaker recognition task. This motivates an alternative strategy to address the problem by attempting to separate the interfering speaker from i-Vectors that are extracted from co-channel signals. This approach is justified from a speaker recognition point of view, since: 1) i-Vectors have become the de-facto algorithm for speaker recognition and formulating separation algorithms which employ i-Vectors allows for a more seamless integration of the speech separation and recognition steps. 2) i-Vectors are low dimensional representations of the entire audio files, and therefore easier to process at some level. Hence, algorithms that use i-Vectors are computationally more attractive compared to applying signal processing algorithms on the original signals/spectrograms.

The approach taken here is to add a speaker dependent term intended to model the interfering speech in the PLDA model for channel compensation. Since this term also corresponds to speaker identities, it should as well be represented within the speaker subspace. When considering i-Vectors extracted from co-channel recordings, interfering speech adds an additional kind of variability to the i-Vector space. This issue is addressed by considering two tactical approaches:

- Constructing co-channel background data to train the PLDA models.
- Modifying the PLDA formulation to fit the co-channel speech paradigm.

The first approach relies on the same ability of the PLDA framework that already compensates for channel mismatch. The formulation of this algorithm is entitled the *co-channel PLDA* method, and represents a combination of the standard PLDA method and simplified PLDA, which uses a full-covariance noise term to characterize the channel interference (paper submitted to ISCA Interspeech-2015 [5]). Modifying PLDA to deal with co-channel data as described results in robust performance under a range of co-channel conditions. Figure 48 shows the resulting EER performance across different signal-to-interference ratios (SIR) for three methods with co-channel data obtained by mixing channels from the Switchboard corpus (i.e., in the LDC Switchboard corpus, which contains 2-way conversations over the telephone, the individual separate sides A and B of the conversations are available – so it is possible to know exactly when overlap speech events occur). The first method (*cleanPLDA*), used as a point of reference, is formed when the co-channel data is added to the trial set without any consideration of speaker interference. The baseline method (*mixedPLDA*) treats co-channel interference the same way as when dealing with channel interference. Adding co-channel data to the development set helps accomplish this aspect. The proposed method (*cchPLDA*), uses the modified PLDA algorithm to remove the secondary speaker interference from the i-Vector space, and thereby improving speaker modeling and recognition performance. To illustrate the performance more broadly, Fig. 49 presents results for the corresponding DET curves.

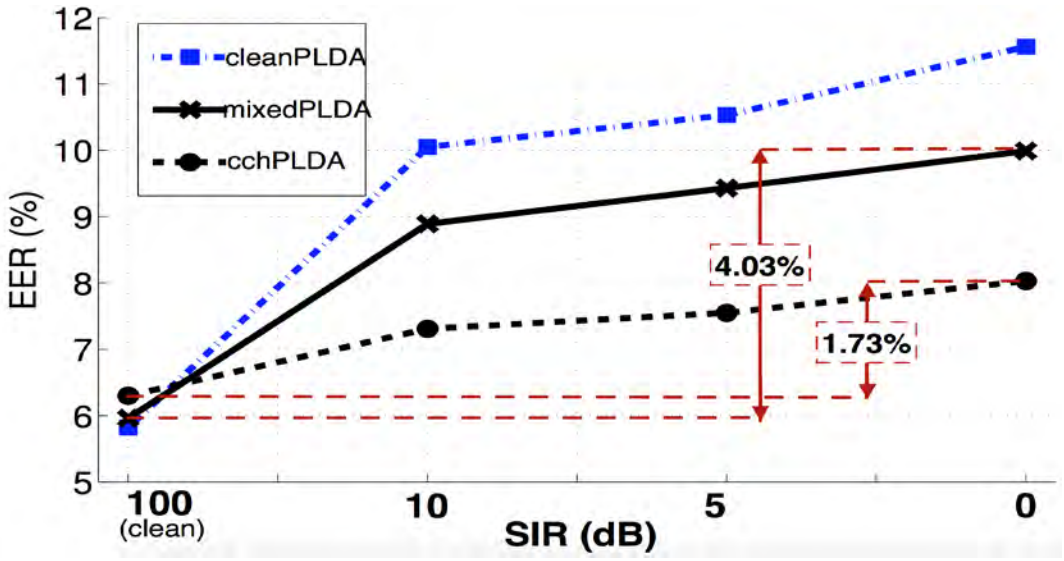


Figure 48: *Speaker Recognition performance as the Signal-to-Interference (SIR) ratio representing the degree of overlap speech from a competing speaker is introduced into the primary speaker audio stream. Three scenarios are shown: (i) cleanPLDA, (ii) mixedPLDA, and (iii) the proposed cchPLDA solution.*

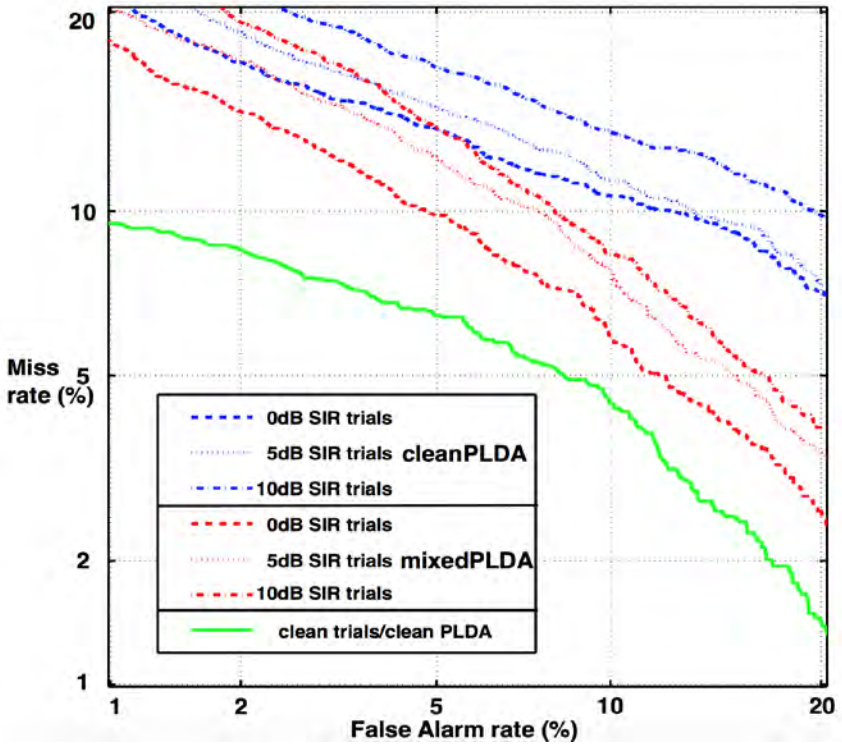


Figure 49: *Resulting DET curves of the PLDA based approaches to address overlap speech in speaker recognition. The "clean trials/clean PLDA" scenario represents the groundtruth condition.*

The success of this modified PLDA algorithm for speaker recognition described is encouraging, but still in its early stages. The intention next is to combine this approach with the recently developed phone-based i-Vector extraction paradigm. This would allow for addressing the problem of co-channel speaker recognition

with assumed knowledge of the conversation transcripts. The idea is to replace the unsupervised clustering of the UBM acoustic space into Gaussian mixtures with a supervised classification obtained on a per tri-phone basis. This has shown to improve i-Vector/PLDA speaker recognition performance. The goal next is to use this approach to force the desired phone labels obtained from the transcripts of the target speaker in co-channel data to indirectly remove the effect of the secondary speaker.

The problem of co-channel/overlap speech is challenging – and during this 36month period, both meaningful and quantitative successful advancements have been made, especially to help with speaker recognition. While future efforts will continue to be applied to improving speaker recognition, the intent is to also explore the ability to improve automatic speech recognition. This problem has recently been the focus of many researchers due to the monaural speech separation challenge. The goal would be to use this experience from overlap detection and more recently separation in the i-Vector space to improve ASR for co-channel speech signals.

Table 26: *Overlap detection error rates as relative overlap speech distortion (in dB) increases*

overlap in test	clean	6dB	3dB	0dB	-3dB	-6dB	-9dB
female	2.02	8.99	11.58	15.59	19.82	24.05	27.88
male	1.27	4.87	6.65	9.97	13.39	17.26	22.6

Table 27: *Impact of overlap speech detection and removal for SID test data*

overlap in train	clean	6dB	3dB	0dB	-3dB	-6dB	-9dB
Female	3.1	9.18	10.5	11.65	12.95	14.43	15.26
Male	2.55	8.27	9.11	10.27	11.16	12.29	13.27

3.1.3: Co-Channel/Overlap Speech Detection and Separation References:

[1] N. Shokouhi, A. Sathyanarayana, S.O. Sadjadi, J.H.L. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," IEEE ICASSP, Vancouver, BC, May 2013.

[2] N. Shokouhi, J.H.L. Hansen, "Overlapped Speech Detection with Applications to Speaker Identification", submitted to IEEE Trans. on Audio, Speech, and Language Processing.

[3] N. Shokouhi, S. O. Sadjadi, and J. H. L. Hansen, "Co-channel speech detection via spectral analysis of frequency modulated sub-bands," *ISAC INTERSPEECH*, Singapore, September 2014.

[4] N. Shokouhi, A. Ziaei, A. Sangwan, J.H.L. Hansen, "Robust Overlapped Speech Detection and its Application in Word-Count Estimation for Prof-Life-Log Data," IEEE ICASSP, Brisbane, Australia, April 2015.

[5] N. Shokouhi, J.H.L. Hansen, "Probabilistic Linear Discriminant Analysis for Robust Speaker Identification in Co-channel Speech," submitted to ISCA INTERSPEECH 2015.

3.2: Prof-Life-Log: Diarization for Real-World Audio Streams: *Speech Diarization / Massive Data Sets*

There is an ever increasing presence of audio content from individuals in their daily lives, including voice mail, voice search, human-to-human and human-to-machine communications. Capture, processing/interpretation, storage, and recall based on personal daily information is a key cognitive demand placed on each of use on a daily basis. Extracting specific knowledge of daily voice interaction represents

new and emerging domains for personal productivity assessment, analysis of conflict resolution, and health care monitoring. While significant progress has been made throughout this 36month period, this section will emphasize advancements during the last 12 month period to address audio analysis as well as improvement of systems and algorithms for massive naturalistic audio data. The Prof-Life-Log corpus (PLL) has been established (and continues to expand) over this current project as shown in Fig. 50, resulting in the following four contributions:

- [Contribution 1] speech activity detection (SAD), (i.e., the task of separating speech from other background sounds);
- [Contribution 2] environment detection and tracking;
- [Contribution 3] statistical word counting, (i.e., the task of counting the number of words spoken); and
- [Contribution 4] speech interaction diarization, (i.e., the task of detection and classifying speech events such as conference calls, teaching, as well as word/topic content).

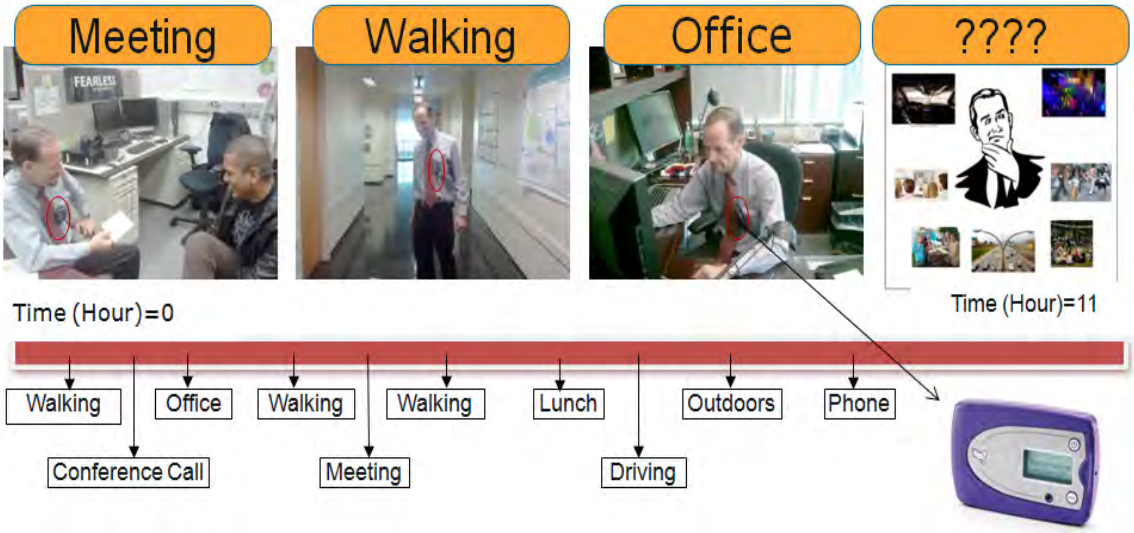


Fig. 50: *Prof-Life-Log data collection using the LENA unit: A single session consists of 16+ hours of audio recording with the speaker constantly carrying the unit. Audio is collected in a wide variety of backgrounds such as Cafeteria, Office, Meeting, Walking, Driving, and others.*

3.2.1: [Contribution 1] Speech Activity Detection (SAD): In [Contribution 1], a new unsupervised threshold estimation technique which was formulated that enhances a state-of-the-art unsupervised SAD algorithm for long-duration non-speech/ speech-sparse segments [1]. The resulting system is shown to achieve a +30% improvement over base-line system on the Prof-Life-Log corpus.

The block diagram of the proposed system vs. the previous baseline SAD (Combo-SAD [2]) is shown in Fig. 51. Combo-SAD was also developed over this project period, but the Threshold Optimized (TO-Combo-SAD) provides for greater performance in extensive naturalistic audio streams.

The proposed technique (i.e., TO-Combo-SAD) as is shown if Fig. 51, first builds a prior model of speech (using external corpora), and dynamically chooses between the prior and posterior speech models (posterior model is built from data by Combo-SAD in an unsupervised fashion). The new method is designed to choose the prior model whenever the posterior model is weak. Therefore, it consistently delivers superior results over Combo-SAD.

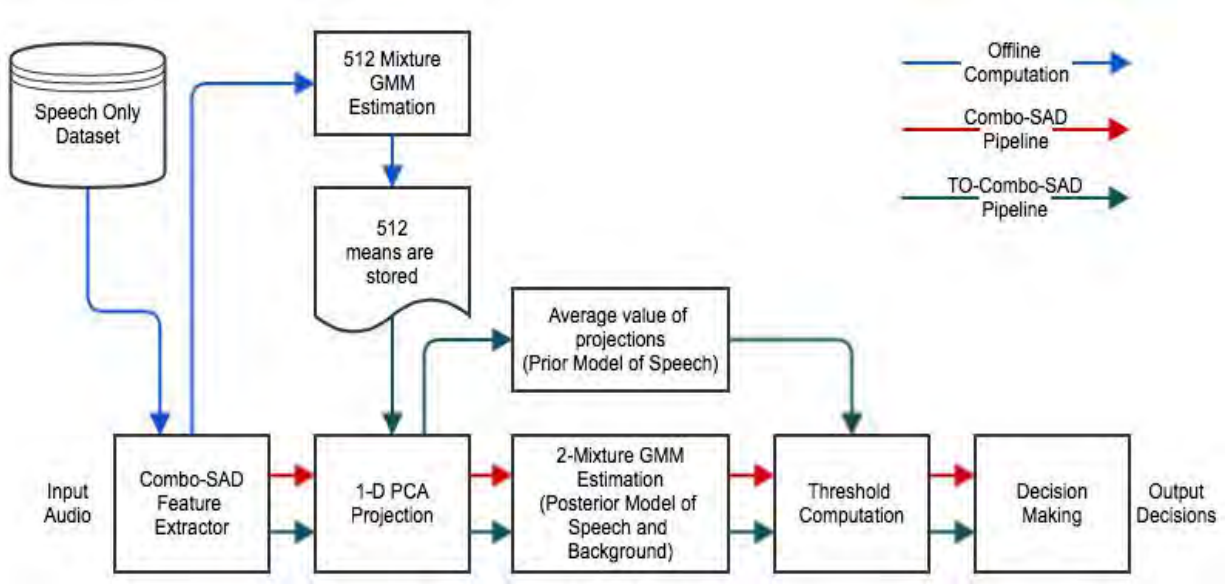


Fig. 51: *TO-Combo-SAD – the threshold optimized Combo-SAD flow diagram*

Fig. 52 shows DET (detection error trade-off) curves for TO-Combo-SAD and Combo-SAD. The DET curves for overall performance are shown along with curves for (i) speech-sparse, (ii) speech-pause balanced, and (iii) speech-dense data. From the figure, it can be seen that TO-Combo-SAD obtains a remarkable improvement over the state-of-the-art unsupervised Combo-SAD baseline for speech-sparse data. Due to the formulation of TO-Combo-SAD system, it is very likely that the prior model of speech is frequently chosen for decision making in speech-sparse region (and this explains the huge performance improvement). Interestingly, it was also observed that moderate improvements were obtained for speech-pause balanced and speech dominant cuts as well. For these two cases, the improvements are seen in terms of false-alarm reduction for higher values of the miss-rate. Overall, the EER drops from around 40% to 10% using the new proposed solution. This configuration would be ideal as a first phase processing step for any bulk audio processing where the recording conditions are unknown or variable.

It should also be noted that it is true that TO-Combo-SAD has addressed the problem related to long duration non-speech segments. However, the problem for harmonic interference still persists. In future efforts, the ability of TO-Combo-SAD to improve performance over harmonic interference, and related research to this problem, will be considered.

3.2.2: [Contribution 2] Environmental Sniffing and Tracking:

In this area, a new framework was developed to enhance state-of-the-art environmental sniffing to detect and track variable length environments. An evaluation over the Prof-Life-Log corpus shows a +40% relative improvement over the previous base-line system.

For this purpose, major efforts were undertaken on a robust environment tracking system that simultaneously segments and classifies audio, and provides time aligned information. It may be useful to think of the proposed system as an environment diarization solution. The proposed system addresses the issue of noise corrupted by speech audio samples utilizing speech activity detection (TO-Combo-SAD). By using the speech *vs.* pause decisions, the audio corresponding to speech decision can be removed to obtain a relatively uncorrupted (or less corrupted) environment audio sample (i.e., performing speech suppression on the audio first, allows for better pure noise/environment classification and tracking).

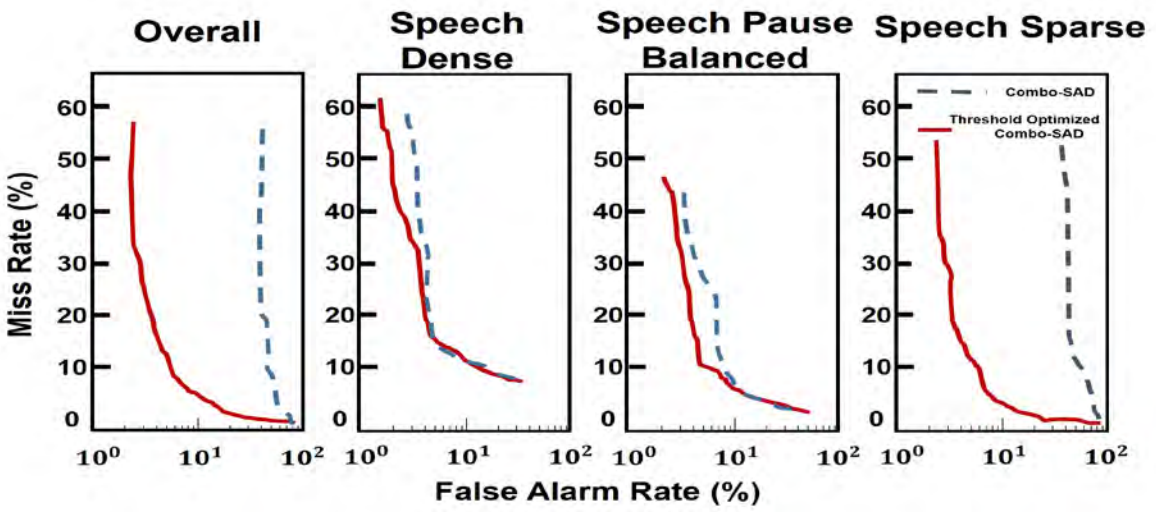


Fig. 52: *DET curves for PLL data utterances.*

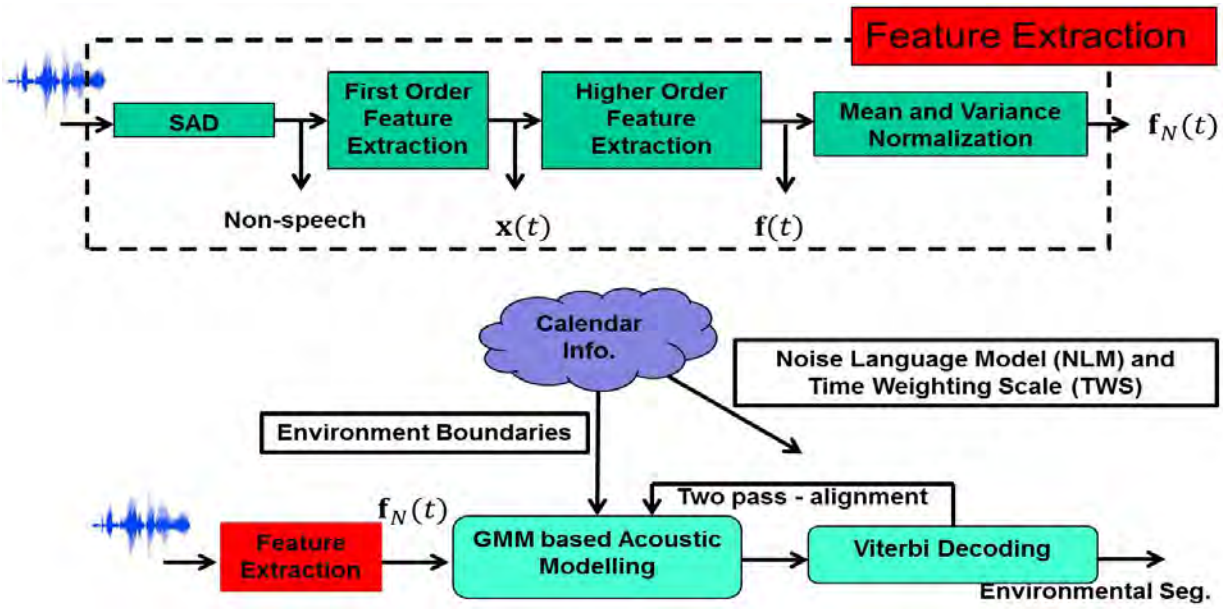


Fig. 53: *Environment sniffing system structure*

Additionally, most environment classification studies assume that ground truth classification labels are available. This is not typically true for personal audio recordings (PARs). The traditional method relies on human annotation to generate ground truth labels. The human annotation method also tends to be largely subjective and error prone. To address this problem, the issue of assigning ground truth labels for various environments in the audio is approached by using a simple yet effective technique that utilizes the primary speaker's calendar (which is readily available for PARs).

Finally, a user is more likely to transition into a new environment the longer he/she stays in the current environment (*i.e.*, a person is more likely to leave the office if they have already been there for 3 hours as opposed to 3 minutes). The use of noise language model such as the one proposed by [3] offers an efficient method to model the mentioned temporal constraints. However, the time context of the noise language model is somewhat narrow, (i.e., it only sees 1 or 2 decisions backwards in time, which typically corresponds to several seconds depending upon the decision frame duration). Here, a method was proposed

to capture a wider time-context in the noise language model. Using a wider time-context allows for smoothing of the decisions more effectively and therefore improves the overall performance of the system. The total system structure and performance comparison based on F-measure criteria are shown in Fig. 53 and 54.

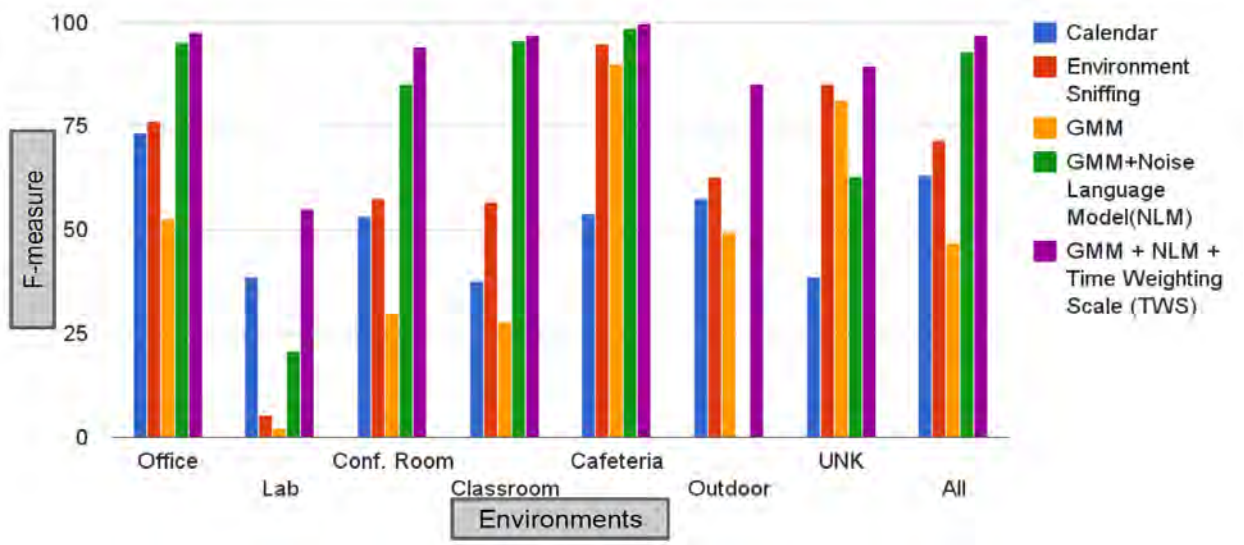


Fig. 54: *Performance of the proposed system compared to baseline systems on 5 hours of Prof-Life-Log audio. Output labels detected by various systems: (a) Ground Truth, (b) Environment Sniffing, (c) proposed GMM, (d) proposed GMM + Noise Language Model (NLM), and (e) proposed GMM + NLM + Time Weighting Scale (TWS). GMM+NLM+TWS system provides best results in terms of classification accuracy.*

From the bar chart in Fig. 54, it can be seen that the calendar accuracy is 63.08%. This accuracy is sufficient for estimating initial model parameters for each environment. This step is similar to a flat initialization in ASR training. Using this initial estimation, the Viterbi decoder does force-align the boundaries for each environment.

By comparing performance of "GMM" only system and "Environmental Sniffing", it is observed that the second system has a +25% absolute greater accuracy rather to the first solution. The reason behind this is based the fact that using a noise language model (NLM) helps greatly for the solution to stay in one environment and not fluctuate between similar environments. Since, a 0.5 second audio segment cannot capture overall environment characteristics, not using the NLM impacts system performance and causes confusion for similar environments such as 'Lab', 'Office', 'Classroom', and 'Conference room'.

Using NLM helps the system to not fluctuate between similar environments but on the other hand, if it is fixed over time, it cause the system ignore short duration events. To address this problem, a time weighting scale (TWS) was used to scale down the NLM impact over time for one environment to help the Viterbi algorithm switch between environments. This factor, TWS, helps the system to achieve a +50% relative lower error rate than using the NLM only.

In summary, it should be noted that to find the mapping between environment and calendar information, it was necessary to use heuristic rules. For example, it is assumed that whenever in calendar, an event related to research meeting is observed, it should be occurred in `office'. This mapping in the future could be built automatically using machine learning approaches. Alternatively, using unsupervised techniques to find this mapping, helps the system to be person independent. Further emphasis on this area of research is a potential direction in future improvements.

3.2.3: [Contribution 3] Word Count Estimation:

In this area, a new syllable rate based word counting system was proposed and employed on naturalistic audio streams within the domain of diarization. In spite of the algorithmic simplicity, the solution shows promising results for daily 8-16 hours audio recordings. For the Prof-Life-Log corpus, in terms of word count error, it is possible to achieve as low as 1% error at the end of a 16-hours day.

Counting the number of words spoken by a person is a rich and valuable piece of information for several applications such as health monitoring, second language learning or language development studies. In spite of word counting importance, in real scenarios, developing a robust word count system that can achieve a high performance with low computational cost is very challenging due to natural behavior of audio recordings. In last year, this problem has been addressed for Prof- Life-Log. To do word count, a new framework has been proposed based on syllable detection which has been shown in Fig. 55.

The new framework consists of five main parts, Speech Activity Detection (SAD), speech enhancement, Primary vs. Secondary speaker detection, syllable detection and LMMSE estimator. For separating speech from non-speech, TO-Combo-SAD has been chosen since it shows promising results for naturalistic audio streams. For speech enhancement, four different well-known algorithms have been evaluated which are spectral subtraction (SS) [4], MMSE [5], pKLT [6] and Wiener filtering [7]. To separate Primary vs. Secondary speaker, first the open source speaker diarization toolbox by LIUM [8] has been used to provide initial diarization. The output of the LIUM toolbox provides an initial hypothesis of the first and second speaker. In the next step, energy for the hypothesized segments is computed. By averaging the segment-level energy estimates for first and second speakers, and then selecting the speaker with higher energy level as primary speaker, primary vs. secondary speaker separation is achieved. Three syllable detection algorithms have been evaluated so far. The first one is mrate [9], second one is the modification of mrate proposed by Shri [10] and the last one is Praat based proposed by De Jong and Wempe [11]. Finally, [10] is selected as the baseline syllable detection system framework based on evaluation results on PLL.

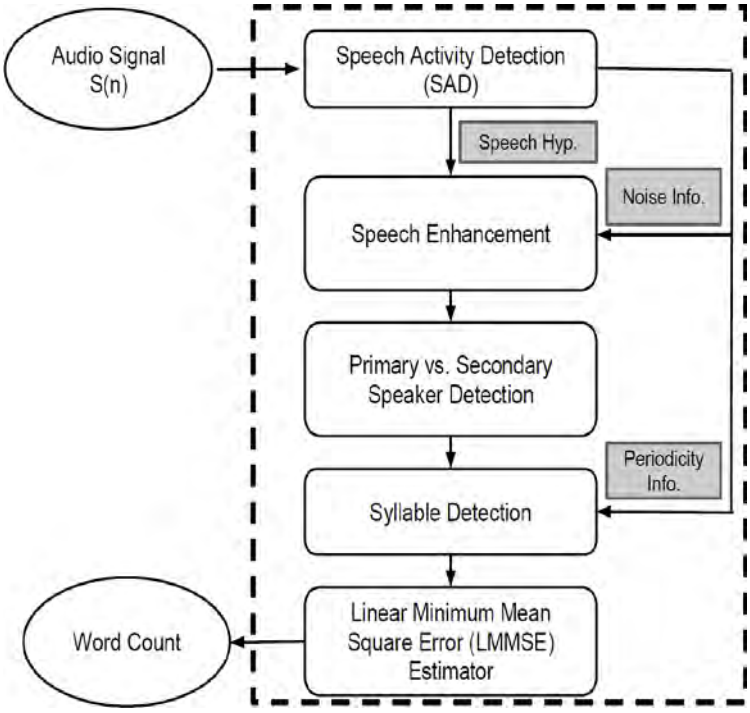


Fig. 55: *Word count system structure*

To evaluate the proposed frame work, Word Count Error (WCE) is defined as:

$$WC_{err} = 100 \times \frac{|\,\#Actual\,words - \#Estimated\,words\,|}{\#Actual\,words}.$$

To assess the enhancement techniques performance on word count estimation for Prof-life-Log, a controlled experiment was used. In this experiment, multiple pure-samples (speech-free) were extracted of the dominant noise-types from a Prof-Life-Log development day, and used these samples were used to corrupt the Hub-5 data. This process allows one to study the performance of the proposed method in noise, in a more controlled fashion. In particular, the following (i) Office, (ii) Cafeteria, (iii) Car and (iv) Outdoor noise-types were selected from development day data from Prof-Life-Log.

These noise-types were chosen since they reflect the diversity of conditions, (i.e., from relatively quiet (office) to extremely noisy (Cafeteria)). Using the FaNT tool, three SNR variants for each noise-type was generated (i.e. , 0dB, 10dB, and 20dB), which produced a total of 12 variants of Hub-5 (4 noise-types and 3 SNR values). Now, it is possible to evaluate the effect of different noise enhancement techniques. For the SS algorithm, based on the equations (3, 4) in [4], the value of **β** is set to 0.05. For MMSE algorithm based on equations (7, 30, 51) according to [5], **q** and **α** are set to 0.2 and 0.98, respectively. Base on equations (4-6 and 8) in [7], **β** and **λ** are set to 0.98 for Wiener filtering. Finally, based on equation 34 in [6], **ν** is set to 0.08. The TO-Combo-SAD weight is set to 0.5 based on the best EER% on the tuning day. The results are shown in Fig. 3-12.

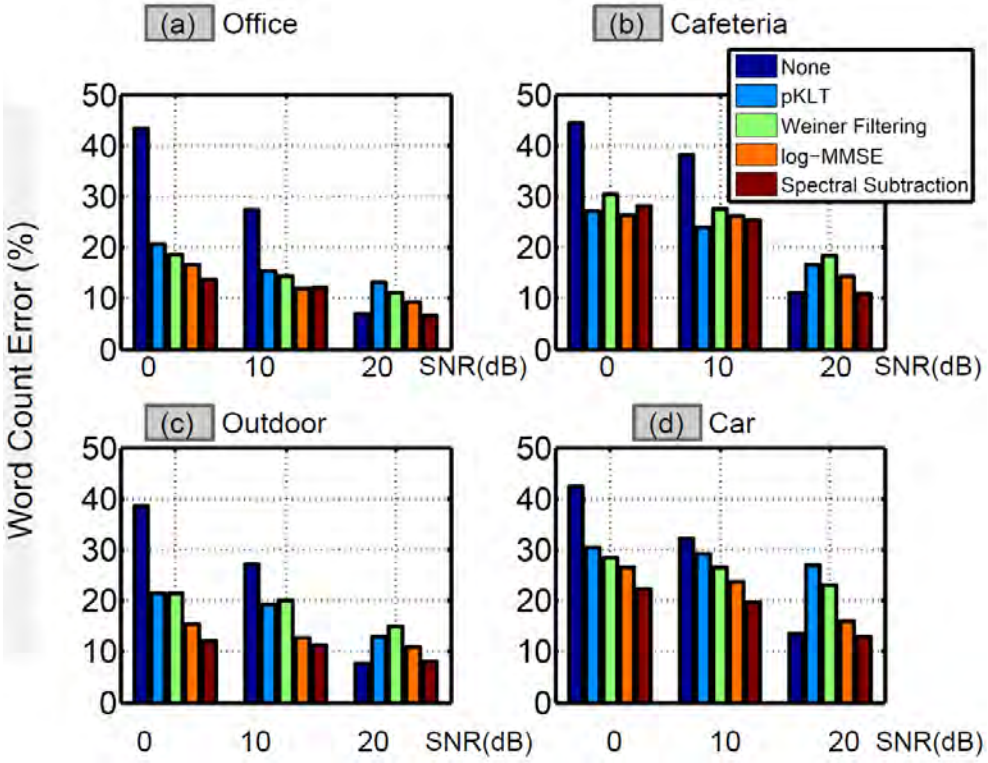


Fig. 56: *Comparing performance of 4 speech enhancement techniques: (i) SS (ii) pKLT (iii) log-MMSE and (iv) Wiener Filtering using word count error (WCE) for noisy Hub-5 dataset. Word count errors when not applying any speech enhancement are also shown and labeled as "None". In general, SS generally provides the lowest WCE.*

From Fig. 56, several facts can be observed. First, in all conditions and SNRs, Spectral Subtraction achieves better performance than the other three algorithms except Cafeteria in which for 0 and 10 dB, MMSE is

working better than SS. The second observation is that, for 20 dB SNR, not doing enhancement is better than applying an enhancement technique. This can be because of distortion they make on almost clean speech signal (20dB SNR). The last observation is that pKLT is always the worst algorithm except in Cafeteria environment, in which Wiener filtering is the worst. Again, knowing when to capture a sample of the background noise and potential noise updated rates can impact performance as well.

Next, it is off interest to assess performance of the proposed framework on Prof-Life-Log evaluation set (12 days with 93 hours transcribed data). For this purpose, the database is categorized into 4 different conditions. Office-Like which consist of office, lab, classroom and conference room; Cafeteria-Like which consists of cafeteria and hallway environments. The third is Outdoor audio segments, and the last is the car environment. All audio segments are approximately 2.5mins duration. The environment-specific performance is shown in the Fig 57.

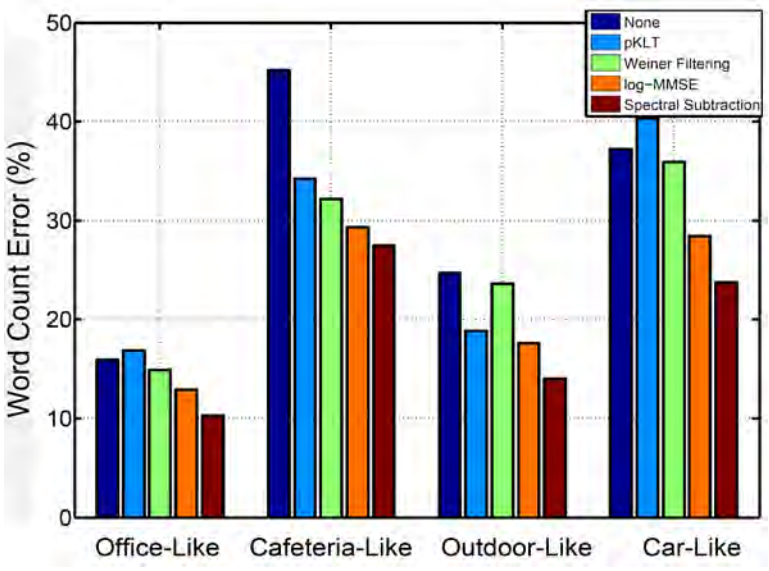


Fig. 57: *Comparing the performance of 4 speech enhancement techniques: (i) SS (ii) pKLT (iii) log MMSE and (iv) Wiener Filtering using word count error (WCE) for Prof-Life-Log. Word count errors when not applying any speech enhancement are also shown and labeled as None. SS provides the lowest WCE across all environments.*

The first observation is that the Spectral Subtraction has the best results on all four conditions. In addition, the overall performance for Office-Like environments (Most dominant environments in evaluation set) is 9.53% (It means that in average 10 words error in actual 100 words). Also for Cafeteria-Like environments which are very noisy as well as Car environment, the overall Word Count Error is around 27% which is very reasonable with respect to very low computational resources our system needs.

In the next experiment, trends in cumulative word count estimation error were analyzed for three workdays (using SS enhancement technique and modified m-rate based syllable detector in the frame work) as shown in Fig. 58.

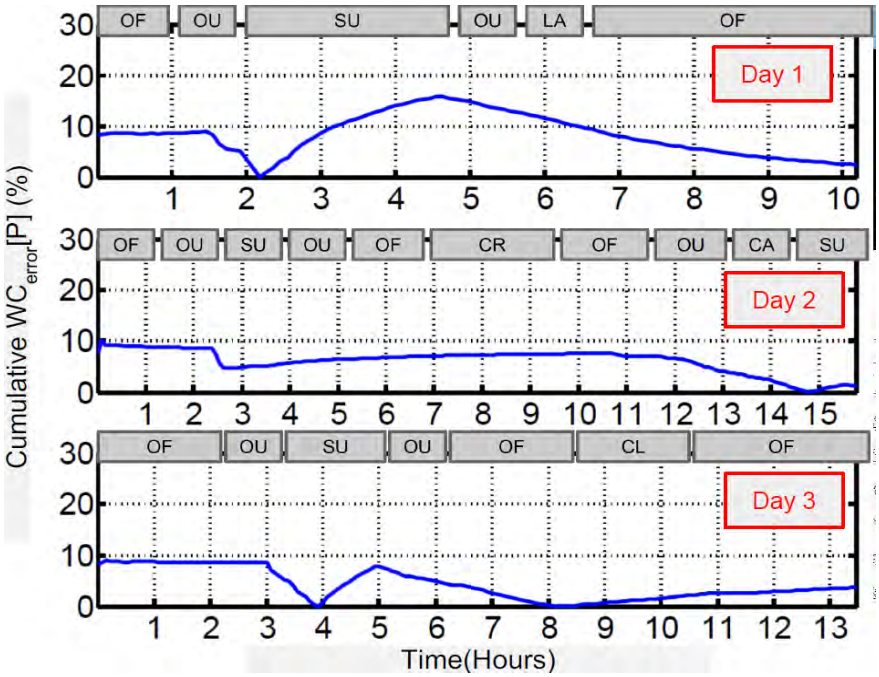


Fig. 58: *Cumulative Word Count errors for three sample days from Prof-Life-Log data. OF: Office, OU: Outside, SU: Cafeteria, LA: Lab, CR: Conference room, CA: Car, CL: Classroom*

When compared to the WCE for segment level, the cumulative WCE generally tends to be lower than the best numbers we obtained at segment level. In other words, the best performance at segment level was 10% WCE (for office-like environment using SS), and the cumulative WCE for all days in Fig. 58 is lower than 10% (with hours 3-to-6 on Day 1 being the only exception). In fact, the final WCE at end of day is 1.7%, 1% and 4% for days 1, 2, and 3, respectively. The reason for lower cumulative WCE is that the system over estimates the number of words in noisy environments such as `cafeteria' or `hallway', and it underestimate the number of words in more relatively quite environments such as `office'. These two errors cancel each other with time. For example, in Fig. 58, day 1, as the primary speaker spends time in `office', the error is almost constant near 8.7%. As soon as the primary speaker leaves his `office' and goes to `cafeteria', the cumulative error first reduces and then increases to about 17.6%. Next, when the primary speaker comes back to a relatively quiet place (`office'), the cumulative error again comes down to near 1.7%.

On day 2, as the primary speaker is in `office', the error is around 9.8%. When the primary speaker goes to `cafeteria', the error reduces and then when he goes to a relatively quieter place (`conference room' and `office'), the error again stabilizes around 8.7%. While he leaves the `office' for his `car' and then `cafeteria', the error reduces to about 1%.

On day 3, the primary speaker spends time in `office' and the error is around 9.45%. When the primary speaker goes to the `cafeteria', the error first reduces and then increases to near 8.7%. Again, when he comes back to `office', error reduces and then stables around 4%.

In the final experiment, calendar event information is combined with word count estimates using the common time dimension. We generalize specific calendar events into 9 general categories for ease of analysis, namely, (i) Conference Call, (ii) Faculty Meeting (primary speaker meets faculty in the department), (iii) Thesis Defense, (iv) Student Meeting (non-research discussion such as logistics, infrastructure, etc.), (v) Staff Meeting (administrative discussions), (vi) Teaching, (vii) Research Meeting

(research discussions with students and staff), (viii) Lab Meeting (large group meeting), (ix) Senior Design Meeting (mentoring discussion), and (x) Alone time (primary speaker is working by himself). Additionally, we used nine days (labeled Days A-to-I) for the analysis.

Fig. 59 shows the word count estimates for each day and event category. Additionally, cumulative daily counts are also shown for every day. Finally, cumulative event word counts across all days are also shown. Some interesting observations can be made based on this visualization. For example, Day H appears to be an outlier in terms of cumulative word count (which is very less compared to other days). On the other hand, most words were spoken on Days B, F and I. In terms of events, most words were spoken in research, student and faculty meetings. As one would expect, no words were spoken during alone time. Thesis defense and senior design meeting had lowest word counts as these events are relatively rare. In terms of similarity, Days A, F, and I seemed to have similar word count profiles. The primary speakers appears to have spent most time talking to students and staff about research on Day B. In summary, it is interesting to note that even a simple measurement such as word count when analyzed with readily available meta-data starts to tell a story.

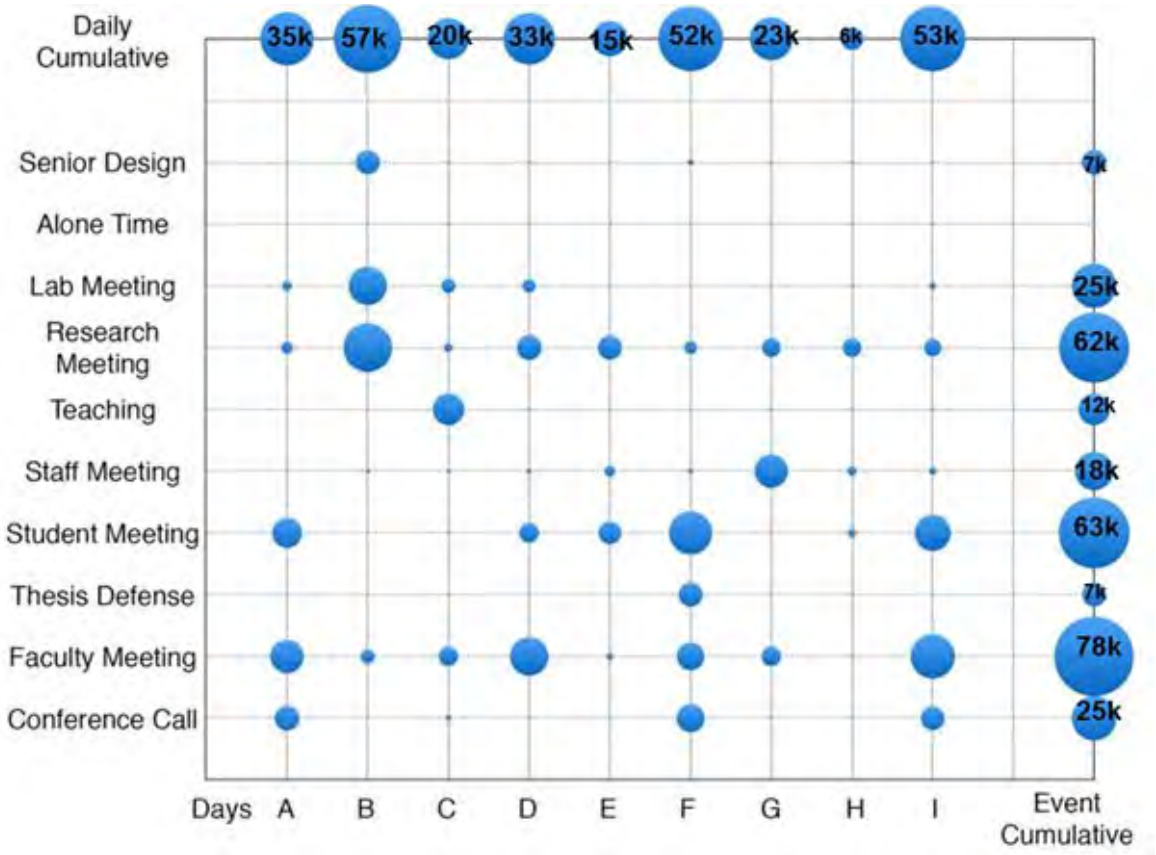


Fig. 59: *Comparison of word count measurements for 9 days and 10 event categories. Additionally, cumulative daily and event counts are also shown.*

In conclusion, it is noted that the system for word count estimation uses simple syllable rate measurement and LMMSE estimator to estimate word count. In the future, alternative features including prosodic and contextual based as well as much more sophisticated machine learning approaches such as Deep Learning could be studied in this area. In addition, new technique in enhancement and speaker diarization could also be a potential direction.

3.2.4: [Contribution 4] Leveraging KWS plus SAD, Environmental Sniffing, and Word Count:

Finally, in [Contribution 4], keyword spotting information is leveraged, in addition to the advancements from [Contribution 1], [Contribution2] and [Contribution 3], to address both detection and classification of speech interaction events such as participating in conference call or teaching in a classroom. Evaluation on the Prof-Life-Log task achieves a system accuracy of 82%.

The proposed method employs speech activity detection (TO-Combo-SAD) and speaker diarization systems to provide high level semantic segmentation of the audio streams. Subsequently, a number of audio, speech and lexical features are computed in order to characterize events in daily audio streams. The features are selected to capture the statistical properties of conversations, topics and turn-taking behavior, which creates a classification space that allows us to capture the differences in interactions. Our experimental results show that the proposed system achieves good classification accuracy on a difficult real-world dataset (i.e., Prof-Life-Log)

The proposed system consists of three major parts, audio pre-processing, feature extraction and classification. The workflow diagram of the system is shown in Fig. 60.

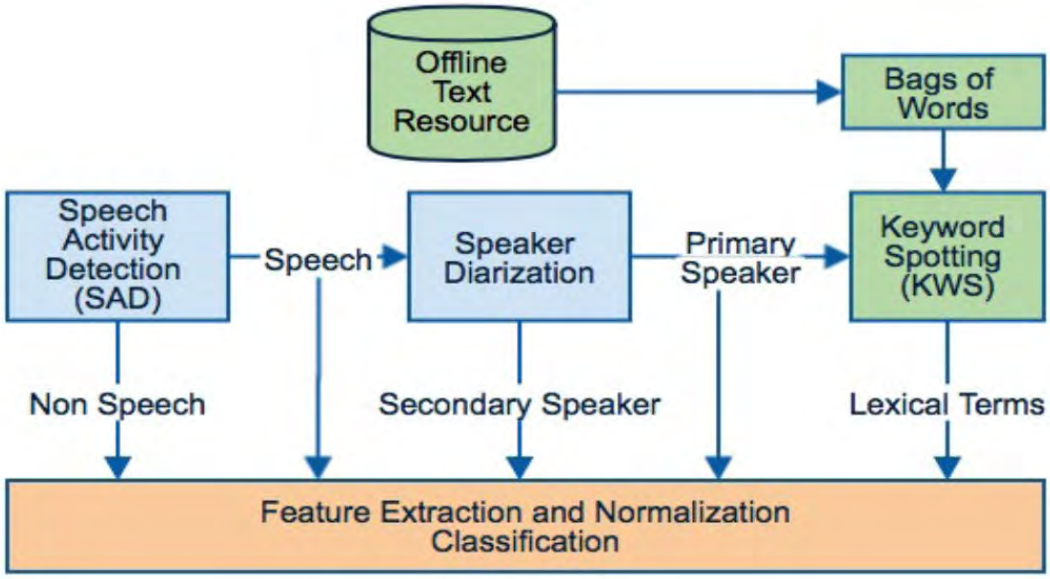


Fig. 60: Block diagram of proposed speech interaction diarization system.

For system evaluation, 5-fold cross validation was performed by splitting the dataset described before into 80% for training and 20% for evaluation. The performance results reported here are averaged over the 5 trials. In this experiment, the impact is analyzed in terms of dimension reduction on classification due to PCA. By varying the principal components used for classification from 1-to-13, it is possible to compute the corresponding values for performance accuracy on the 5-way classification task. Since the underlying events are unlikely to vary fast with time, temporal constraints are imposed on the SVM output decisions by employing median filtering (the SVM decisions are first chronologically ordered). Altogether, the measured impact of applying 3, 5, and 7 window median filter on raw SVM outputs are achieved, with performance results shown in Fig. 61. From the figure, it is observed that the classification performance first steadily increases as the PCA dimensions used for classification are increased, then plateaus out, and finally decreases slightly. The best performance is seen for the first 8 dimensions. Additionally, this trend is seen for all variants of temporal constraints that we applied on the data. Finally, it can be seen that the 5-point median filter seemed to work best, and corresponds to an overall accuracy of about 82%.

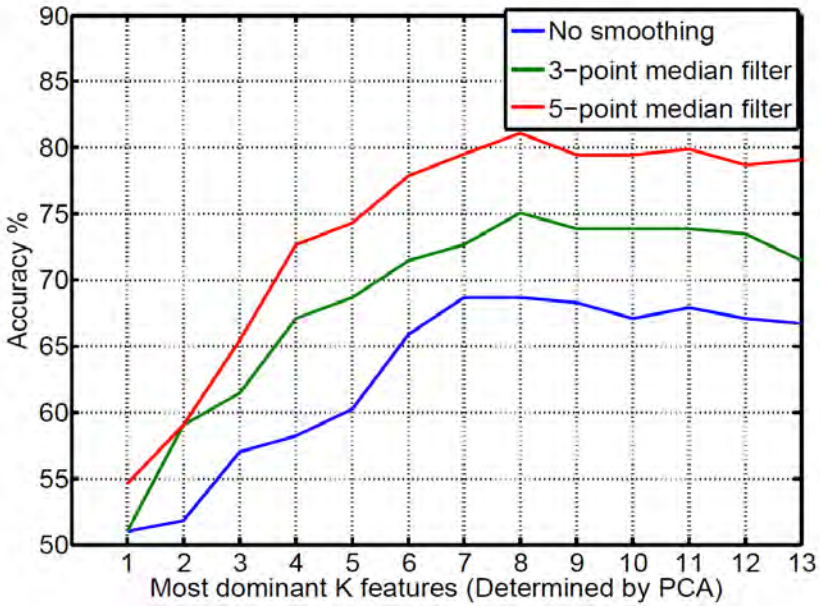


Fig 61: Performance accuracy for the proposed system.

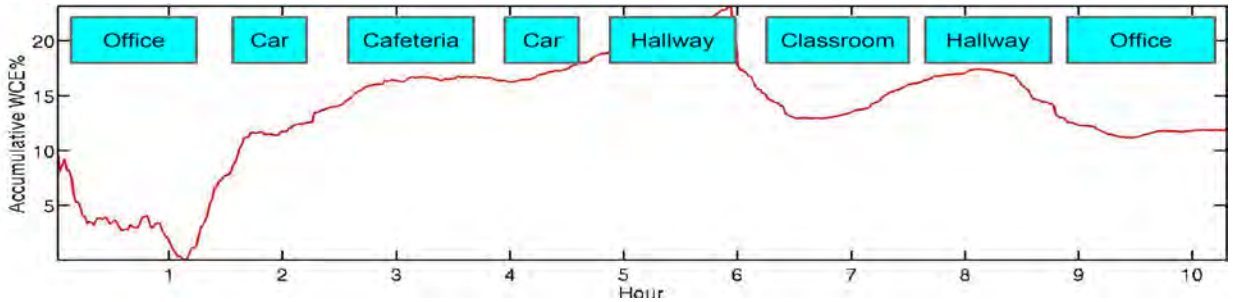


Fig 62: Sample Daily performance of Word Count Error over 12hour period.

For future work, the next direction would be to extract audio, speech and lexicon based features used in the system and correlate them with other higher level analysis. For example, it is possible to study the relation between these features for a full-day and stress level or behavior analysis of the person. Also, studying the relation of these features and educational analysis is another direction we are looking for. For example, how much can these SLT features estimate the student success in class? Human performance analysis is possible then by employing such analysis on personal audio streams.

3.2.5: References cited in this sub-section (3.2):

[1] A. Ziaei, L. Kaushik, A. Sangwan, J. H. L. Hansen and D. Oard, "Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions", ISCA, Interspeech 2014, pp. 1544-1548, September 2014

[2] S. Sadjadi, J. H.L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," IEEE Signal Processing Letters 20 (3), 197-200, March 2013

[3] M. Akbacak, J. H.L. Hansen, "Environmental sniffing: Noise knowledge estimation for robust speech systems," IEEE Transactions on Audio, Speech, and Language Processing 15 (2), 465-477, Feb. 2007

[4] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in Proc. IEEE ICASSP, Washington, DC, pp. 208–211, Apr. 1979.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. ASSP , vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[6] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," IEEE Trans. SAP , vol. 11, no. 6, pp. 700–708, Nov. 2003

[7] P. Scalart and J. Vieira-Filho, "Speech enhancement based on a priori signal to noise estimation," in Proc. IEEE ICASSP'96, Atlanta, GA, pp. 629–632, May 1996.

[8] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, S. Meignier, "An Open-source State-of-the-art Toolbox for Broadcast News Diarization," ISCA Interspeech'13, Lyon France, 25-29 Aug. 2013
[9] N. Morgan, E. Fosler-Lussier, "Combining multiple estimators of speaking rate," ICASSP'98, Vol. 2, pp. 729-732, 1998.
[10] D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," IEEE TASL Proc., 15(8), 2190-2201, 2007.
[11] N.D. Jong, T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," Behavior research methods, Springer, 2009.

3.3: Speech Analysis for Diarization – Some Additional Findings

In this section, some additional findings relating to speech and speaker analysis traits in the Prof-Life-Log corpus as massive daily audio streams is considered. References which contain further details are found at the end of this section.

- *Analysis of speaker and environment characteristics in Prof-Life-Log corpus*: the focus here has been on characterization of speakers and environments captured in the Prof-Life-Log database. *Environmental characteristics* are analyzed through long-term spectra and derived parameters (spectral center of gravity (SCG) and spectral energy spread (SES)). It is observed that SCG and SES consistently vary with varying environments and may benefit automatic environment detection (Figure 63).

- *Vocal effort as a function of environment*: the relation between the level/type of background noise and vocal effort is called Lombard function. This current study has found nearly linear relation between the noise level and vocal intensity in three conditions (Walking, Car, Cafeteria). Note that in low noise environments, the level of noise does not impact the intelligibility of speech communication and hence, small variation of the noise level does not induce strong vocal effort changes when compared to noisy environments. We have observed this phenomenon in our study - the slope of Lombard function is much smaller when transitioning between Office and Walking compared to Walking/Car/Cafeteria (Figure 64).

- *F_1-F_2 formant vowel space, fundamental frequency, pitch patterns, and speech rhythm* were studied. The analyses suggest that the acoustic-phonetic characteristics of speech production change when switching between environments. Somewhat surprisingly, even though having by nature unique physiological characteristics and talking manners, the primary and secondary speakers were found to choose the same strategies when altering their mean fundamental frequency and speech rhythm when switching between environments (Figure 65). It is not clear to what extent these changes are due to the mutual adaptation of the communication parties and to what extent due to the environmental characteristics. We assume it may be a combination of these two factors.

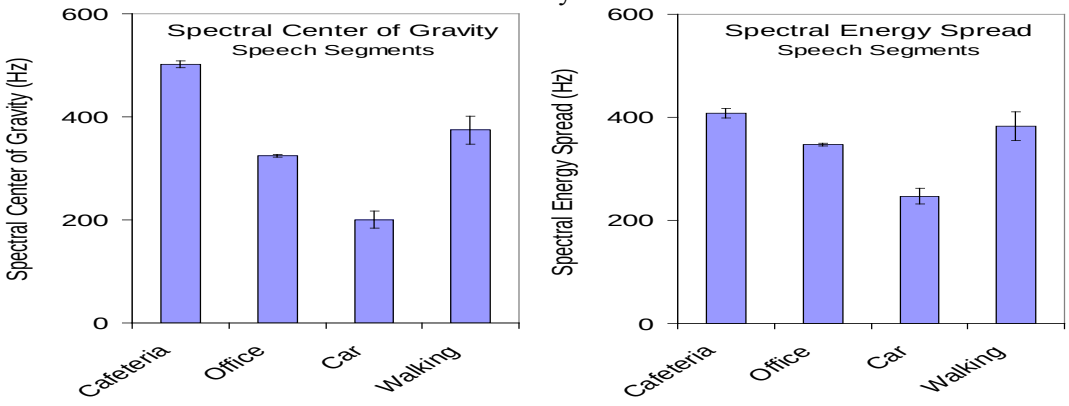


Figure 63: *Mean values of spectral center of gravity (SCG) and spectral energy spread (SES) in various environments. Error plots delimit 95% confidence intervals.*

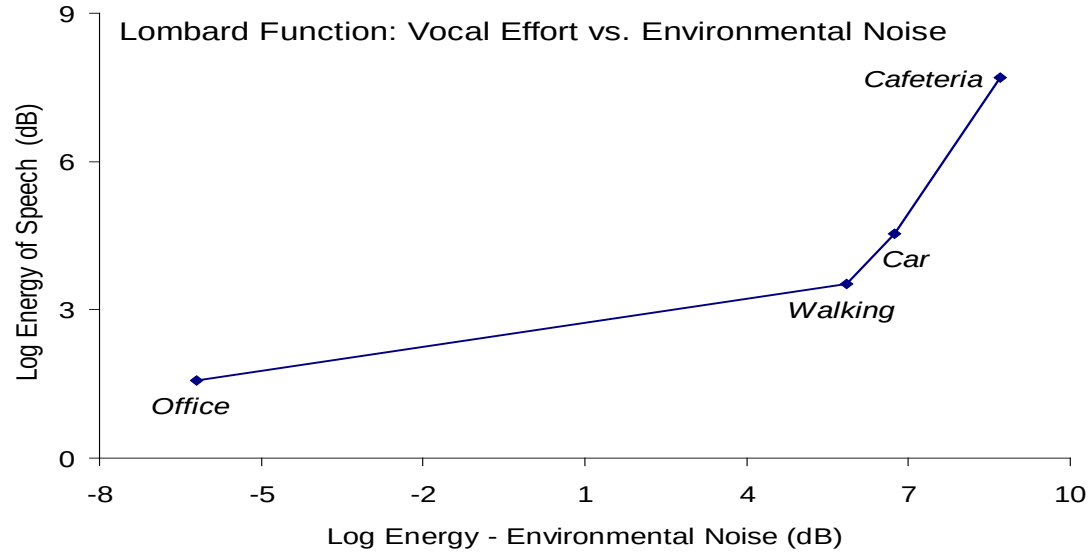


Figure 64: *Lombard function - relation between vocal intensity and type and level of environmental noise.*

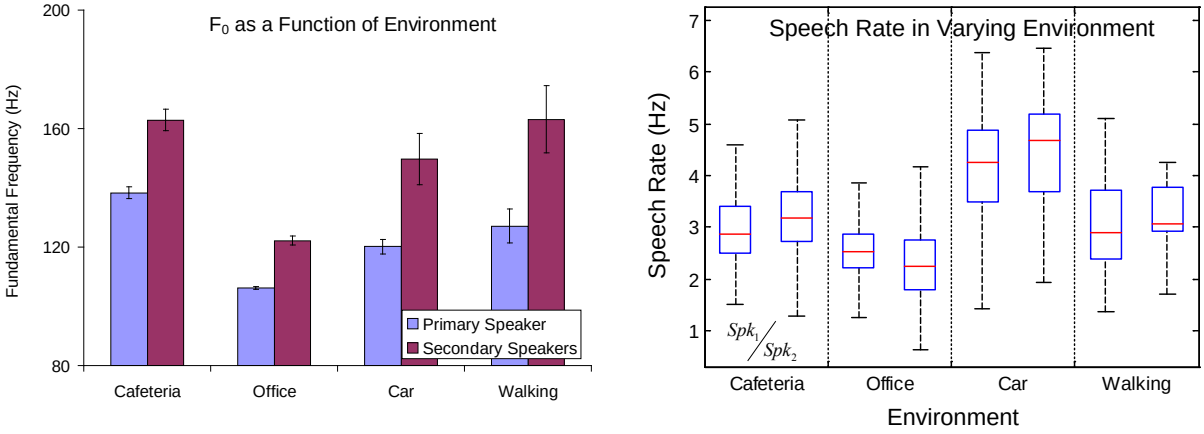


Figure 65: *Mean fundamental frequency (F_0) and speech rate/rhythm in primary and secondary speakers in changing environment.*

References for Sec. 3.3

[Boril13] Boril, H., Ziaei, A., Hansen, J. H. L. (2013). "Prof-Life-Log: Production of Conversational Speech as a Function of Varying Environment," in LENA International Conference, Poster Presentation, April 28-30 (Denver, Colorado).

[Hautamaki13] Hautamaki, V., Lee, K. A., van Leeuwen, D., Saeidi, R., Larcher, A., Kinnunen, T., Hasan, T., Sadjadi, S. O., Liu, G., Boril, H., Hansen, J. H.L., Fauve, B. (2013). "Automatic Regularization of Cross-Entropy Cost for Speaker Recognition Fusion" in Proc. of Interspeech'13, 1609-1613, August 25-29 (Lyon, France).

[Saeidi13] Saeidi, R., Lee, K. A., Kinnunen, T., Hasan, T., Fauve, B.,Bousquet, P.-M., Khoury, E., Sordo Martinez, P. L., Kua, K., You, C., Sun, H., Larcher, A., Rajan, P., Hautamaki, V., Hanilci, C., Braithwaite, B., Gonzales-Hautamaki, R., Sadjadi, S. O., Liu, G., Boril, H., Shokouhi, N., Matrouf, D., El Shafey, L.,Mowlaee, P., Epps, J., Thiruvaran, T., van Leeuwen, D. A., Ma, B., Li, H., Hansen, J. H. L., Bonastre, J.-F., Marcel, S., Mason, J., Ambikairajah, E. (2013). "I4U Submission to NIST SRE 2012: A Large-Scale Collaborative Effort for Noise-Robust Speaker Verification," in Proc. of Interspeech'13, 1986-1990, August 25-29 (Lyon, France).

[Liu13] Liu, G., Hasan, T., Boril, H., Hansen, J. H. L. (2013). "An Investigation on Back-End for Speaker Recognition in Multi-Session Enrollment," IEEE ICASSP'13, 7755-7759, May 26-31 (Vancouver, Canada)

[Hasan13] Hasan, T., Sadjadi, O., Liu, G., Shokouhi, N., Boril, H., Hansen, J. H. L. (2013). "CRSS systems for 2012 NIST Speaker Recognition Evaluation," IEEE ICASSP'13, 6783-6787, May 26-31 (Vancouver, Canada).
[Hasan12] Hasan, T., Liu, G., Sadjadi, S. O., Shokouhi, N., Boril, H., Ziaei, A., Misra, A., Godin, K. W., Hansen, J. H. L. (2012). "UTD-CRSS Systems for 2012 NIST Speaker Recognition Evaluation," NIST 2012 Speaker Recognition Evaluation Workshop, Dec. 11-12, 2012 (Orlando, Florida).

Task 4 – Automatic Speech Recognition/Keyword Spotting (ASR/KWS):

4.1: Keyword Spotting (KWS): Over this 36month period, an extensive effort has been undertaken to research Keyword Spotting (KWS) technology with the goal of building systems that operate efficiently and effectively on practical naturalistic audio data. In order to meet this goal, the strategy has been to measure progress against a variety of corpora such as Hub-5, NOTEL, Prof-Life-Log, Apollo Space Missions, UT-Opinion, YouTube, etc. The methods, algorithms and systems built by us during the course of this investigation have shown good performance across the mentioned variety of data. Figure 66 highlights the progress in KWS across various corpora over the past 36month period.

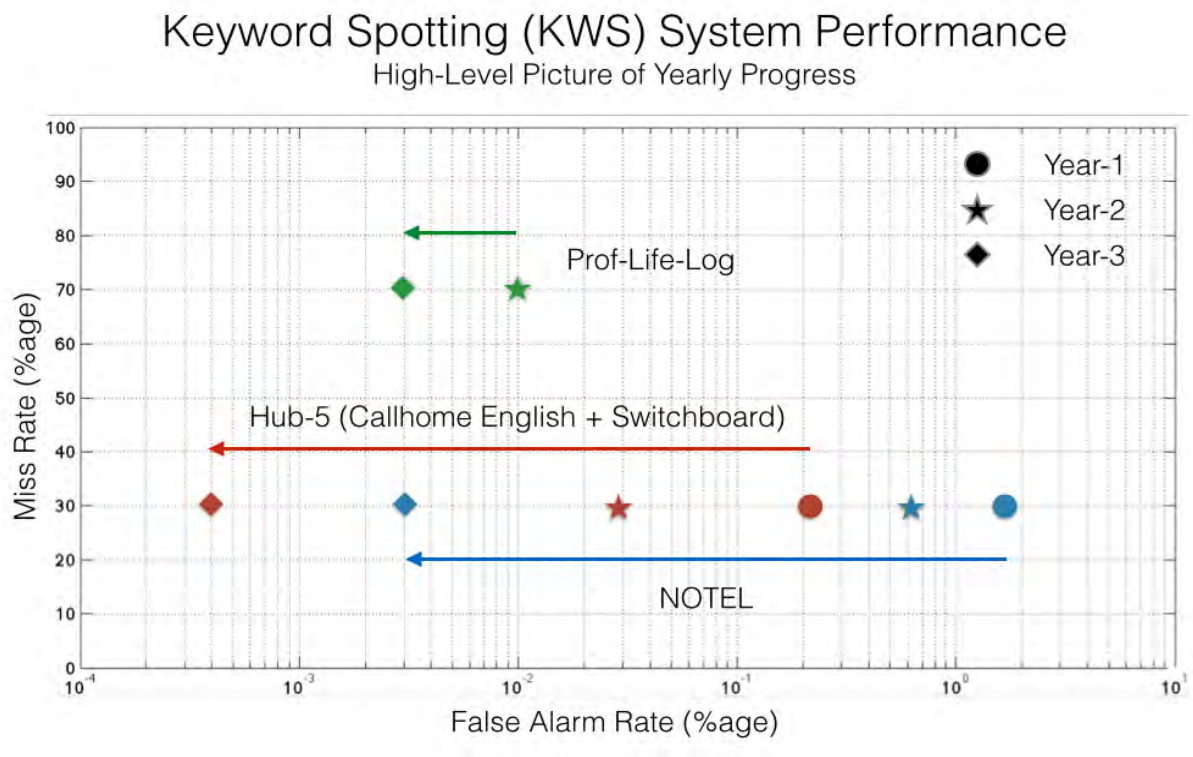


Figure 66: *KWS yearly performance improvements*

Figure 66 presents a high-level view of the performance of CRSS-UTDallas KWS systems on 3 corpora, namely, NOTEL (speech from non-native English speakers under mixed noise conditions), Hub-5 (standard ASR task in the research community), and Prof-Life-Log (continuous naturalistic audio data captured in real-world conditions). In the figure, the performance improvements have been captured on a year-to-year basis. In what follows, the major technical contributors towards improvement over this three year period are described. During this period, regular code deliveries have been transferred during periodic site visits.

4.1.1: KWS: Acoustic, Language, Keyword Search: a major emphasis has been to investigate acoustic modeling, language modeling and keyword search algorithms in order to improve KWS system

performance. In the automatic speech recognition (ASR) front-end, feature processing using LDA/MLLT (Linear Discriminant Analysis/Maximum Likelihood Linear Transform) has consistently improved both ASR (in terms of lowering word error rate) and KWS (lowering equal error rate) results. LDA/MLLT tries to project features into a discriminative space and reduce the feature dimensions, and this seems to help speech recognition (and KWS). Deep learning methods for feature processing such as bottleneck features have also proved to be helpful in reducing errors. Bottleneck features follow the same principle, (i.e., feature space reduction and projection into discriminative space).

For acoustic modeling, progressive decrease in error rates were shown as the system configuration/solution moved from Speaker Adaptive Training (SAT) using fMLLR to Subspace Gaussian Mixture Models (SGMMs). Additional improvement in performance was seen when moving to Maximum Mutual Information (MMI) criterion for training (MMI is a discriminative method for training). Finally, deep learning methods such as convolutional neural networks (CNNs) provide further gains on top of the best MMI systems developed by CRSS-UTDallas. The research has also been able to verify the positive impact of deep learning methods on both NOTEL and Hub-5 tasks. However, the measurements are yet to be made for Prof-Life-Log and other naturalistic corpora.

From the language modeling perspective, most improvements come from data augmentation for training. CRSS built web crawlers to scrape relevant textual data from the web to capture secondary knowledge to improve speech/language technologies. As an example, CRSS obtained close to 1billion words of text that closely match conversational telephony speech (CTS). Additionally, CRSS also has close to 1billion words of text that contains opinionated data such as reviews (downloaded from websites such as Glassdoor.com, Amazon.com, Tripadvisor.com, Hotels.com, etc.). Finally, CRSS also has close to 200M words of text for Indian English that was downloaded to support language model building for NOTEL task.

The combination of the mentioned approaches provided CRSS with nearly half the performance improvement shown in Fig. 66 on the NOTEL and Hub-5 tasks. Word error rates (WER) were obtained for corpora such as: Hub-5 switchboard task with about 20% WER, and NOTEL task with about 34% WER. The NOTEL WER is the best that CRSS has seen in the literature.

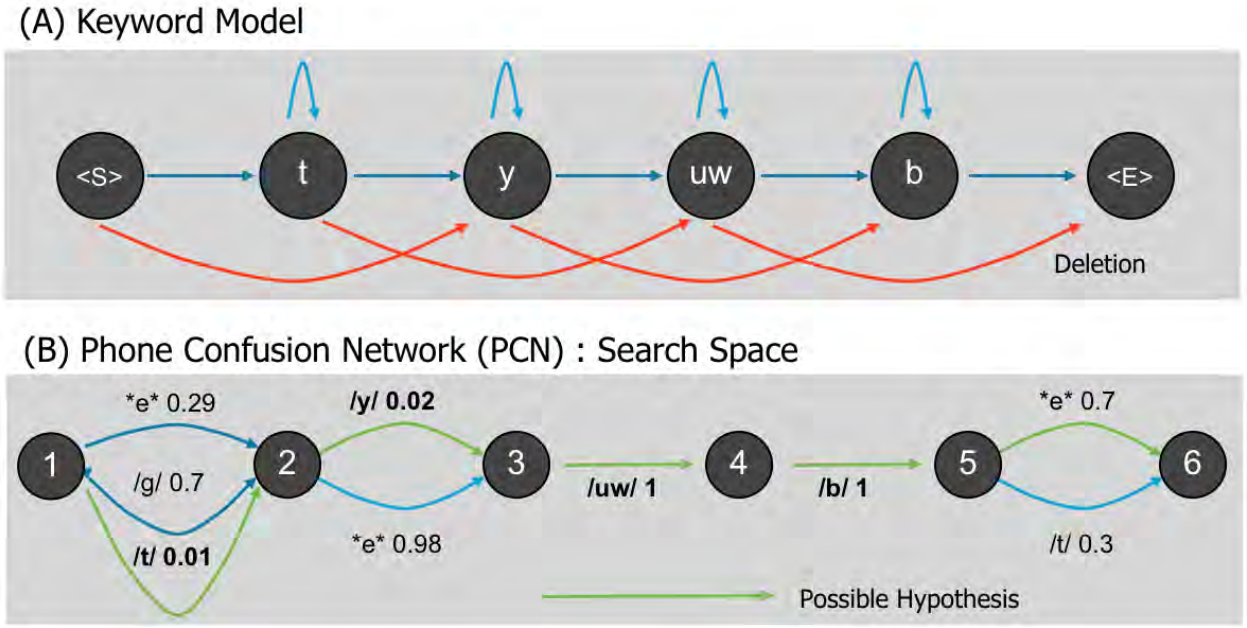


Figure 67: *(A) Keyword Model and (B) Searching for the keyword in Phone Confusion Network*

Keyword search techniques typically operate on speech recognition output. Word lattices are the most common starting point for KWS search techniques. Phone lattices are a good alternative as they offer more flexibility in the search space. They are especially useful for looking for out of vocabulary (OOV) words. On the other hand, phone lattices can be too flexible and result in large number of false-positives. One way to retain the flexibility of phone search space and avoid large number of false-positives is to first create ASR word lattice and then convert it into phone lattice. Since the speech recognizer first decodes speech into words, it avoids creating meaningless combinations of phonemes (as is often seen in phone decoding). In our experience, this method works best when used in conjunction with bi-gram language models. It seems like bi-gram provide the right level of "controlled-flexibility".

Unlike lattice based search methods, we proposed a new algorithm for searching phone confusion networks (PCN, see Fig. 67(B)). The method is called PCN-KWS. PCNs are typically computed from phone lattices. In PCNs, the recognition output is presented in form of a graph with a fixed number of nodes from start to finish. Several alternative paths are presented to move from one node to next. Each path captures a phoneme along with posterior probability (can be thought of as confidence that the phoneme was detected).

Figure 67(A) shows the model of a keyword used in PCN-KWS. The keyword is modeled as acyclic graph, and allows for phoneme deletion and insertion. The PCN-KWS algorithm attempts to search the keyword in the PCN structure. By allowing flexibility for phoneme deletion and insertion, the algorithm is more robust to ASR errors. We have observed that the PCN algorithm contains complementary properties to traditional word-lattice based search methods. For this reason, we have always experienced improved results when the search outputs for the two systems are fused. The word-based techniques are more rigid (hence, they have higher miss-rate but lower false-alarms). PCN-KWS is very flexible (so it has lower miss-rate but higher false-alarms). The combination of the two methods tends to provide the best of both worlds.

Figure 68 illustrates this point in context of a KWS task for Prof-Life-Log. In this task, we employ a word-based search technique that uses FST (it is useful to note that the FST algorithm is distributed with Kaldi) along with PCN-KWS algorithm. Additionally, we divide the keywords to be searched into two groups, in-vocabulary terms (IV keywords, i.e., these terms are modeled by ASR lexicon and language model), and out-of-vocabulary terms (OOV, i.e., these terms do not appear in the ASR lexicon and language model). We run the FST and PCN-KWS search for IV and OOV terms, and the results are captured in Fig 3. It is also useful to note that due to the nature of Prof-Life-Log (PLL) data, even the IV terms are not very strongly modeled in the ASR lexicon and language model (as PLL contains a variety of topics and naturalistic conversational spontaneous speech). From the figure, it is seen that the rigid FST method shows high miss-rate for both IV and OOV terms (in fact, the DET curve stops abruptly and the full range of performance cannot be obtained). On the other hand, the PCN-KWS delivers a full range of performance. As expected, the performance for IV terms is better than OOV terms. In summary, PCN-KWS is more flexible and indispensible when working in open-ended domains (such as Prof-Life-Log). It is also useful to note that on other tasks such as Hub-5, the gap between IV and OOV performance tends to be much larger.

A key learning from the previous experiment is that modeling keywords in the ASR lexicon and language model can deliver the largest improvements in operation. This is a manageable tasks if the keywords are known a~priori (for example, the DARPA RATS program ran with the same assumption that the list of keywords were available upfront). In another experiment, we tried to investigate a practical method of exploiting the knowledge of the keyword list in building a KWS system for Prof-Life-Log (PLL). PLL contains lots of conversational data that can be classified as research vs. non-research discussions. We ask the question that can we automatically classify conversations into research vs. non-research using keywords. This method involves KWS where we search every 5 minutes of audio for research keywords and make a classification decision based on the density of the detected keywords. This task is especially interesting because research terms tend to be mostly OOV (being specialized vocabulary). In an effort to build this system, we downloaded all technical papers from 3 Interspeech conferences, and extracted the

text. Using part-of-speech (POS) tagging, we identified all noun terms (which usually are good candidates for keywords). Furthermore, word frequency information was used to further prune this list and come up with a final list of technical terms. This technical term list is now treated as our keyword list. Additionally, the text from Interspeech was used to build a language model that was then interpolated with CRSS's background language model to deliver a new hybrid language for the ASR. In this manner, potential keywords that were more likely to appear in research discussion within PLL, were modeled in the ASR/KWS system.

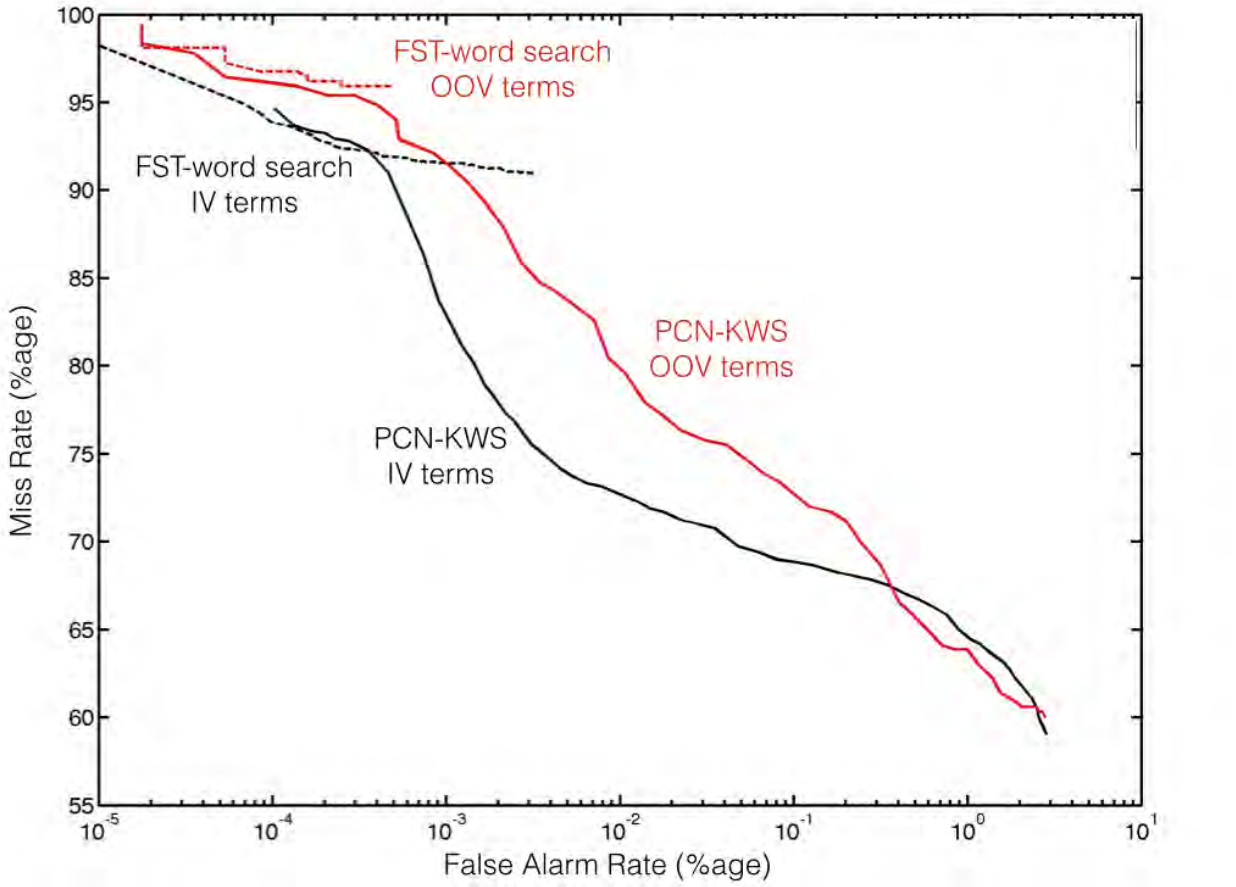


Figure 68: *Comparing relative benefits of FST-based word search and PCN-KWS when looking for in-vocabulary and out-of-vocabulary terms in Prof-Life-Log.*

Figure 69 shows the KWS performance for this task. The system was employed separately using (i) the background LM (build for CTS data) and (ii) hybrid LM (interpolation of Interspeech language model and background language model). The improvements are shown for common keywords (terms that are of non-technical nature), technical keywords (terms extracted from Interspeech and chosen as research keywords) and for a combination of non-technical and technical keywords in Fig. 69. The performance for technical keywords increased as a result of the modeling effort. Interestingly, the performance on non-technical (common) keywords also increased. This result seems to suggest that more accurate modeling of the language can provide a general boost in performance.

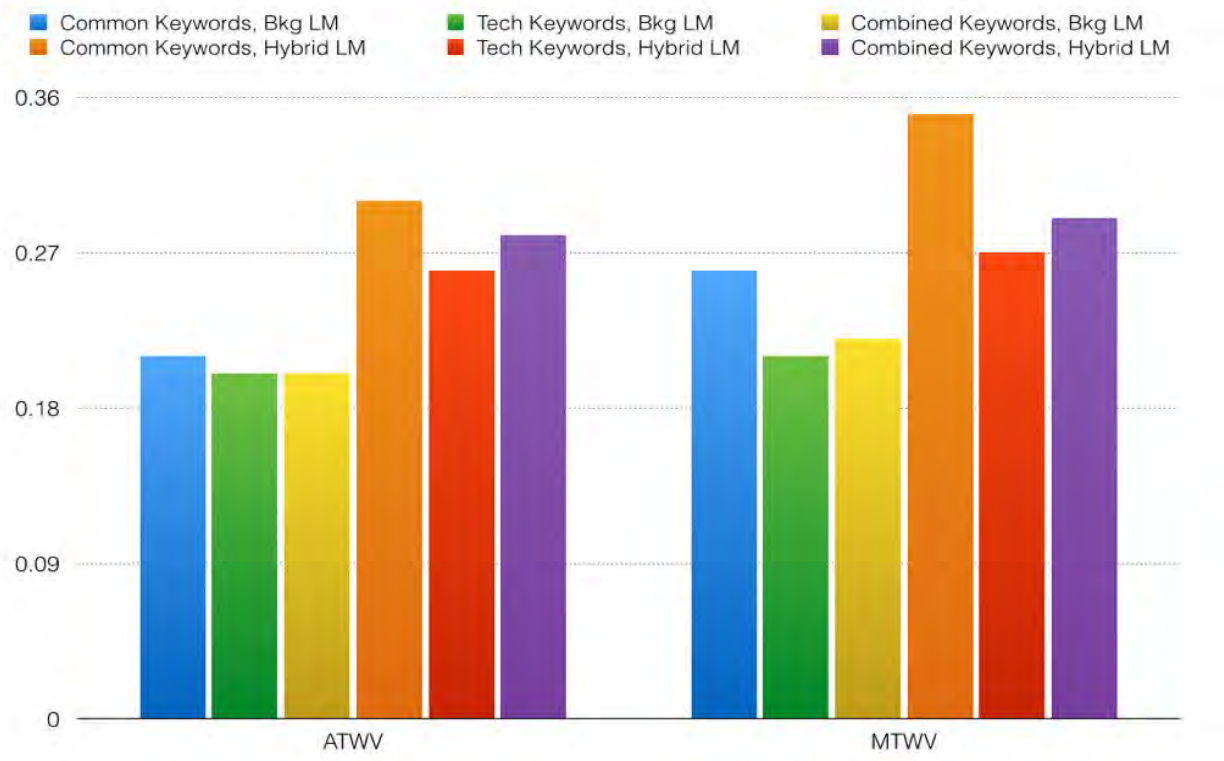


Figure 69: *Average Term Weighted Value (ATWV) and Maximum Term Weighted Value (MTWV) for Prof-Life-Log Technical terms search task.*

4.1.2: KWS Applications:

CRSS has extensively used the developed KWS system to drive sentiment detection in audio. Most existing methods for audio sentiment analysis use automatic speech recognition (ASR) to convert speech to text, and feed the textual input to text-based sentiment classifiers. Our study shows that such methods may not be optimal, and we propose an alternate architecture where a single keyword spotting system (KWS) is developed for sentiment detection. In the new architecture, the text-based sentiment classifier is utilized to automatically determine the most powerful sentiment-bearing terms, which is then used as the term list for KWS. In order to obtain a compact yet powerful term list, a new method is proposed to reduce text-based sentiment classifier model complexity while maintaining good classification accuracy. Finally, the term list information is utilized to build a more focused language model for the speech recognition system. The result is a single integrated solution that is focused on vocabulary that directly impacts classification [5,6].

The new method exploits the fact that the lexical evidence for sentiment in spoken comments is sparse and depends on a relatively smaller focused vocabulary. Accurate sentiment detection relies on a small fraction of the speech recognition transcript, because sentiment-bearing vocabulary tends to be sparse in spoken opinions. For example, in a statement like “I ordered a pepperoni pizza last night and it was wonderful”, only 1 out of 11 word conveys sentiment. While this may not be true for every comment, sparseness is generally prevalent in spoken comments. Given this nature of spoken comments, it would be reasonable to assume that sentiment detection is tolerant of high word error rates (WERs). In other words, sentiment detection accuracy depends on being able to reliably detect a very focused vocabulary in the spoken comments. Therefore, keyword spotting (KWS) technology seems to be better suited for sentiment detection, as opposed to full-transcript ASR.

In order to build an effective KWS system, a compact yet effective keyword list is first needed. The textual

features extracted by most text-based sentiment classification system are a good starting point to generate a keyword list. However, the learning paradigm for most of these systems tends to be greedy and generates a very large number of features. To mitigate this problem, an iterative technique is proposed that can reduce the feature size (and consequently model complexity) without significantly sacrificing performance accuracy. Additionally, the term list is incorporated in the speech recognition language model to assist in better KWS by ensuring none of the vital terms is OOV (out of vocabulary). The mentioned innovations deliver a single integrated system.

The proposed solution is evaluated on videos from YouTube.com and UT-Opinion corpus (which contains naturalistic opinionated audio collected in real-world conditions). Experimental results show that the KWS based system significantly outperforms the traditional architecture in difficult practical tasks. Figure 4-5 shows the sentiment detection performance for the proposed KWS based system and compares it to the traditional ASR approach. As seen in the figure, the ASR approach is barely better than a random system (which guesses negative vs. positive sentiment for the audio file). The EER (equal error rates) for YouTube and UT-Opinion datasets are shown (the entire DET curve for UT-Opinion is also shown). We obtain 10% and 32% EER for YouTube and UT-Opinion, respectively. Our YouTube data contains relatively longer audio files when compared to UT-Opinion corpus, and the sentiment tends to expressed more strongly (negative or positive). This makes UT-Opinion data more challenging, which can be seen in the results. Overall, KWS proves to a more viable method for addressing sentiment detection than ASR alone. It is also possible that this experience may extend to a variety of other tasks – here, at least one more case is considered based on detecting research vs. non-research discussion in Prof-Life-Log.

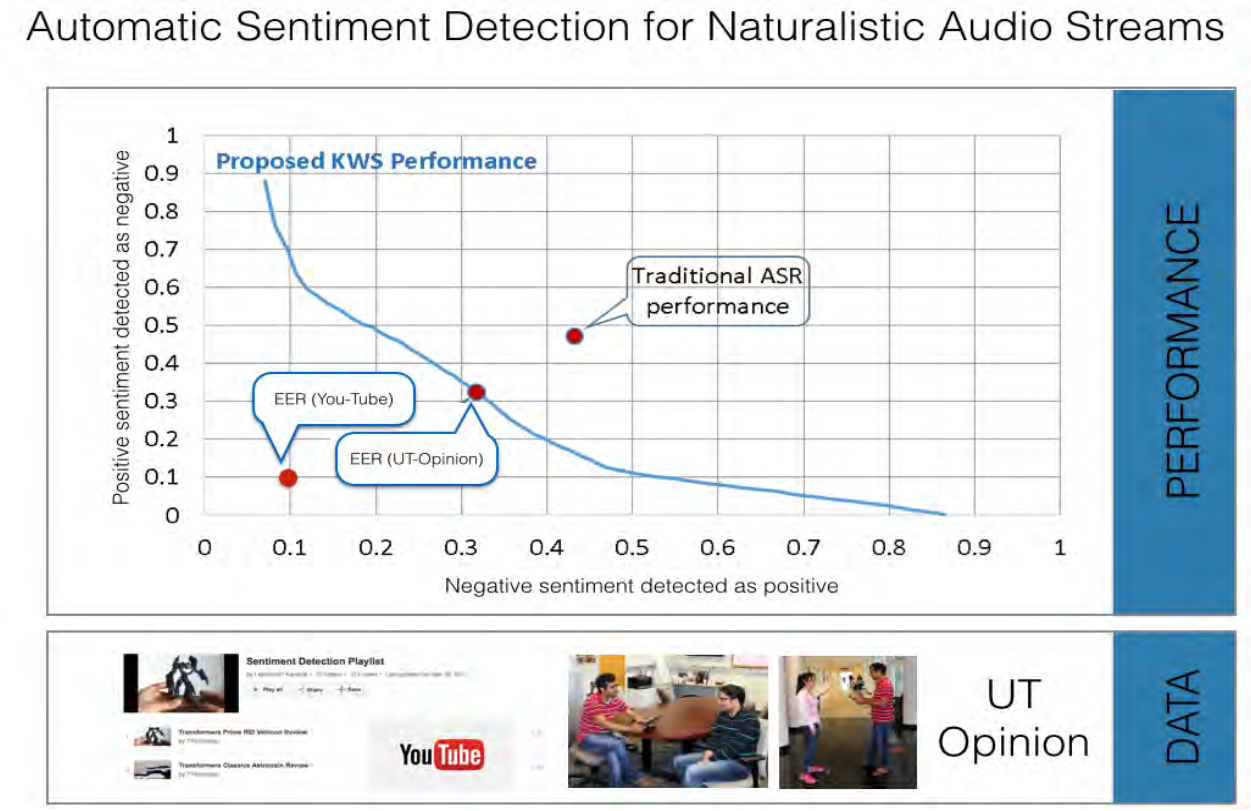


Figure 70: *Sentiment Detection Performance. System uses Keyword Spotting to detect sentiment-bearing terms in the audio.*

Another interesting dataset that we have analyzed using KWS is the NASA Apollo 11 space mission. The audio data is very challenging in the sense that it contains time-varying environmental noise, channel distortions, specialized vocabulary and naturalistic spontaneous conversational speech [8,9,10,11]. There are several research tasks that can benefit from KWS in the NASA corpus, (e.g., document linking, information retrieval, sentiment detection, topic monitoring, behavior monitoring, etc.).

An experiment is describd that was performed using the lunar landing part of the Apollo 11 space mission. The objective was to perform conversation analysis (and KWS system was used for this purpose). Before we can pass the audio data to KWS, the data is prepared using SAD (Speech Activity Detection) and Speaker Diarization algorithms. Data preparation separates speech from noise, and further segregates speech for the 4 speakers (3 astronauts and capsule commander), and sets the stage for conversation analysis. In the next step, the data is sent to the ASR system, and the output lattices are processed by FST-based word search and PCN-KWS algorithms. The sentiment-bearing terms are generated offline using a large repository of opinionated data such as reviews. By searching for these terms in the audio file, each 2-minute chunk of data is assigned a positive vs. negative sentiment rating. The information is successively aggregated to yield results for bigger chunks of time. Additionally, keywords are also used to detect topics. Simultaneously, other analyses such as word count, conversation turn-taking, etc. are also performed. In this manner, a rich repository of information is collected for every 2-minute interval of time.

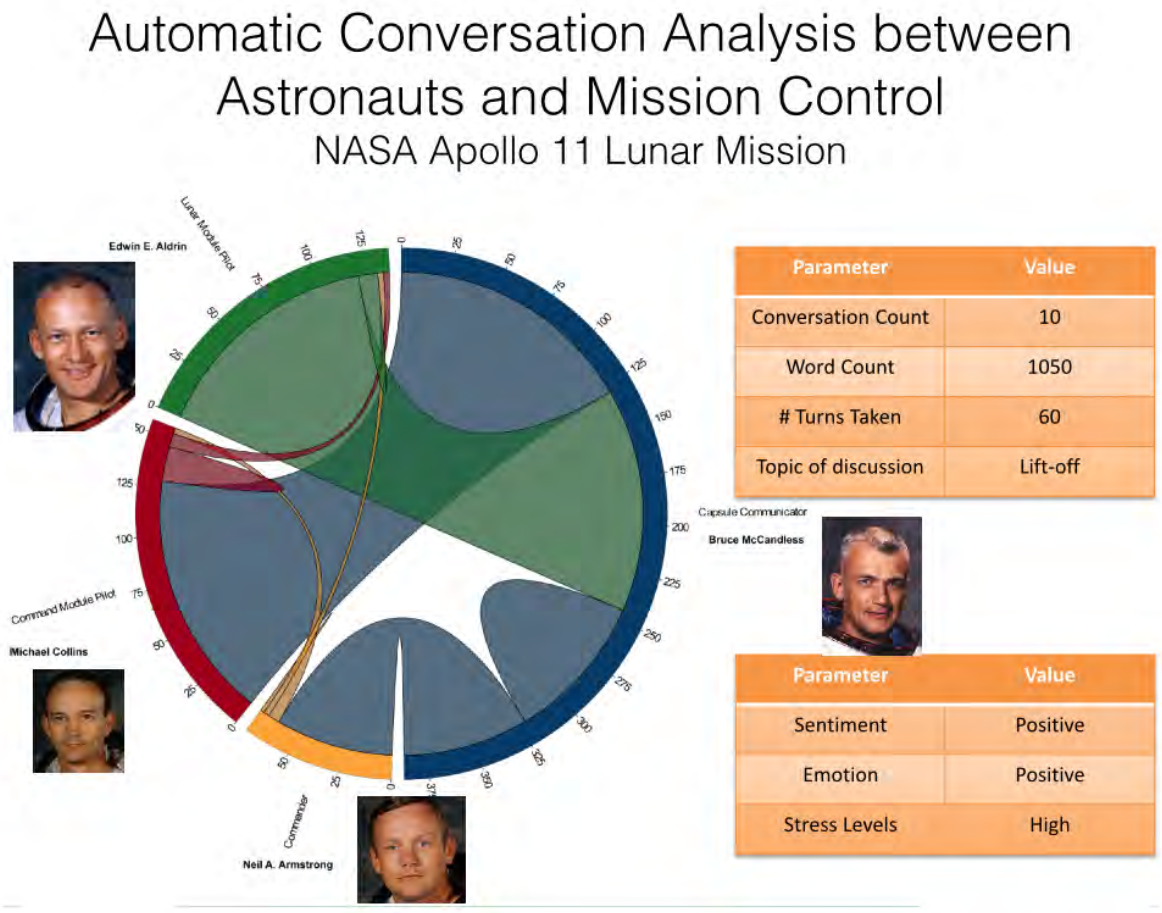


Figure 71: *Analysis of NASA Apollo 11 Space Mission conversations between astronauts and capsule commander. Analysis illustrates turn-taking, word count and sentiment detection.*

Figure 71 shows the graphical representation of the analysis performed for a sample 2-minute data chunk. The chord diagram illustrates globally the volume of directed conversation between any 2 individuals. Additionally, the overall sentiment level during this time is also noted (table summary to the right). Other statistics such as word count, conversation count, etc. are also shown. In summary, the power of KWS in being able to support sophisticated analysis of this nature has been demonstrated. This solution has also been delivered to AFRL during regular meetings.

4.1.3: KWS Publications from 36month effort:

[1.] A. Sangwan and J.H.L. Hansen, "Automatic Analysis of Mandarin Accented English using Phonological Features," Speech Communication, Vol. 54, No. 1, 2012, pp. 40-54

[2.] F. Williams, A. Sangwan and J.H.L. Hansen, "Automatic Accent Assessment Using Phonetic Mismatch and Human Perception," IEEE Transactions on Audio, Speech and Signal Processing, Vol. 21, No. 9, 2013, pp. 1818-1829.

[3.] T. Hasan, H. Boril, A. Sangwan and J.H.L. Hansen, "Multi-modal highlight generation for sports videos using an information-theoretic excitability measure," EURASIP Journal on Advances in Signal Processing, 2013.

[4.] T. Hasan, H. Boril, A. Sangwan and J.H.L. Hansen, "A multi-modal highlight extraction scheme for Sports Videos using an Information-Theoretic measure," ICASSP, 2012.

[5.] L. Kaushik, A. Sangwan and J.H.L. Hansen, "Sentiment Extraction from Natural Audio Streams," ICASSP, 2013.

[6.] L. Kaushik, A. Sangwan and J.H.L. Hansen, "Automatic Sentiment Extraction from YouTube Videos," Workshop on Automatic Speech Recognition and Understanding (ASRU), 2013.

[7.] A. Sangwan, H. Boril, T. Hasan and J.H.L. Hansen, "A Multimodal System for Automatic Sports Highlights Generation: Getting the good parts you want to your mobile platform with limited bandwidth!," Workshop on Spoken Language Technology (SLT), 2014.

[8.] A. Sangwan, C. Yu, L. Kaushik, A. Ziaei and J.H.L. Hansen, "Speech Processing Technology for Human Behavior and Performance Monitoring: Recent Algorithmic Advancements, Solutions and Challenges," NASA Workshop on Human Research Program (Integrated Pathway to Mars), 2015.

[9.] A. Kline, Z. Terlizzesse, K. Schrader, R. Pabba, L. Kaushik, A. Sangwan and J.H.L. Hansen, "Apollo Archive Explorer: An online tool to explore and study space missions," NASA Workshop on Human Research Program (Integrated Pathway to Mars), 2015.

[10.] D. Oard, A. Sangwan and J.H.L. Hansen, "Reconstruction of Apollo Mission Control Activity," First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH), 2013.

[11.] J. Malionek, D. Oard, A. Sangwan and J.H.L. Hansen, "Linking Transcribed Conversational Speech," 36th International ACM SIGIR Conference on Research and Development of Information Retrieval, 2013.

4.2: Distance Based Advancements for ASR:

In this area, the focus has been to develop strategies which can address reverberation and noise stemming from a single microphone based distance capture of speech. The idea is that the individual would be speaking normally, and not specifically directing his/her speech to the distance microphone (i.e., this would suggest an altered speech pattern to direct ones speech to the microphone; but instead to simply capture speech from an individual who would be speaking normally at a distance).

Recent advancements in Automatic Speech Recognition (ASR), particularly in the area of deep learning, has resulted in ASR systems achieving improved accuracy which is beginning to enable their use in many real-world applications. However, this gain has primarily been seen only for close-talking recordings. The error rates of ASR systems when using distant-talking (far-field) microphones remains high, usually twice as high as the close-talking error rates. Therefore, in this area the focus has been on solutions for improving the robustness of distant speech recognition systems.

Major challenges in far-field ASR include:

- Room Reverberation (with either no knowledge or knowledge of room impulse response)
- Additive environmental noise (possibly non-stationary; multi-tiered noise, etc.)
- Speaker movements and head orientation (potentially rapid changes in direction of speech)

Room reverberation, in particular, is the primary challenge in far-field ASR which acts as non-stationary and non-Gaussian noise which is correlated with the desired speech signal. This renders most conventional noise-robustness approaches ineffective for handling reverberation.

Existing approaches for far-field ASR can be categorized in three broad groups:

1. Signal or feature enhancement
2. Robust feature extraction
3. Model adaptation or improved Acoustic Model development

4.2: Distance Based ASR: Front-End & Backend Solutions

4.2.1: NMF-based front-end processing

Using a simple but effective model of reverberant speech in the magnitude spectrum domain, a multichannel front-end processing approach has been proposed using two variants of nonnegative matrix factorization (NMF) [1,2], which achieves considerable robustness in highly reverberant and noisy conditions. The reverberant speech model describes each filterbank output as a convolution between the clean speech subband signal and the time-frequency envelope of the room impulse response (RIR). Figure 72 shows the overall multichannel front-end used for far-field speech recognition.

A voice activity detection (VAD) unit initially distinguishes between speech and silence (noise-only) frames in each channel. If the frame is tagged as non-speech (i.e. noise), a spectral representation of the frame is used to update a noise dictionary. If the frame is tagged as speech, Sparse Decomposition (SD) by supervised NMF is performed on its magnitude spectrum using elements of both speech and noise dictionaries as fixed bases. While the noise dictionary elements are collected online to account for non-stationary environmental noise, the speech dictionary is created in advance using a speech database. This stage suppresses the noise component in each channel. The noise-enhanced spectral representations are then jointly processed by a convolutive nonnegative tensor factorization (CNTF) algorithm [2], which decomposes the magnitude STFTs into the underlying clean speech component (common between all channels) and the RIR components of each individual channel. This estimate of the clean speech spectrogram is then used to extract features for ASR.

Although the described algorithm is a multichannel solution, it is fundamentally different from conventional multichannel processing algorithms (i.e. beamforming solutions) in that it operates on the magnitude spectrogram and is independent of signal phases. This has the advantage of eliminating the need for speaker location (which is difficult to estimate in a reverberant environment), and also enables the algorithm to work on a distributed array with arbitrary random locations of the microphones. There are a number of scenarios where the microphone pickups can be in random or distributed but not equal locations.

Figure 73 shows the performance of the proposed front-end for robust ASR on the DIRHA-GRID corpus. This is a multi-room, multichannel small-vocabulary corpus that has been collected in an apartment with four different rooms, and with significant nonstationary environmental noise and very high T60 values. It is clear that the proposed CNTF solution has a significant impact on ASR based WER, and further improvement is achieved when MLLR adaptation is also included. Increasing the number of processing channels from 1 to 2 or 4, always provides gains, but the gain is even more significant with CNTF processing incorporated.

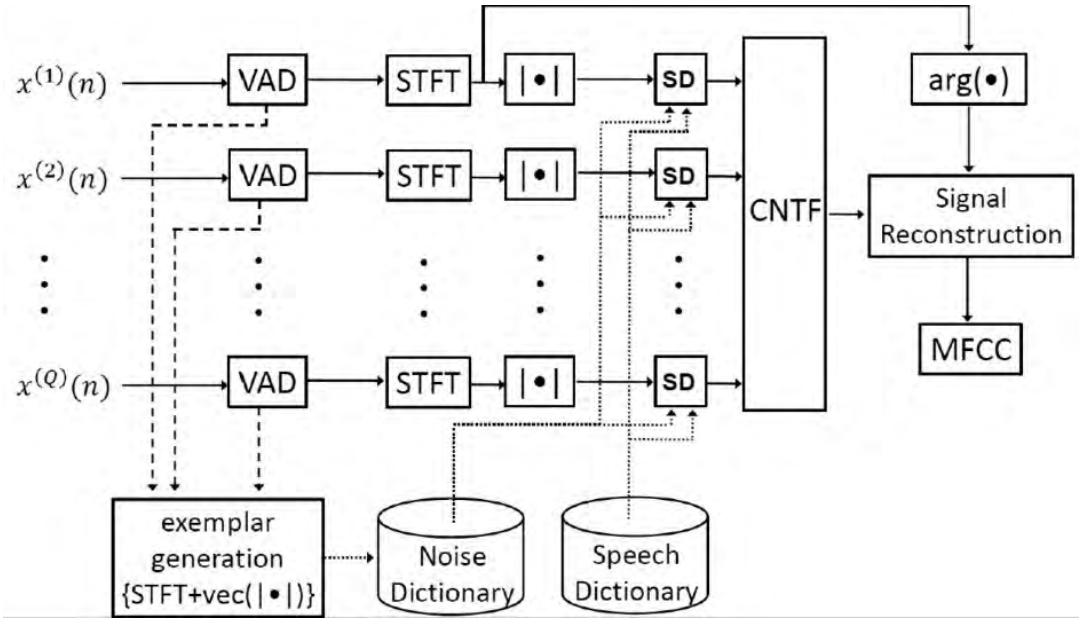


Figure 72: *Overall multichannel front-end for robust far-field ASR*

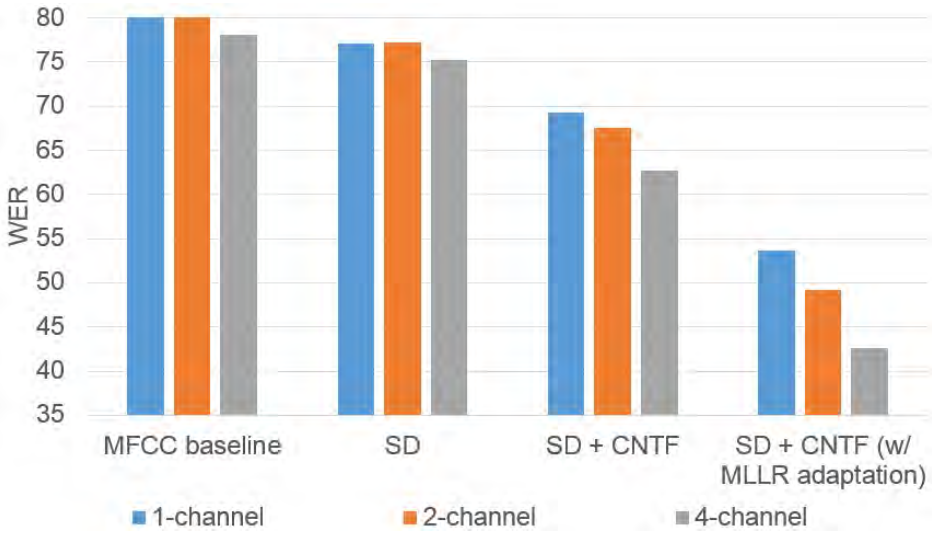


Figure 73: *Results of ASR experiments on the DIRHA-GRID corpus*

4.2.2: DNN-based front-end processing

The use of deep learning in speech recognition systems has resulted in significant improvements in accuracy. There are three major ways to use a DNN in an ASR system:

1. DNN as a feature transformation from noisy/reverberant frames to corresponding clean frames.
2. Generating new robust features (tandem and bottleneck features).
3. DNN as acoustic model in the back-end (hybrid approach).

Figure 74 demonstrates the use of DNN process as a feature transformation in the front-end.

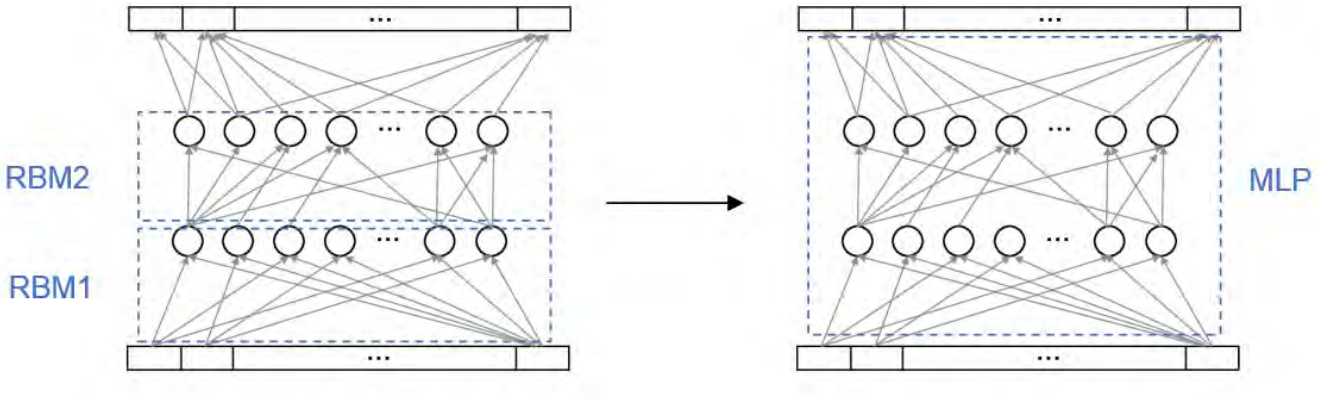


Figure 74: *Use of a DNN for feature transformation in the ASR system front-end*

The experiments indicate that if adequate synchronous noisy/clean speech signals are available for training (stereo training data), using such a deep network in the front-end can provide significant improvements in accuracy, outperforming most other front-end enhancement strategies.

The following Table 28 demonstrates the power of DNN transformations in reverberation-robust ASR for a medium vocabulary task (Note: here the RIRs from the Aachen impulse response (AIR) database have been used to contaminate TIMIT sentences with reverberation at different T60 values).

Table 28: *Distance Based ASR using CNMF and DNN processing for TIMIT data corrupted by Aachen Impulse Response (RIR) scenarios as a reverberant room corpus*

	Room1 (T60=800ms) d=2.25m	Room1 (T60=800ms) d=4m	Room1 (T60=800ms) d=7.1m	Room2 (t60=850ms) d=3m	Room3 (t60=230ms) d=1.9m
Baseline	**37.3**	**51.8**	**63.5**	**64.6**	**15.0**
CNMF [2]	**19.7**	**28.6**	**39.5**	**40.7**	**12.4**
DNN	**14.2**	**17.0**	**24.3**	**24.2**	**12.7**

A rather surprising observation was the network's great ability to generalize to completely unseen RIRs. The blue column in Table 28 indicates the RIR using which the DNN was trained. None of the other RIRs have been seen in the training phase. However, the DNN is able to achieve similar results for other RIRs in the same room and even for RIRs in other rooms with completely different acoustic characteristics.

4.2.3: DNN-based Acoustic Modeling

The emergence of deep neural network (DNN) acoustic models for ASR has now changed the directions of research in far-field ASR. A DNN-HMM acoustic model shows considerable inherent robustness to variations in the input data, thus automatically achieving a degree of robustness which minimizes the significance of many previous front-end approaches. The following Figure 75 shows the basic steps in training a DNN-HMM acoustic model (hybrid approach).

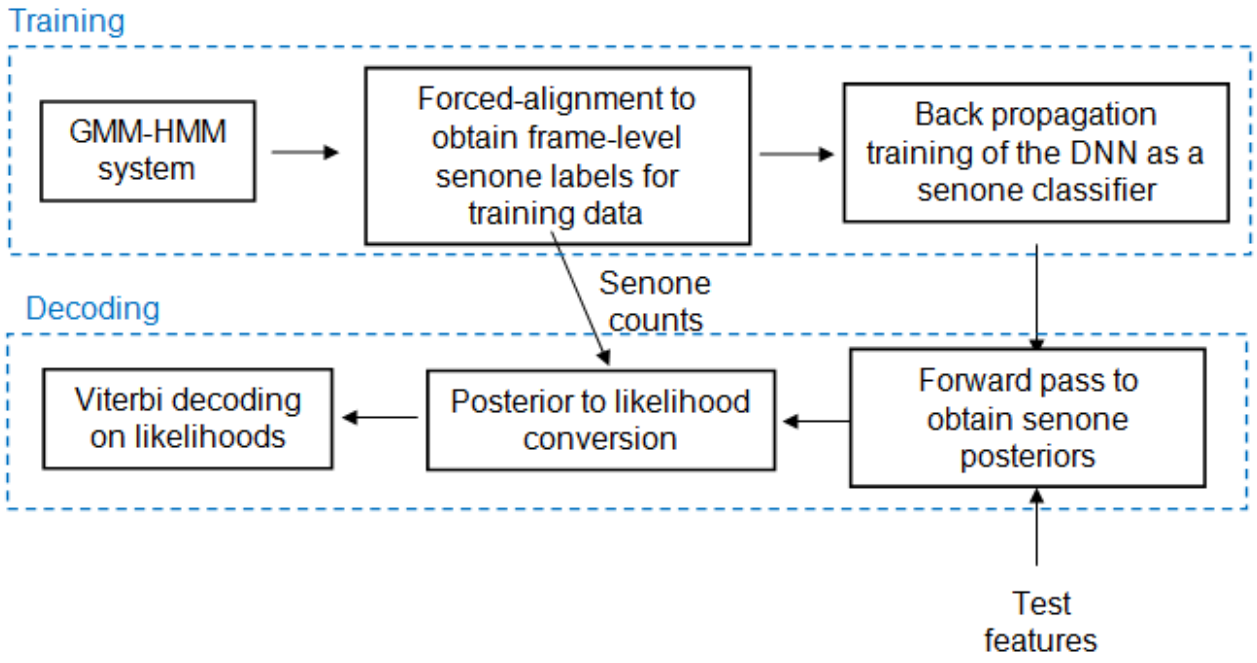


Figure 75: *Basic steps in Training/Decoding for the DNN-HMM acoustic model (hybrid approach)*

An evaluation of the proposed DNN-HMM system performance in an LVCSR task on the Fisher corpus was performed. A 150 hour subset of the Fisher corpus was used to train a DNN with 6 layers each containing 1024 nodes. The following table shows WERs from recognition experiments on a 5 hour test subset from the Fisher data. All experiments are performed using in-house DNN training tools developed at CRSS (together with the Viterbi decoder from Kaldi).

Table 29: *WER recognition performance using FISHER corpus with GMM-HMM baseline and the proposed DNN-HMM solution.*

Acoustic model	*Type of feature processing*	*WER (%)*
GMM-HMM	Raw MFCC	51.0
	LDA	47.9
	LDA+SAT (speaker-independent decode)	44.9
	LDA+SAT (decode w/ fMLLR adaptation)	36.1
DNN-HMM	Raw MFCC (speaker-independent decode)	34.2

The results indicate that even though a GMM-HMM system can benefit from different types of conventional feature processing including Linear Discriminant Analysis (LDA), speaker adaptive training (SAT), and decoding using fMLLR, a simple DNN-HMM system without any such processing achieves better performance with only raw MFCCs and without any speaker-specific processing. Note that further improvement can be achieved by training a DNN on LDA+fMLLR features from the GMM system. Here assessment is only on the performance of the DNN on raw MFCCs independently (without any GMM-dependent feature processing).

Experiments have also been performed on the reverberant and noisy data from the IARPA ASPIRE[1] challenge. The data consists of 5 hours of speech from 30 different speakers, collected using different far-field microphones and contaminated with additive noise at different SNRs. Here, the focus is to assess the robustness of the DNN-HMM system without any explicit enhancement or dereverberation processing. Therefore, the same clean-trained models are used from the Fisher experiments.

The improved robustness of the DNN-HMM system is due to the fact that the multiple layers of nonlinear processing in a DNN provide higher-level features in the upper layers that are much more invariant to small variations in the input (WERs are shown in Table 30).

Table 30: *WER recognition performance for IARPA ASPIRE Data – consisting of far-field microphones and additive background noise.*

Acoustic model	WER (%)
GMM-HMM + LDA + SAT	77.8
DNN-HMM (Raw MFCC)	68.2

4.2.4: DNN-based Acoustic Modeling – Next Steps:

- *DNN adaptation for environmental mismatch*. We have shown that by using a small number of adaptation utterances from the target environment and simple discriminative adaptation algorithms for a deep neural network acoustic model, relative improvements of up to 15% can be expected compared to the baseline DNN system [3].
- *Using alternative network architectures* such as convolutional neural networks (CNN) and recurrent neural networks (RNN). RNNs in particular are attractive choices for reverberant ASR since they explicitly model the inter-frame correlations.
- *Effective use of multichannel data* for a DNN-based ASR system. Current approaches include simple concatenation of features from different channels, conventional beamforming followed by training DNNs on the beamformed data, or training DNNs on the data from all different channels. Further advancements could explore weighting features based on environmental conditions, speaker conditions, or domain/style {i.e., monologue, 2-way conversation, prompted, etc.) structure.

4.2.5: Distance Based Publications:

[1.] S. Mirsamadi and J.H.L. Hansen, "Multichannel feature enhancement in distributed microphone arrays for robust distant speech recognition in smart rooms," IEEE Spoken Language Technology (SLT) Workshop, 2014.

[2.] S.M. Mirsamadi, J.H.L. Hansen, "Multichannel Speech Dereverberation based on Covolutive Nonnegative Tensor Factorization for ASR applications" ISCA Interspeech, 2014, pp. 2828 – 2832, Singapore, 14-18, 2014.

[3.] [3] S. Mirsamadi and J.H.L. Hansen, "A study on deep neural network acoustic model adaptation for robust far-field speech recognition" – submitted to ISCA Interspeech 2015.

[1] http://www.iarpa.gov/index.php/working-with-iarpa/prize-challenges/306-automatic-speech-in-reverberant-environments-aspire-challenge

4.3: Whisper Based Processing for ASR: In this section, advancements in whisper based processing for speech recognition are presented. Whisper speech is extremely challenging for ASR, so the ability to extract/identify any amount of content is viewed as a major accomplishment. It should be noted that word-error-rates (WERs) for speech in this content can be close to 100%, so any specific words can be viewed as a success (i.e., it will not ever be possible to achieve 0% WER on whisper speech when human listeners not present during speech production/conversation cannot hope to know groundtruth themselves).

4.3.1: Whisper Based Speech - Overview

Whisper represents an effective mode of communication in scenarios where the communicator does not wish to disturb uninvolved parties, or where a private information needs to be exchanged. Clearly, this makes whisper perfectly suited for human-machine interaction, especially with hand-held devices such as smartphones being used in open-office settings, company meetings, or public places. Unfortunately, a majority of current speech technology is designed and fine-tuned for modal (neutral) speech and breaks when faced with the acoustic-phonetic differences introduced by whisper.

In the voiced portion of modal speech, an air flow from the lungs results in vibration of the vocal folds within the larynx. These vibrations serve as the excitation to the vocal tract. In whispered speech, the glottis is kept open and an audible turbulent flow produced by passing air serves as the source for the articulators [1]. Besides the lack of periodic excitation from the glottal vocal folds, other prominent differences between modal speech and whisper can be observed in prosodic cues [2], phone durations [3], energy distribution between phone classes, spectral tilt, and formant locations due to different configurations of the vocal tract [1,4-11], resulting in altered distributions of phones in the formant space [12].

In this study, the focus is on the design of affordable strategies that would alleviate the mismatch between neutral-trained ASR models and incoming whispered speech with minimalistic requirements on the whispered adaptation data. While large vocabulary speech recognition (LVCSR) of whisper with neutral-trained models may seem largely unrealistic with current technology, we will show that in modest tasks with a constrained lexicon and language model, neutral-trained ASR models can be successfully adapted towards whisper to both significantly reduce whisper recognition errors, and at the same time accommodate neutral speech recognition without the need of external neutral/whisper segmentation. Indeed, for applications such as voice control of smart-phones/sending pre-set texts messages, constrained ASR maybe quite suitable.

In this task, four approaches are explored to address the whispered speech recognition:

1) In the first approach reconfigures the front-end of the ASR engine by changing the filterbanks to match the whisper speech characteristics better.
2) In the second method, formant upward shifts in whispered speech are compensated using vocal tract length normalization (VTLN) and Shift methods.
3) Finally, the last two approaches enable production of large quantities of whisper-like (pseudo-whisper) utterances from easily accessible modal speech recordings while requiring only a small amount of un-transcribed whisper samples to learn the target whisper domain characteristics. The generated pseudo-whisper samples are then used to adapt neutral ASR models to whisper. The two proposed methods utilize either:
    a. a vector Taylor series (VTS) algorithm or,
    b. a denoising autoencoder (DAE) solution.

4.3.2: Neutral/Whispered Speech Corpus

The speech corpus used in this study are drawn from the UT-Vocal Effort II (VEII) database [13]. The focus is on the read portion of VEII where each subject reads 41 TIMIT sentences [14] while switching between neutral and whispered speech modes. A subset of neutral and whispered TIMIT sentences from 39

female and 19 male speakers are used in the corresponding experiments. The recordings were downsampled to 16 kHz (from their original 44.1kHz rate). In the ASR experiments, the TIMIT [14] database is used for acoustic model training and baseline evaluations. The content of the VEII and TIMIT data sets used in this study is detailed in Table 31.

Table 31: *Speech corpora statistics used in this study; M/F - males/females; Train - training set; Adapt – model adaptation/VTS-GMM set; Ne/Wh - neutral/whispered speech; #Sents - number of sentences; Dur - total duration in minutes. Closed Speakers - same speakers (different utterances) in Adapt/Test; Open Speakers - different speakers in Adapt/Test.*

			# Sessions			
Corpus	Set	Style	M	F	# Sents	Dur
TIMIT	Train	Ne	326	136	4158	213
	Test	Ne	112	56	1512	78
VEII *Closed Speakers*	Adapt	Ne	19	39	577	23
		Wh			580	34
	Test	Ne			348	14
		Wh			348	21
VEII *Open Speakers*	Adapt	Ne	13	26	766	30
		Wh			779	45
	Test	Ne	5	13	351	14
		Wh			360	20

4.3.3: Recognition of Whispered Speech:

For all ASR experiments, a gender-independent speech recognizer was trained on 3.5 hours of TIMIT recordings (see Table 31). Here, 3-state left-to-right triphone HMMs with 8 Gaussians per state are used to model 39 phone categories (including silence). Front-end feature vectors are extracted using a 25 ms/10 ms windowing of a 16 kHz/16 bit audio signal and comprise 39 static, delta, and acceleration coefficients processed with cepstral mean normalization. The recognizer is implemented using the CMU Sphinx 3 toolkit [15].

In all experiments, the TIMIT acoustic models are MLLR-adapted in a supervised fashion towards the VEII acoustic/channel characteristics using the neutral adaptation sets detailed in Table 31. Based on the experiment, also the whispered portion of the adaptation set is used. The experiments are carried out on closed and open speaker test sets to evaluate how the potential benefits of the discussed methods transfer between the two application domains.

4.3.4: ASR of Whispered Speech: Baseline Setup

The neutral TIMIT acoustic models were adapted to the neutral VEII adapt set (duration of 23 minutes - see Table 31) and tested on whispered speech. The performance for the MFCC front-end and TIMIT LM dropped to 67.7% WER. This is not very surprising, given the considerable mismatch between the acoustic classes in neutral and whispered speech, especially when all voiced phones become unvoiced. In this sense, the acoustic mismatch is too prominent to perform any reasonable medium sized vocabulary recognition task of the whispered speech using simply neutral acoustic models. However, as discussed in the introduction, there are applications where recognition with a constrained grammar/language model may be meaningful, especially for whispered speech. To mimic such tasks, we restrict the lexicon/language model to approximately 160 words that cover the complete vocabulary of the VEII neutral and whispered test set. Results for the TIMIT models adapted with the VEII neutral adapt set and tested with closed speaker test set of neutral and whispered speech using the constrained lexicon are shown in Table 32 respectively. While the whisper set WER is still high, the task starts to be more applicable in real world environment.

4.3.5: ASR of Whispered Speech: Modified Front-Ends

Previous studies on robust ASR for stressed speech have reported performance gains when altering configurations of the front-end feature extraction filterbanks [16-18]. Inspired by [18], our first step is to replace the Mel filterbanks (FB) in MFCC and PLP features with a bank of triangular filters uniformly distributed over a linear frequency axis. In this case, the band cutoffs are located at the center frequencies of the adjacent filters [17]. The FB low and high cutoff frequencies are set to ~133 Hz and ~6855 Hz in all cases. The results for selected FB configurations are shown in Table 32, where 20Uni denotes a FB of 20 uniformly distributed filters. The modified PLP configurations are with bypassed equal loudness and power-intensity processing.

Table 32: *Performance of traditional front-ends; closed speaker test set; WER (%).*

Train	Adapt	Test	MFCC	PLP
TIMIT	-	TIMIT	6.0	6.6
	Ne	Ne	**5.2**	5.4
		Wh	27.0	24.6
	Wh	Wh	**18.2**	22.0

It can be seen that for MFCCs, the uniform triangular FB causes a slight WER degradation for the neutral set in both closed and open speaker test scenarios while providing a dramatic WER reduction for whisper (from 27.0 to 19.5% WER in closed test set, and from 38.5 to 30.2% WER in open test set from Table 33). For PLP, the triangular FBs reduce WER on both neutral and whispered speech. The PLP-20Uni features with neutral acoustic models provide comparable performance for whisper as the original MFCC system adapted to 34 minutes of transcribed whispered speech, which is quite encouraging. In addition, when applied to the original TIMIT train/test task, PLP-20Uni reduces the neutral WER to 5.5% (comparing with the 1st row in Table 32).

Table 33: *Performance of proposed strategies; WER (%).*

Speaker Scenario	Test Set	MFCC	MFCC 20Uni	PLP	PLP 20Uni	PLP 20Uni-Redist	PLP 20Uni-5800	PLP 20Uni-Redist-5800
Closed	Ne	5.2	**3.8**	5.4	4.0	4.1	4.5	3.9
	Wh	27.0	19.5	24.6	18.2	17.3	14.0	**13.7**
Open	Ne	6.3	5.8	7.1	5.2	5.6	5.5	**5.0**
	Wh	38.5	30.2	35.4	27.6	27.7	**22.9**	23.4

4.3.6: ASR of Whispered Speech: Changing Sub-Band Resolution

The previous section demonstrated a substantial whisper WER reduction due to replacement of the front-end FB. In this section, an approach is considered which reconfigures the filterbank resolution to further accommodate whispered speech. In [16], the authors analyzed the relevance of spectral subbands to speech recognition by training acoustic models on individual band energies of the filterbank. Subsequently, based on the band-specific WER, the filterbank was redistributed to increase its resolution in the most relevant parts of the spectrum.

In this section, we utilize a similar approach, with the difference that rather than training models on an output of a single filter at a time, we preserve the entire filterbank and only omit one filter in each iteration. Our baseline front-end in this experiment is PLP-20Uni. Fig. 76 presents WER contours for the neutral and whispered speech closed test set scenario (for a system adapted to neutral VEII set). The neutral and whisper WER contours suggest that the importance of the spectral components falling into the bands 3-8 is shared for both neutral and whispered speech.

Also, the WER vs. omitted frequency band curves in Fig. 76 suggests a rather ambiguous contribution of the highest frequency components to the neutral and whisper recognition performance. Based on this observation, we propose a modification of the previous features, PLP-20Uni-Redist-5800. This approach limits the linear filterbank in PLP-20Uni-Redist to the range of 133-5800 Hz. Based on these results, the rest of the experiments utilize PLP-20Uni-Redist-5800 features, unless stated otherwise.

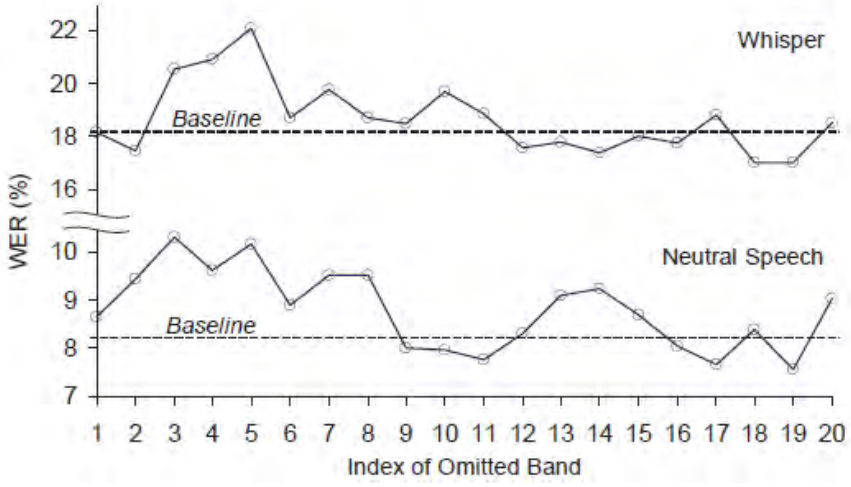


Figure 76: *WER vs. omitted filterbank bands; closed speaker test set.*

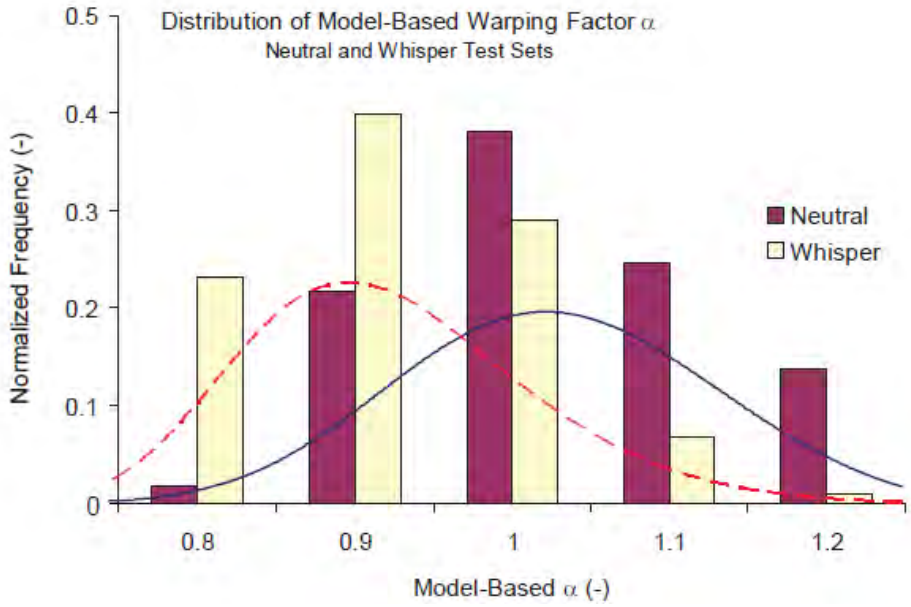


Figure 77: *Distribution of α's in neutral and whisper MVTLN. Full line - neutral samples; dashed line - whispered samples.*

4.3.7: ASR of Whispered Speech: Model & Feature-Based Compensation Methods

METHOD 1: VTLN and Shift Algorithm

As was shown in [7-10, 19], one of the differences between neutral and whispered speech is the shift of the formants to higher frequencies, especially the first and second ones. Similarly, formant changes are

observed among different speakers in neutral speech mode due to different vocal tract shape of each speaker. One of the standard methods to normalize this difference among speakers is maximum likelihood unsupervised VTLN which is performed by maximizing the likelihood function for each speaker using a simple frequency warping function that can be applied by modifying the filterbank in frequency domain [20]. Past studies have shown that VTLN is helpful in compensating for formants shifts caused by Lombard effect [21], which are in a way similar with those in whispered speech (upward shifts in F1 and F2).

However, since there are different shift rates in low and high formants of whispered speech, an alternative frequency axis warping, which is not passing through the coordinate origin, might be more suitable to shift the whispered formants toward neutral ones [18]. This alternative frequency warping is denoted as Shift.

Figure 77 presents the VTLN choices of α during decoding of neutral and whispered test sets. For the plot purposes, the counts of the 9 α candidates were accumulated into a 5-bar histogram. As the figure shows, the maximum for the neutral samples is at 1. The variance of the distribution reflects the VTLN effort to compensate for the vocal tract differences in the test set individuals. The α distribution maximum for whisper is at 0.9, where the corresponding high cutoff frequency of the extraction filterbank is increased by a factor of ~1.11 - resulting in filter bank stretching. This confirms that VTLN is trying to compensate for the upward formant shifts in whisper by compressing the speech spectrum in the frequency domain.

Figure 78 shows the formant space for neutral, whisper and (model domain) M.D. VTLN-transformed whispered speech. In the figure, there is a large mismatch in the vowel space of neutral and whispered speech (switching of /eh/ and /uh/ phones). However, (model domain) M.D. VTLN approach is successful in moving the formants back to their original places. Still a considerable distance between neutral and whispered samples is observed in the formant space.

METHOD 2: VTS Algorithm

Past studies on whispered speech recognition [9, 10, 22, 23] suggest that neutral-trained model adaptation towards whisper is effective in reducing the acoustic mismatch between the two speech modalities. However, for a successful supervised adaptation, a sufficient amount of transcribed whisper adaptation data is required. In this and the following section, two strategies are proposed that require only a small amount of un-transcribed whispered utterances to produce a large population of pseudo-whisper samples from available neutral speech. The pseudo-whisper samples are used for effective neutral ASR model adaptation towards whisper. This is motivated by the fact that large corpora of transcribed neutral speech are easily accessible to system designers while transcribed whisper is rare and difficult to acquire.

In the VTS method, the environment is modeled as a speech signal corrupted by channel effect and an additive stationary noise [24, 25]. In this study, the same idea is applied to transform neutral speech to its pseudo-whisper samples. Meaning, neutral speech is modeled as whispered speech being passed through a channel with an added noise.

The process is outlined in Fig. 79. First, a small amount of unlabeled whisper samples are used to train a whisper Gaussian mixture model (GMM). Subsequently, we utilize this GMM in the VTS scheme to extract transforms for broad phone classes (voiced and unvoiced) for the neutral utterances drawn from the `adaptation' set (see Table 30). The transforms are estimated on an utterance level. Phone boundaries in the neutral utterances are estimated using forced alignment (since transcriptions for adaptation data are available). Finally, the transforms are applied at the utterance level to produce pseudo-whispered samples.

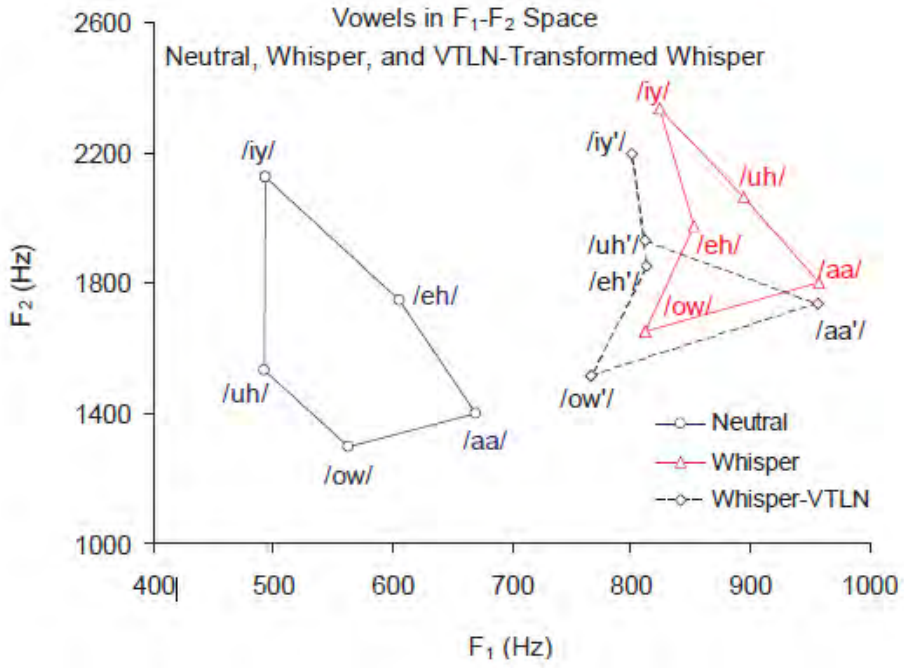


Figure 78: *VTS* Vowel distributions in F1-F2 formant space; neutral, whisper, and VTLN-transformed whisper samples from closed speakers sets.

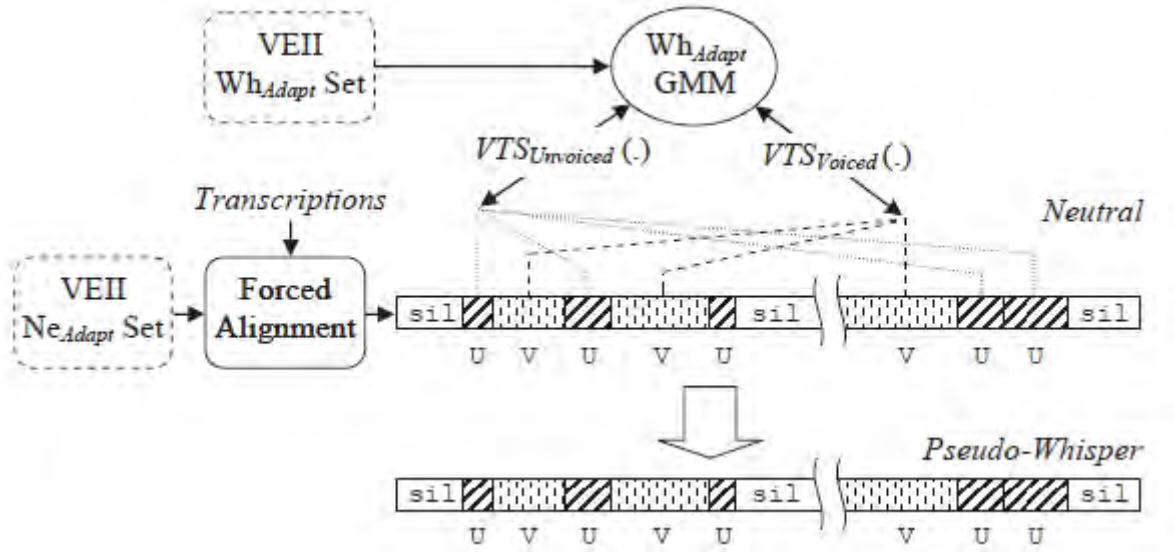


Figure 79: *VTS-based generation of pseudo-whisper samples using whisper GMM and samples from neutral Adapt set. Voiced- and unvoiced-specific VTS transforms are applied in the example.*

In the first experiment to evaluate the VTS method performance, the efficiency of VTS-transformed data for model adaptation is compared when using transformations derived from broad phone classes (voiced and unvoiced). For this experiment, the cases when only one phone class is transformed at a time is compared for pseudo-whisper speech generation, and the case when both classes are transformed at the same time using their respective transformations. It can be seen that for both closed and open speakers scenarios, the WERs follow the same trend.

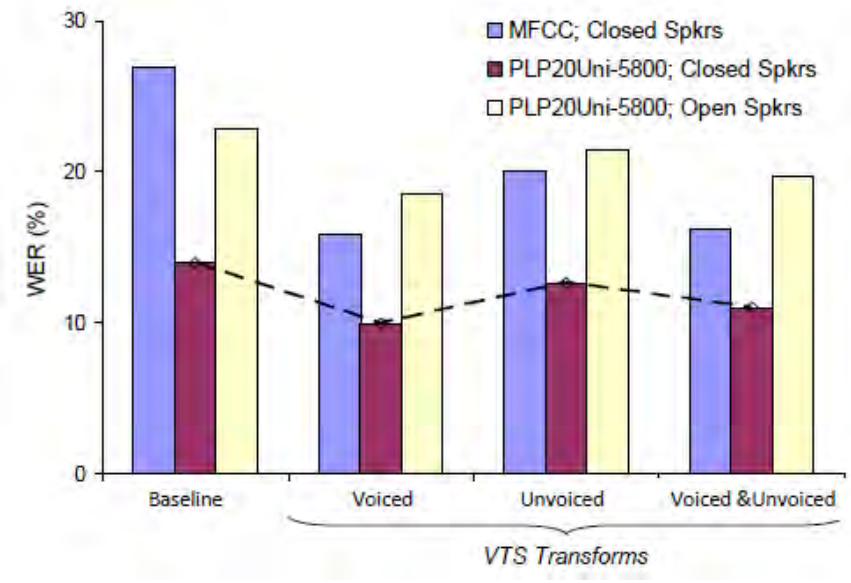


Figure 80: *Performance of VTS with voiced-, unvoiced-, and "voiced- and unvoiced"-specific transforms applied.*

Next, the effects of adaptation set size is studied on recognition performance for the two proposed solutions. In the first approach denoted MLLR, the samples are used in MLLR adaptation to transform the neutral TIMIT models towards VEII channel/acoustics characteristics and whispered speech. In the second solution denoted VTS, TIMIT-trained models are MLLR adapted toward pseudo-whisper generated samples using VTS (see Fig. 81). While both setups have access to the same original adaptation data sets, the VTS configuration can effectively produce as many pseudo-whisper samples as available in the neutral set. Figures 82(a) and 82(b) compare performances on closed and open speaker sets for both neutral and whisper data. It can be seen that performances are identical for MLLR and VTS for an empty whisper adaptation set, and that VTS starts showing superior performance in all other conditions.

4.3.8: ASR of Whispered Speech: Denoising Autoencoder (DAE) Algorithm

In this section, the DAE network is introduced to generate pseudo-whisper samples (see Fig. 81). An autoencoder is an artificial neural network trained to reconstruct its input [26]. An autoencoder tries to find a deterministic mapping between input units and hidden nodes by means of a nonlinear function. DAE have been recently used in speech recognition for denoising and dereverberation of speech [27, 28].

Similar to the VTS approach [29], the idea is to assume that neutral speech samples are statistically corrupted version of whispered speech. We consider two approaches to transform neutral speech to their corresponding pseudo-whisper ones (Fig. 83): (1) *feature-based*: DAE produces pseudo-whisper cepstral samples on frame bases; (2) *statistical-based*: DAE produces statistical characteristics of cepstral features, means and variances, which are then used to transform the whole phone segment to its pseudo-whisper equivalent.

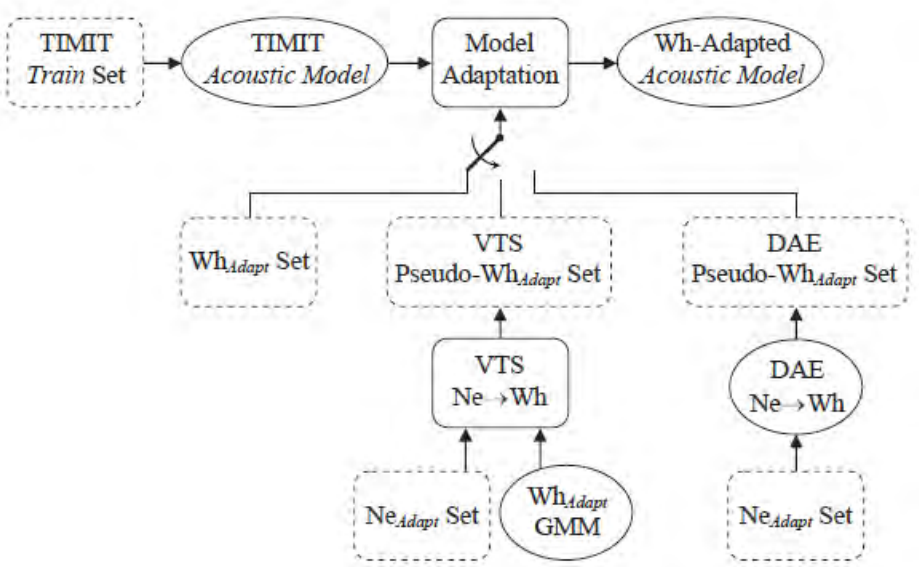


Figure 81: *Three approaches to whisper model adaptation.*

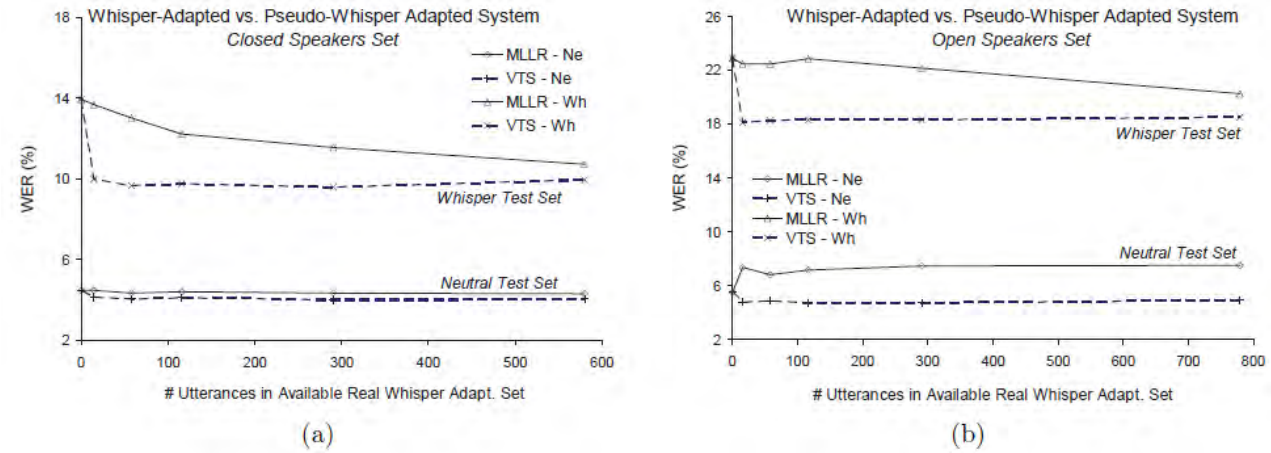


Figure 82: *Comparison of model adaptation on whisper and on VTS-generated pseudo-whisper samples; (a) closed speakers set; (b) open speakers set*

Figures 84(a) and 84(b) illustrate the detailed steps of the two DAE methods. Both approaches use transcribed neutral and whisper samples drawn from the Adapt set (see Table 30). Neutral or whispered frames assigned to a specific phone (e.g. /aa/), by means of forced alignment, are grouped together to form a single phone stream. Also, it is noted that usually the length of each phone in neutral and whisper domains are not the same, therefore the training will stop when the last frame of the shorter stream is reached.

Two scenarios are considered for DAE-based pseudo-whisper generation: (i) *supervised* - it is assumed that word- level transcriptions for the real whisper adaptation set are available; here, forced alignment using the neutral ASR system and the available transcriptions is carried out to estimate phone boundaries in the whisper adapt set; (ii) *unsupervised* - whisper transcriptions are not available. The latter configuration either disregards phone boundaries in the whisper adaptation samples (denoted Random in Table 34) or relies on the neutral ASR engine to estimate the phonetic content and its boundaries in the whisper adaptation set (denoted Neutral ASR Alignment).

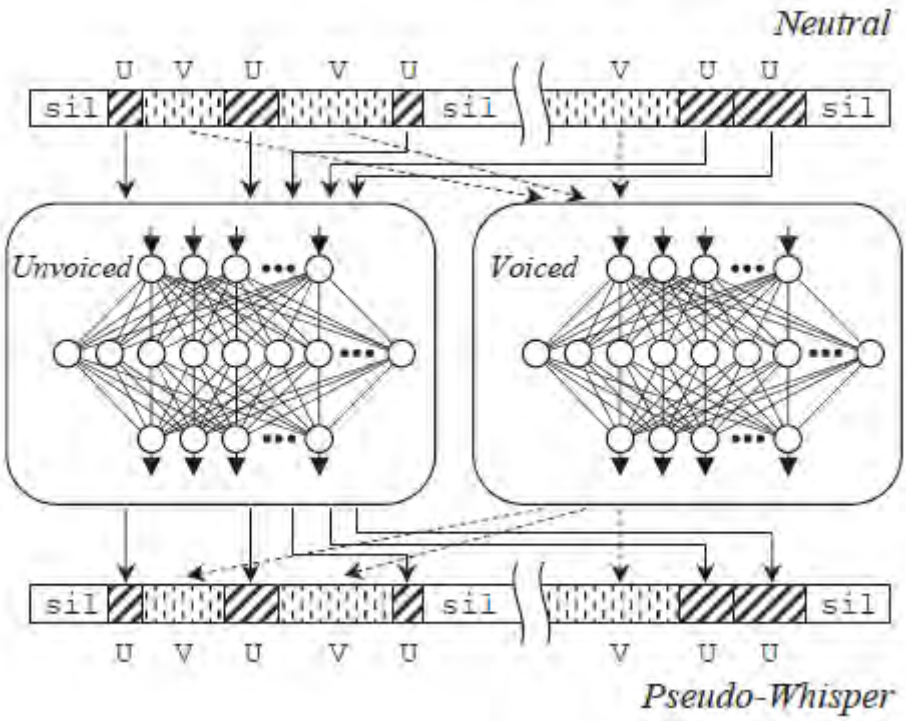


Figure 83: *DAE-based generation of pseudo-whisper samples using unvoiced- and voiced-specific nets trained on Adapt set. In feature-based approach, DAE directly generates pseudo-whisper cepstral frames; in statistical-based approach, DAE produces phone segment statistics that are then used to transform neutral phone segments to pseudo-whisper.*

Table 34 summarizes a point measurement results (290 real whisper adaptation samples available) for all DAE setups. The first two result rows compare supervised feature-based (Feat.) and statistical-based (Stat.) systems that utilize forced alignment (F.A.) on the whisper adaptation set. The feature-based DAE takes one feature frame at a time as an input and simultaneously produces one output feature frame, while the statistical-based DAE is trained on phone-segment statistics as inputs and targets. It can be seen that the ASR systems adapted on pseudo-whisper produced by the two supervised DAE approaches reach comparable performance, with the feature-based DAE being more successful in all conditions besides the closed speakers whisper scenario. For this reason, and to limit the amount of experiments, only feature-based DAE is considered for the unsupervised scenarios.

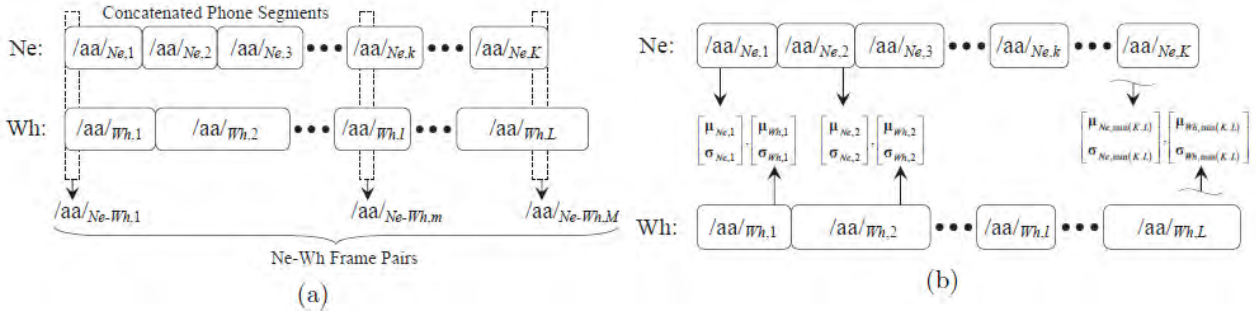


Figure 84: *Data segmentation for DAE fine-tuning; (a) feature-based approach; (b) statistical-based approach. This is repeated for all phone classes.*

In the unsupervised section of Table 34, several configurations of the Random and Neutral ASR Alignment setups are considered. In the Random scenario, the neutral adaptation set is still labeled by means of forced alignment. Instead of labeling whispered adaptation samples and splitting them into phone-specific target streams, they are concatenated into a single, phone-independent whisper stream. When training the unvoiced and voiced DAEs, the network inputs are presented with the neutral samples from the respective broad phone categories while the samples from the concatenated whisper stream form the DAE targets. Besides training the DAE to perform a frame-to-frame mapping during its training (first row of the Random DAE results), also cases where a mean cepstral vector extracted from a 5- or 11- frames long sliding window from the concatenated whisper stream was provided as a target (Avg5Frames and Avg11Frames rows) are considered. The assumption here is that averaging the neighbouring whisper segments may provide more stable targets for the DAE training. Lastly, a Random DAE setup where 11 neighboring frames from the neutral stream are provided simultaneously at the DAE input, is considered to investigate the effects of temporal context. As can be seen in Table 34, the Random DAE WERs on whispered test sets are in general higher than those of the supervised DAEs, which suggests that the partitioning of whisper into two rather than one broad phonetic classes is beneficial. Averaging the adjacent frames in the target stream had a positive impact on DAE training as it converged more than twice as fast compared to using per-frame targets. Providing the broader temporal context resulted in slight whisper WER reduction compared to all other Random DAE setups.

In the Neutral ASR Alignment scenario, similar experiments with extending the input temporal context and smoothing the output targets are carried out with the difference that the target averaging here is performed on the level of the whole phone segment estimated from the ASR alignment (AvgPhoneSeg). As shown in the penultimate results row, this setup reaches comparable WER on the open speakers task vs. the supervised feature-based system. The final row of Table 33 shows the additional benefit of incorporating the Model-Domain (M.D.) Shift in the unsupervised scheme (no additional gain for Whisper in the Open case, but clear gains in neutral and whisper in closed as well as neutral in open).

Table 34: *Performance of supervised and unsupervised DAE strategies; WER (%).*

Phone-To-Phone Mapping		DAE Input	DAE Output	Frequency Transform	Closed Ne	Closed Wh	Open Ne	Open Wh
Supervised	F.A.; *Feat.* DAE	1	1	None	**3.4**	9.8	**4.9**	**18.0**
	F.A.; *Stat.* DAE	PhoneSeg	PhoneSeg		4.1	**9.4**	5.0	18.9
Unsupervised	Random (*Feat.* DAE)	1	1		3.4	10.3	**5.1**	20.5
		1	Avg5Frames		**3.2**	10.7	5.3	20.2
		1	Avg11Frames		3.3	10.4	5.2	20.1
		11	Avg11Frames		3.5	**10.2**	5.4	**19.3**
	Neutral ASR Alignment (*Feat.* DAE)	1	1		**3.0**	**8.4**	**4.0**	19.2
		11	AvgPhoneSeg		3.7	10.3	5.2	**18.2**
		11	AvgPhoneSeg	Shift (M.D.)	3.2	8.9	4.1	**18.2**

Figures 85(a) and 85(b) compare performance of ASR systems adapted to pseudo-whisper from the two unsupervised DAE setups (last two rows of Table 34) to the baseline MLLR system and the system adapted to VTS pseudo-whisper. The notation Ne/Wh in the trend captions denotes the neutral or whispered test set, and MShift refers to the model domain Shift transformation. It can be seen that the proposed VTS and DAE adaptation schemes provide considerable WER reduction over the traditional adaptation on the available whisper samples. In addition, both VTS and DAE benefit from being combined with the Shift transform in most of the evaluation conditions. In the open speakers whisper task, VTS with Shift slightly outperforms the two DAE solutions and Shift somewhat reduces DAEs performance for larger adaptation set sizes.

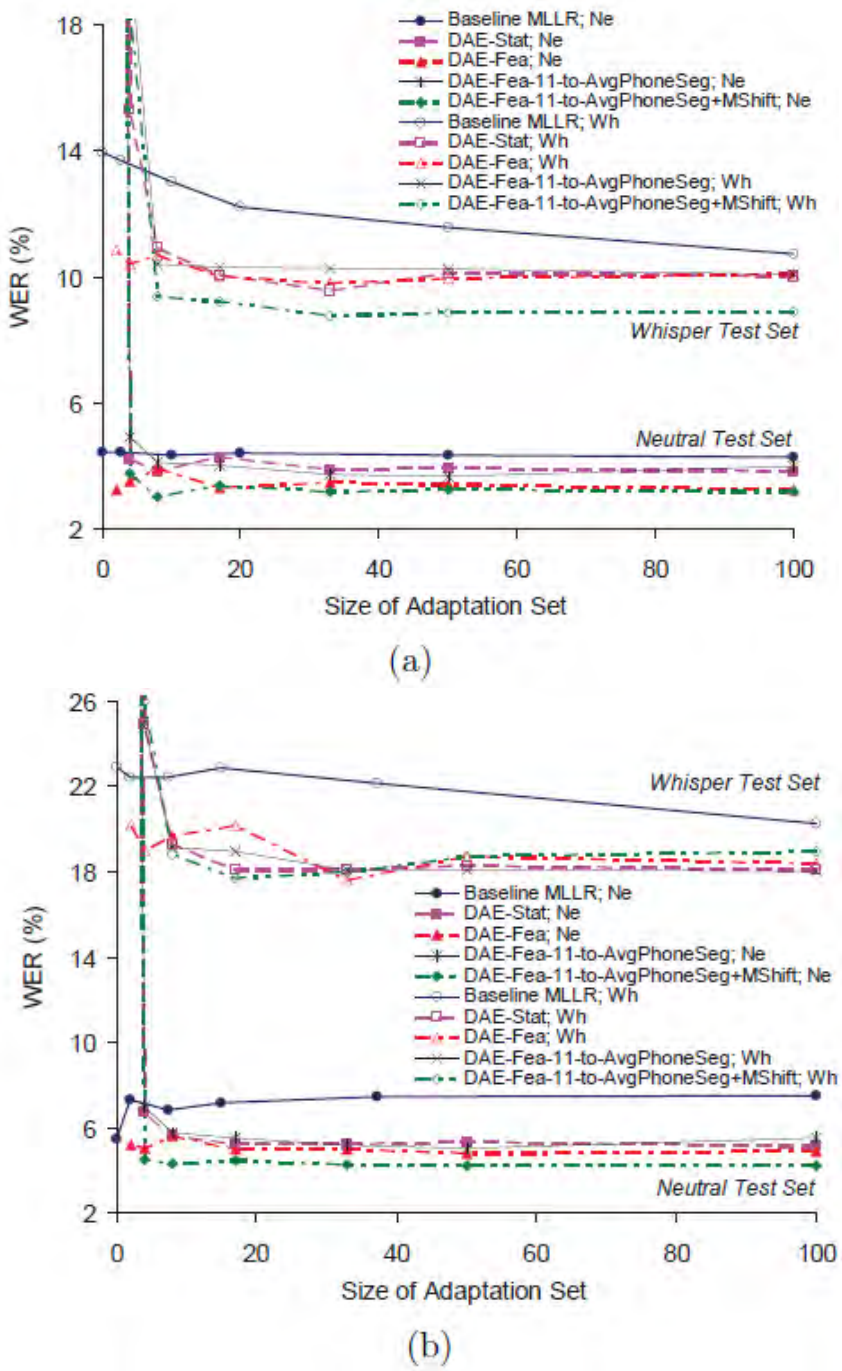


Figure 85: *Comparison of model adaptation on whisper (MLLR), as well as supervised and unsupervised DAE-generated pseudo-whisper samples;(a) closed (b) open speaker test sets; DAEs with 300 hidden neurons.*

4.3.9: ASR of Whispered Speech: Summary of Experimental Results

Experimental results from this sub-task has compared WERs of baseline systems, modified front-ends, frequency transformations, and the best VTS and DAE setups. It can be seen that the modified front-ends notably reduce the errors of the baseline MFCC and PLP setups, and further benefit when combined with the VTLN and Shift transformations. The best DAE configuration (Stat. DAE + Shift (M.D.)) outperforms the PLP baseline by 15.8% absolute on the closed speakers, and by 17.4% absolute WER on the open

speakers whisper task. In addition, the same DAE outperforms a system sharing the same front-end (PLP-20Uni-Redist-5800), but whose acoustic models were adapted in a supervised way to the transcribed real whisper, by 2.7 % on closed speakers and 4.2% absolute WER on open speakers whisper task. The best VTS setup (VTS + Shift (M.D.)) outperforms PLP by 15.6% on the closed speakers and 18.0 % on the open speakers whisper task, and a PLP-20Uni-Redist-5800 system adapted to transcribed real whisper by 2.6% and 4.7%, respectively. It is observed that in spite of their conceptual differences, VTS and DAE provide mutually competitive performance improvements across all tasks.

4.3.9: ASR of Whispered Speech: Publications from this effort:

[1.] Ghaffarzadegan S., Boril H., Hansen J. H. L., "Generative Modeling of Pseudo-Whisper for Robust Whispered Speech Recognition," submitted to IEEE Trans. on Audio, Speech & Language Processing.

[2.] Ghaffarzadegan S., Boril H., Hansen J. H. L., "UT-VOCAL EFFORT II: Analysis and Constrained-Lexicon Recognition of Whispered Speech", IEEE ICASSP 2014.

[3.] Ghaffarzadegan S., Boril H., Hansen J. H. L., "Model and Feature Based Compensation for Whispered Speech Recognition"," ISCA INTERSPEECH 2014.

[4.] Ghaffarzadegan S., Boril H., Hansen J. H. L., "Generative Modeling of Pseudo-Target Domain Adaptation Samples for Whispered Speech Recognition," IEEE ICASSP 2015.

4.3.9: ASR of Whispered Speech: References for this Sub-Task:

[1] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," Medical Engineering & Physics, vol. 24, no. 7, pp. 515-520, September 2002.

[2] W. F. L. Heeren and C. Lorenzi, "Perception of prosody in normal and whispered French," The Journal of the Acoustical Society of America, vol. 135, no. 4, pp. 2026-2040, 2014.

[3] P. X. Lee, D. Wee, H. S. Y. Toh, B. P. Lim, N. Chen, and B. Ma, "A whispered Mandarin corpus for speech technology applications," ISCA INTERSPEECH`14, Singapore, September 2014, pp. 1598-1602.

[4] C. Zhang, T. Yu, and J. H. L. Hansen, \Microphone array processing for distance speech capture: A probe study on whisper speech detection," Asilomar Conference on Signals, Systems and Computers (ASILOMAR), 2010, pp. 1707-1710.

[5] X. Fan and J. H. L. Hansen, "Acoustic analysis for speaker identification of whispered speech," IEEE ICASSP'10, 2010, pp. 5046-5049.

[6] X. Fan and J. H. L. Hansen, "Speaker Identification within whispered speech audio streams," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1408-1421, 2011.

[7] T. Ito, K. Takeda, and F. Itakura, \Acoustic analysis and recognition of whispered speech," IEEE ASRU'01, 2001, pp. 429-432.

[8] I. Eklund and H. Traunmuller, "Comparative study of male and female whispered and phonated versions of the long vowels of Swedish," Phonetica, pp. 1-21, 1997.

[9] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," Speech Communication, vol. 45, no. 2, pp. 139 - 152, 2005.

[10] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2011.

[11] M. Matsuda and H. Kasuya, "Acoustic nature of the whisper," EUROSPEECH'99, 1999, pp. 133-136.

[12] H. R. Sharifzadeh, I. V. McLoughlin, and M. J. Russell, "A comprehensive vowel space for whispered speech," Journal of Voice, vol. 26, no. 2, pp. e49-e56, 2012.

[13] C. Zhang and J. H. L. Hansen, "Advancement in whisper-island detection with normally phonated audio streams," ISCA INTERSPEECH-2009, 2009, pp. 860-863.

[14] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," Speech Comm., vol. 9, no. 4, pp. 351 - 356, 1990.

[15] C. M. University, "CMUSphinx - Open source toolkit for speech recognition," 2013. [Online]. Available: http://cmusphinx.sourceforge.net/wiki

[16] S. E. Bou-Ghazale, J.H.L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," IEEE Trans. on Speech & Audio Processing, vol. 8, no. 4, pp. 429-442, 2000.

[17] H. Boril, P. Fousek, and P. Pollak, "Data-driven design of front-end filter bank for Lombard speech recognition," Interspeech-06, Pittsburgh, Pennsylvania, 2006, pp. 381 -384.

[18] H. Boril and J.H.L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," IEEE Trans. Audio, Speech, Lang. Proc., vol. 18, no. 6, pp. 1379-1393, Aug. 2010.

[19] S. Ghaffarzadegan, H. Boril, and J.H.L. Hansen, “UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech," IEEE ICASSP, 2014.
[20] L. Lee and R. C. Rose, “Speaker normalization using efficient frequency warping procedure," in IEEE ICASSP, 1996, pp. 353-356.
[21] H. Boril, “Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora," Ph.D. dissertation, CTU in Prague, Czech Rep., http://www.utdallas.edu//hynek, 2008.
[22] A. Mathur, S. M. Reddy, and R. M. Hegde, “Significance of parametric spectral ratio methods in detection and recognition of whispered speech," EURASIP Journal on Advances in Signal Processing, vol. 2012, no. 1, pp. 1-20, 2012.
[23] C.-Y. Yang, G. Brown, L. Lu, J. Yamagishi, and S. King, “Noise-robust whispered speech recognition using a non-audible-murmur microphone with VTS compensation," iSymp. Chinese Spoken Language Processing (ISC-SLP), 2012, pp. 220-223.
[24] P. J. Moreno, B. Raj, and R. M. Stern, “A vector Taylor series approach for environment-independent speech recognition," IEEE ICASSP-96, 1996, pp. 733-736.
[25] A. Acero, L. Deng, T. T. Kristjansson, and J. Zhang, “HMM adaptation using vector Taylor series for noisy speech recognition." ISCA INTERSPEECH, 2000, pp. 869-872.
[26] P. Vincent, H. Larochelle, Y. Bengio, P. Manzagol, “Extracting and composing robust features with denoising autoencoders," Inter. Conference on Machine Learning, ICML '08. New York, NY, 2008, pp. 1096-1103.
[27] X. Feng, Y. Zhang, and J. R. Glass, “Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," IEEE ICASSP.
[28] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, \Reverberant speech recognition based on denoising autoencoder." ISCA INTERSPEECH. ISCA, 2013, pp. 3512-3516.
[29] S. Ghaffarzadegan, H. Boril, and J.H.L. Hansen, “Model and feature based compensation for whispered speech recognition," in Interspeech 2014, Singapore, Sept 2014, pp. 2420-2424.

Task 5 – Speaker State Assessment / Environmental Sniffing (SSA/EnvS):

In this task, a number of sub-areas were considered. In essence, many of the tasks represent high risk, but potentially high reward based speech/language research. The topic addressed involved aspects relating to automatic identification and knowledge extraction of both speaker characteristics as well as environmental structure known as environmental sniffing. The specific sub-tasks include: (i) speaker height estimation from speech, (ii) nonlinear distortion detection with emphasis on peak clipping detection, (iii) gender identification in noise, and (iv) Lombard effect processing for speech recognition. Additional work on deep neural networks (DNNs) using graphical processing units (GPUs) was also considered, and can be found in the corresponding publications.

5.1: Speaker Height Estimation from Speech:

Estimating speaker height can assist in voice forensic analysis and provide additional side knowledge to benefit automatic speaker identification, or acoustic model selection for automatic speech recognition. In this study, a statistical approach to height estimation that incorporates acoustic models within a non-uniform height bin width Gaussian Mixture Model structure, as well as a formant analysis approach that employs linear regression on selected phones are presented. The accuracy and trade-offs of these systems are explored by examining the consistency of the results, location and causes of error, as well a combined fusion of the two systems using data from the TIMIT corpus. Open set testing is also presented using the MARP corpus and publicly available YouTube audio to examine the effect of channel mismatch between training and testing data and provide a realistic open domain testing scenario. The proposed algorithms achieve a highly competitive performance to previously published literature. Although the different data partitioning in the literature and this study may prevent performance comparisons in absolute terms, the mean average error of 4.89 cm for males and 4.55 cm for females provided by the proposed algorithm on TIMIT utterances containing selected phones suggest a considerable estimation error decrease compared to past efforts.

5.1.1: Speaker Height – Background: A majority of studies on height estimation from voice rely on the assumed correlation between individual's height and vocal tract length (VTL), supported by the evidence from magnetic resonance imaging (MRI) (Fitch and Giedd, 1999). Among other speech production features, low frequency energy (van Dommelen and Moxness, 1995), glottal pulse rate (Smith et al., 2005), subglottal resonances (Arsikere et al., 2012), fundamental frequency (Lass and Brown, 1978; Kunzel, 1989; van Dommelen and Moxness, 1995; Rendall et al., 2005; Ganchev et al., 2010a), formants (van Dommelen and Moxness, 1995; Rendall et al., 2005; Greisbach, 1999), and Mel frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC) (Pellom and Hansen, 1997; Dusan, 2005) were studied in the context of height.

One of the first studies in the area of automatic height estimation from speech used a statistical approach based on a Gaussian mixture model (GMM) class structure with 19 static MFCCs as the feature vector (Pellom and Hansen, 1997). Using the TIMIT corpus, an accuracy of 70% was achieved within 5 cm but it should be noted that the speaker independent height models were trained on selected sentences from all available TIMIT speakers and hence, the evaluation set contained samples from the same speakers (yet different sentences).

The proposed solution for speaker height estimation consists of two parallel solutions which are ultimately fused together to form an overall estimated height value. Figure 86 presents the flow diagram.

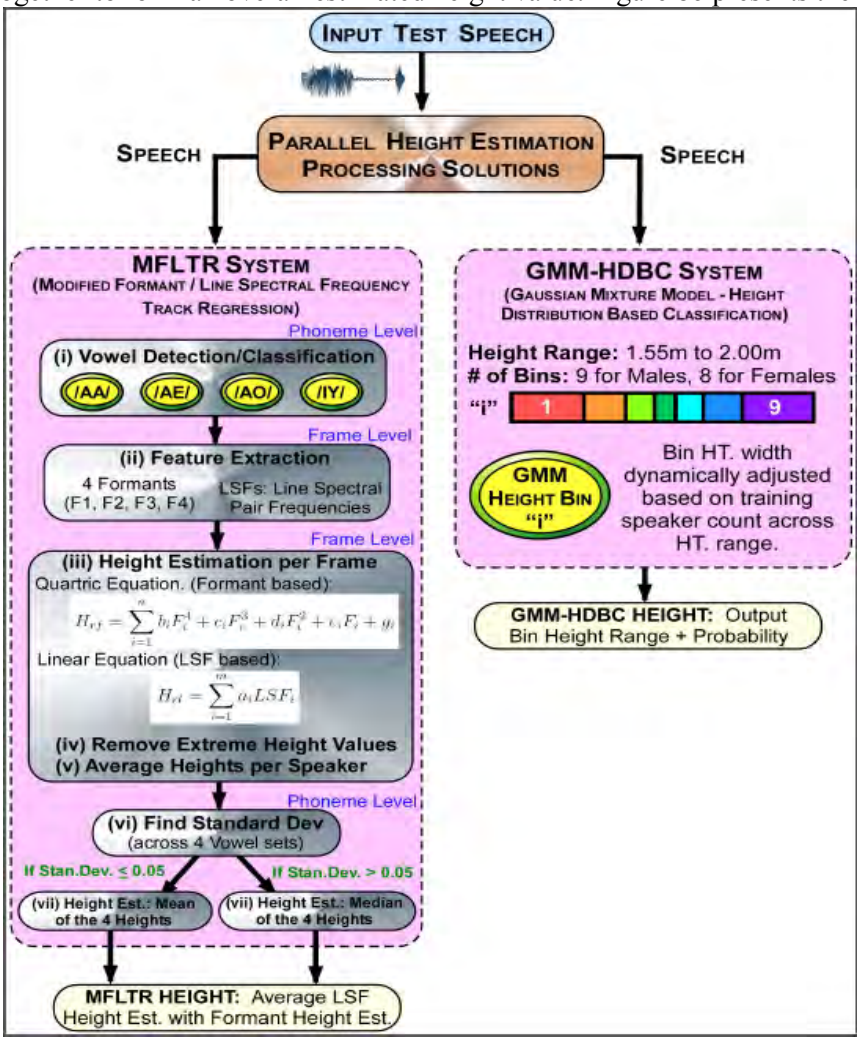


Figure 86: *Overall flow diagram of CRSS-UTDallas Height Estimation solution from speech. The Modified Formant Track/LSF Regression solution (MFLTR) is on the left, and the GMM Height Distribution Based Classification (GMM-HDBC) is shown on the right.*

5.1.2: Speaker Height – Proposed MFLTR solution:

Modified Formant/Line Spectral Frequency Track Regression – MFLTR

The first proposed height estimation solution is MFLTR employs both LSF based height estimation as well as direct formant location estimation. When examining the LSFs (line spectral pair frequencies), the formant information can be inferred by two closely paired LSFs (Itakura, 1975; Crosmer and Barnwell, 1985). Once the raw formant locations and LSFs have been calculated, they can be smoothed over time in order to reduce estimation errors that are known to occur. The modified formant track/LSF regression algorithm (MFLTR) for height estimation is based on solving an equation that represents the height of a speaker in terms of the first four formants along with 18 LSFs, and then performing a post-processing clean-up phase for the height estimates.

For system development, a total of 268 male and 127 female sessions from the TIMIT corpus sampled at 16 kHz were utilized in the evaluations. The gender-dependent sets were split approximately in half to form training and test sets with non-overlapping speakers.

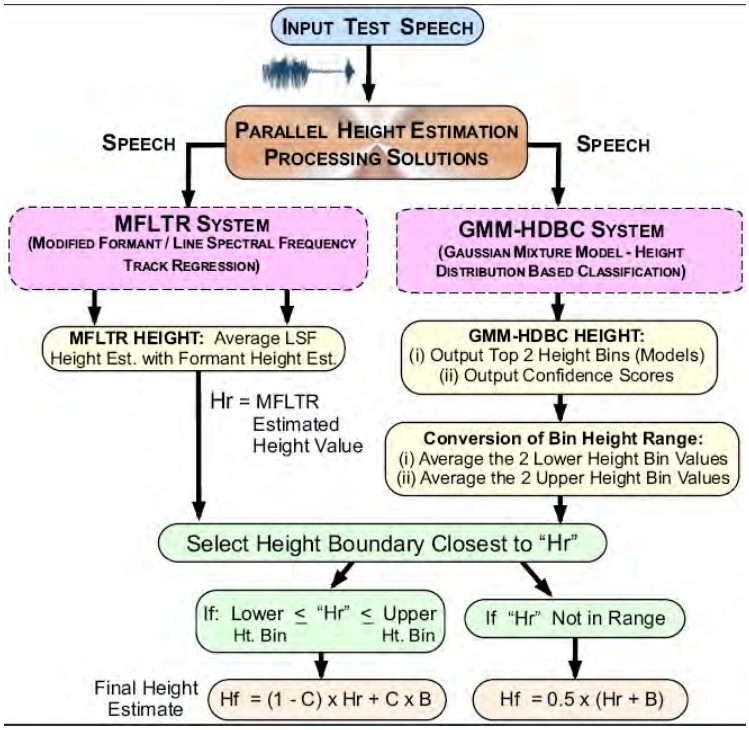


Figure 87: *Algorithm details for the Modified Formant Track/LSF Regression solution (MFLTR) and GMM Height Distribution Based Classification (GMM-HDBC) solutions.*

5.1.3: Speaker Height – Proposed GMM-HDBC solution:

A second alternative height estimation scheme, GMM Height Distribution Based Classification (GMM–HDBC), was formulated based on statistical modeling concepts. The feature used for this method consists of 19 static MFCC coefficients including normalized energy. MFCCs have been shown in a previous study to be effective in reflecting a speaker's height (Pellom and Hansen, 1997; Dusan, 2005). The normalized energy is included in order to accommodate thresholding-based silence and low energy speech segments since those are not expected to provide any useful information.

A balanced height coverage for both training and evaluation is assured. This configuration will reduce the speaker dependency and ensure an effective height class estimate for each speaker. The nine centroids in meters for males are as follows: (1.635, 1.73, 1.75, 1.78, 1.8, 1.83, 1.85, 1.88, and 1.935), while the eight centroids for females are: (1.51, 1.6, 1.63, 1.65, 1.68, 1.7, 1.73, and 1.79). It seems useful to also include a

confidence measure to communicate how likely that height class might be for the user. The confidence measure used here is the probability closeness measure (Rabiner and Schafer, 2011).

The MFLTR algorithm produces a specific height estimate for each speaker, while the GMM-HDBC method assigns a height class along with a confidence score. The first step in combining the two methods is to find the lower and upper height boundaries for the top 2 height classes and average the two lower and upper boundaries together. Next, the height value from the MFLTR method is compared to the new upper and lower boundary to determine which is closer. This results in a compromised height estimate since it effectively averages the closest boundary height B with the estimated MFLTR height output H_R. With this method, there will be a single height result per speaker, H_F.

5.1.4: Speaker Height Evaluation – MFLTR Results

The results for the MFLTR method are displayed in Tables 35 and 36. Mean absolute error, MAE (in cm), is the measure chosen to reflect performance of the MFLTR method since it has been used in previous studies for height estimation (Ganchev et al., 2010a; Mporas and Ganchev, 2009; Arsikere et al., 2012, 2013; Williams and Hansen, 2013). MAE was calculated on a per speaker basis.

Table 35: *Comparison of Height Estimation MAEs (mean average error, in cm) for the MFLTR method at the level of individual phonemes – LSF vs. Formants.*

Male				
	LSF			
Phoneme	/AA/	/AE/	/AO/	/IY/
MAE (cm)	5.39	5.29	5.53	5.07
	Formants			
Phoneme	/AA/	/AE/	/AO/	/IY/
MAE (cm)	5.14	5.26	5.45	5.49

Female				
	LSF			
Phoneme	/AA/	/AE/	/AO/	/IY/
MAE (cm)	5.55	5.61	6.32	5.08
	Formants			
Phoneme	/AA/	/AE/	/AO/	/IY/
MAE (cm)	5.06	4.72	5.44	5.35

Table 36: *Comparison of Height Estimation MAEs (mean average error, in cm) for MFLTR method after Phoneme Combination – LSF vs. Formants.*

Male			
	LSF	Formants	Combined
MAE (cm)	4.92	5.06	4.93
Female			
	LSF	Formants	Combined
MAE (cm)	5.23	4.80	4.76

Table 37: *Height Estimation MAEs (mean average error, in cm) AFTER Fusion of MFLTR and GMM-HDBC methods.*

		Fusion
MAE (cm)	Male	4.89
	Female	4.55

Error in terms of MAE (in cm) ranges from 5.14-5.53 cm for males and 4.72–6.32 cm for females. Again note, these height estimates are based on a single vowel (~0.25–0.5 sec). In Table 36, the results shown are obtained after the phoneme level analysis in Figure 87. Here, the solution combines estimates from the one-to-four phonemes to obtain an overall result for the LSF feature as well as the formant feature. The combination result is the final accuracy of the MFLTR method which achieves very effective performance for males (4.93 cm) and females (4.76 cm).

These results comfirm that the majority of the speakers have error less than 5 cm (59.2% of the male speakers, 55.9% of the female speakers). The correlation coefficient for males was determined to be 0.26 while for females the correlation coefficient is 0.34. These coefficients are not considerably high but when considering both males and females together, the correlation coefficient is 0.72 which demonstrates a better relationship between estimated and actual height.

5.1.5: Speaker Height Evaluation – GMM-HDBC Results

The results for the statistical GMM-HDBC height estimation method are determined by considering accuracy within a 5 cm range. For each confidence measure, only those speakers with at least that number are considered. As a result, a reduced number of speakers are included in the results as the confidence measure increases. The vertical lines represent 25% speaker elimination and 50% speaker elimination, respectively.

5.1.6: Speaker Height Evaluation – Fusion of MFLTR and GMM-HDBC

Having demonstrated individual MFLTR and GMM-HDBC performance, the fusion of these systems are now considered. The fusion result is shown in terms of MAE in order to compare performance with MFLTR. The fusion result is tabulated in Table 37. The combined fusion method achieves an MAE with the highest accuracy out of all methods. The GMM-HDBC method when combined with the MFLTR method provides an added level of assurance when the phoneme results are combined. When both males and females are grouped together, the overall correlation coefficient increases to 0.73. A height estimation evaluation was also performed on the MARP corpus, where height ranges were estimated from repeated sessions on males and females. Performance showed some variability, but in general they were consistent.

5.1.7: Speaker Height Evaluation – Open Test Set (TV and Movie Actors)

As a final exploration, the individual height estimation solutions are evaluated on open public speaker data. In order to accomplish this last evaluation, 8 male and 8 female movie/TV actors were chosen to be test speakers due to the availability of speech from interviews, movies, etc. The speech was drawn from YouTube where at least one of the 4 vowels, /AA/, /AE/, /AO/, and /IY/, had to be included in the specific test speech. The speech data was chosen to have minimal background noise. For male actors, the MFLTR method produced individual errors ranging from 1.58 cm to 6.53 cm. (see Table 38). The highest error occurred with the tallest speaker. For females, the MFLTR performed similarly with individual errors ranging from 1.7 cm to 8.29 cm. The highest errors generally also occurred for the tallest females as well.

5.1.8: Speaker Height Evaluation – Conclusion

In this sub-task, the problem of accurate height estimation from speech was investigated. Two alternative solutions were developed for engaging in automatic speaker height estimation as well as a fusion of the two individual methods. The first method, MFLTR, obtains a point estimate of height for each speaker but requires an occurrence of at least one of 4 specific vowels in the test sample. The proposed GMM-HDBC statistical method is text independent but rather than exact height, it assigns a height bin class representing a range of heights. This classification method also produces a complementary confidence measure. To utilize the complementary information produced by the two methods, a fusion system was also developed. The fusion system produces a single height estimate per speaker and improves the accuracy of the MFLTR regression method by utilizing the additional height bin class information and confidence score. An error analysis in the MFLTR, GMM-HDBC, and fusion systems was performed to provide better understanding of the respective performances. Compared to previous investigations on height estimation, these systems are at least equal or in most cases outperform previous methods in terms of MAE. The MFLTR method achieved an MAE of 4.93 cm and 4.76 cm, and the fusion method achieved an MAE of 4.89 cm and 4.55 cm for males and females from the TIMIT database, respectively.

Table 38: *Height Estimation MAEs of open independent subjects (males shown here, females also available in publication). Audio tracks taken from YouTube of actors/actresses from TV and movies with known/published heights from their biographies.*

	Actual Height (m)	MFLTR (m) [Error (cm)]	GMM-HDBC (m) [Error from median (cm)]
		Male	
Adam Baldwin	1.93	1.8647 (-6.53)	1.765-1.79 (-15)
Christian Kane	1.78	1.7396 (-4.04)	1.9-1.96 (+15.5)
David Boreanaz	1.85	1.8174 (-3.26)	1.765-1.79 (-7)
Jim Parsons	1.86	1.7955 (-6.45)	1.765-1.79 (-8)
Johnny Galecki	1.65	1.6480 (-.2)	1.57-1.715 (IN)
Nicholas Brendan	1.80	1.8413 (+4.13)	1.765-1.79 (-2)
Seth Green	1.63	1.6458 (+1.58)	1.57-1.715 (IN)
Simon Helburg	1.70	1.6769 (-2.31)	1.765-1.79 (+8)

Further details are presented in this submitted publication:

- J.H.L. Hansen, K. Williams, H. Boril, "Speaker Height Estimation from Speech: Fusing Spectral Regression and Statistical Acoustic Models," submitted to Journal of the Acoustical Society of America, Aug. 2013; Revised June 2014; Revised Nov. 2014. Accepted July 17, 2015.

5.1.9: References for Speaker Height Evaluation

- Arsikere, H., Leung, G., Lulich, S., and Alwan, A. (2012). "Automatic height estimation using the second subglottal resonance", IEEE ICASSP-2012, pp. 3989–3992, 2012.
- Arsikere, H., Leung, G. K., Lulich, S. M., Alwan, A. (2013). "Automatic estimation of subglottal resonances from adults speech with application to speaker height estimation", Speech Communication 55, 51 – 70.
- Crosmer, J. and Barnwell, T.P., I. (1985). "A low bit rate segment vocoder based on line spectrum pairs", IEEE ICASSP-1985, volume 10, 240–243.
- Dusan, S. (2005). "Estimation of speakers height and vocal tract length from speech signal", ISCA INTERSPEECH-2005, pp. 1989–1992, Lisbon, Portugal, 2005.
- Eide, E. and Gish, H. (1996). "A parametric approach to vocal tract length normalization", IEEE ICASSP-96, volume 1, 346–348.
- Eyben, F., Wollmer, M., and Schuller, B. (2009). "OpenEAR – introducing the Munich open-source emotion and affect recognition toolkit", Inter. Conf. Affective Computing & Intelligent Interaction, 2009, 1–6.
- Fitch, W. T. and Giedd, J. (1999). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging", Journal of the Acoustical Society of America 106, 1511–1522.
- Ganchev, T., Mporas, I., and Fakotakis, N. (2010a). "Audio features selection for automatic height estimation from speech", in Artificial Intelligence: Theories, Models and Applications, edited by S. Konstantopoulos, et al., vol. 6040 - Lecture Notes in Computer Science, 81–90 (Springer).
- Ganchev, T., Mporas, I., and Fakotakis, N. (2010b). "Automatic height estimation from speech in real-world setup", EUSIPCO'2010, 800–804 (Aalborg, Denmark).
- Greisbach, R. (1999). "Estimation of speaker height from formant frequencies", Forensic Linguistics 6, 265–277.
- Hansen, J. H. L. (1988). "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition", Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, (p. 47).
- Hasan, T., Sadjadi, O., Gang, L., Shokouhi, N., Boˇril, H., and Hansen, J. H. L. (2013). "CRSS systems for 2012 NIST Speaker Recognition Evaluation", in IEEE ICASSP 2013, 6783–6787 (Vancouver, Canada).
- Itakura, F. (1975). "Line spectrum representation of linear predictor coefficients of speech signals", Journal

of the Acoustical Society of America 57, 35.
- Kunzel, H. J. (1989). "How well does average fundamental frequency correlate with speaker height and weight?", Phonetica 46, 117–125.
- Lass, N. J. and Brown, W. S. (1978). "Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies", The Journal of the Acoustical Society of America 63, 1218–1220.
- Mporas, I. and Ganchev, T. (2009). "Estimation of unknown speakers height from speech", International Journal of Speech Technology 12, 149–160.
- Pellom, B. L. and Hansen, J. H. L. (1997). "Voice analysis in adverse conditions: the Centennial Olympic Park Bombing 911 call", Proceedings of the IEEE 40th Midwest Symposium on Circuits and Systems 1997, volume 2, 873–876.
- Perry, I., Brestoff, J., and Van der Broeck, J. (2011). "Challenging the role of social norms regarding body weight as an explanation for weight, height, and bmi misreporting biases: Development and application of a new approach to examining misreporting and misclassification bias in surveys", BMC Public Health 11, 331–341.
- Rabiner, L. and Schafer, R. (2011). Theory and Applications of Digital Speech Processing, chapter Algorithms for Estimating Speech Parameters, 548–645
- van Dommelen, W. A. and Moxness, B. H. (1995). "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour", Language and Speech 38, 267–287.
- Williams, K. and Hansen, J. (2013). "Speaker height estimation combining GMM and linear regression subsystems", IEEE ICASSP-2013, pp. 7552–7556, Vancouver, Canada, 2013.

5.2: Tool Development for Speech Corpus Analysis: Clipping and the ClipDaT toolkit

The focus in this area has been to finalized tool development to explore environment specific traits which could impact speech system performance. These would include: (i) clipping detection, (ii) signal drop-out detection, (iii) SNR or background noise classification, (iv) reverb/echo detection, etc. A full system/code delivery for ClipDaT was made to USAF in January 2015.

5.2.1: Peak Clipping - Overview

Audio peak clipping occurs when the volume of the audio signal being recorded exceeds the input voltage range of the microphone's pre-amplifier given the current gain for analog-to-digital (A/D) conversion. When this occurs, the pre-amplifier voltage becomes saturated, and unable to provide an accurate discrete representation for reliable A/D conversion. This causes the "peak" of the signal to not be reproduced by the pre-amplifier, which means this "peak" as well as all portions of the signal above the maximum voltage of the pre-amplifier to be clamped to the maximum of the signal as it passes through the analog-to-digital converter. This means that the natural shape of the speech waveform is not properly represented in the discrete signal; instead, a plateau appears, and the information contained in the higher amplitude samples is lost. The manifestation of this loss of data, and the introduction of this plateau shape comes in the form of non-linear distortion, especially in higher frequencies, resulting in audible artifacts in the recorded audio.

Naturalistic or uncontrolled recordings can contain clipping caused by many different sources, including loud but short impulse-type noises, such as hard impact sounds, many construction tool noises, a gunshot, and other unseen events such as screaming, sirens for police, fire, or ambulance, or car horns. These examples are more likely to be found in spontaneous, real-world data collections, but studio-recorded corpora are not free from this phenomenon. A subject coughing, sneezing, touching a microphone, or adjusting their clothing, if a lapel-mounted microphone is being used, can all cause clipping to be introduced into the recording. Due to the relatively uncontrolled nature of many of the audio sources being used to drive speech processing research today, clipping is an important phenomenon to investigate.

5.2.2: Peak Clipping – Analysis

The figure below (Fig. 88) provides graphical demonstrations of the impact of clipping for three different phoneme classes: vowel /a/, nasal /n/, and fricative /s/. Respectively, the /a/ phoneme was taken from the word "hot", /n/ from the end of "destination", and the /s/ from "east". These visuals were created by isolating the three phonemes in question, and increasing the gain of each until 10% of the audio samples migrated to an extreme.

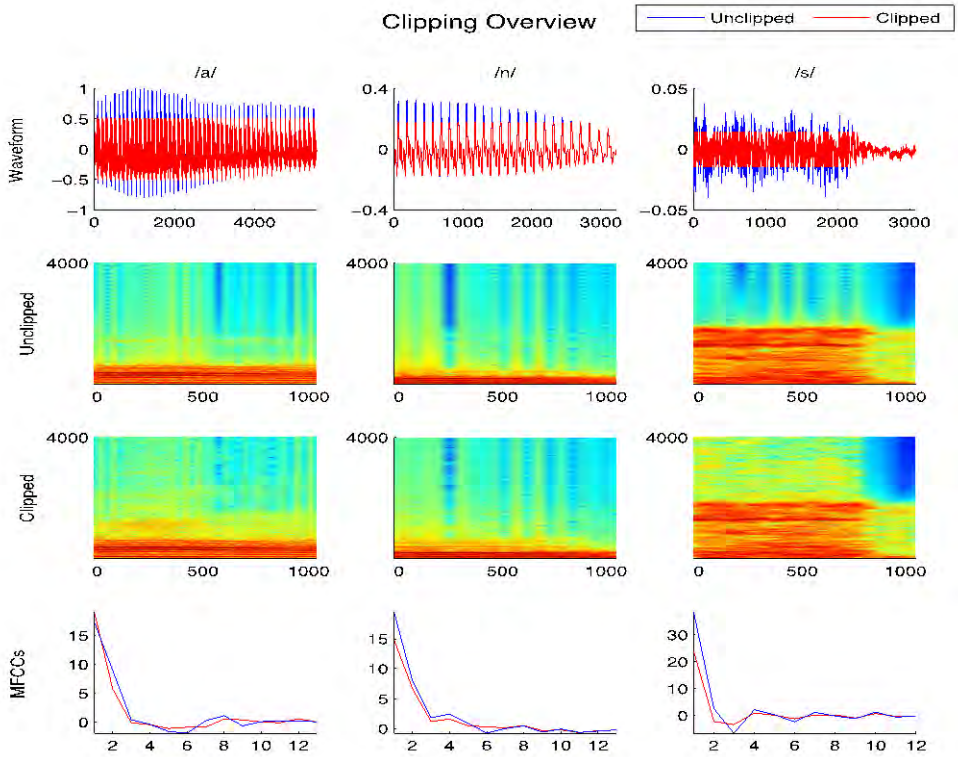


Figure 88: *Clipping Analysis: (i) time waveform, (ii) unclipped and clipped spectrogram analysis, and (iii) MFCC feature response for unclipped and clipped data*

The first row of Figure 88 presents time waveforms of the aforementioned phonemes, with blue representing the original unclipped signal, and red the corresponding clipped version. The second row contains spectrogram representations of the original unclipped waveforms in the first, versus clipped waveform spectrograms in the third row. Comparing the corresponding spectrogram plots of each phoneme, it is very clear that the largest impact of clipping is noise appearing in the higher frequencies of the spectrogram. It is interesting to note that the lower frequencies, and perhaps more importantly, the overall shape of the speech energy information remains, and the formant structure appears to be present and intact. This would imply that the audio is still very intelligible, and that, for this level of clipping at least, the content is not degraded to the point that it is not still easily understood.

Finally, the fourth row of the figure displays the overall average Mel-frequency cepstral component (MFCC) plots for these same three pairs of waveforms. These plots represent a small slice of the data that would be processed and used by classifiers in automatic speech-processing systems to represent speakers or phonemes. Comparisons across these plots provide some idea of how different clipped versus unclipped files would be interpreted by automatic speech systems. Clearly, there is also *some* impact on MFCC feature vectors due to clipping.

5.2.3: Peak Clipping – Impact on Speech Quality & Human Perception

A speech corpus was created to allow for controlled testing of clipping's impact on different factors. It was important to start from a clean source, so this corpus was made by artificially adding clipping to the TIMIT corpus, at four levels; 0.5%, 1%, 5%, and 10% clipping-to-speech ratio, which is the percentage of samples

recognized as speech by SAD that are clipped, meaning a higher number suffers from more clipping distortion. This audio was then evaluated with four speech quality measures: NIST signal-to-noise ratio, waveform amplitude distribution analysis (WADA) signal-to-noise ratio, sources-to-artifacts ratio (SAR) from Blind Source Separation Eval toolbox, and perceptual evaluation of speech quality (PESQ). A higher indicated score means a higher measured quality for each of these measures. In addition, a human-listening trial with 15 participants was performed, where they were presented with 20 sentences of varying clipping contamination, and asked to rate the audio quality of the utterances on a scale from 1-5, ranging from (1) Bad – Clipping/distortion is very annoying, to (5) Excellent – clipping/distortion is imperceptible.

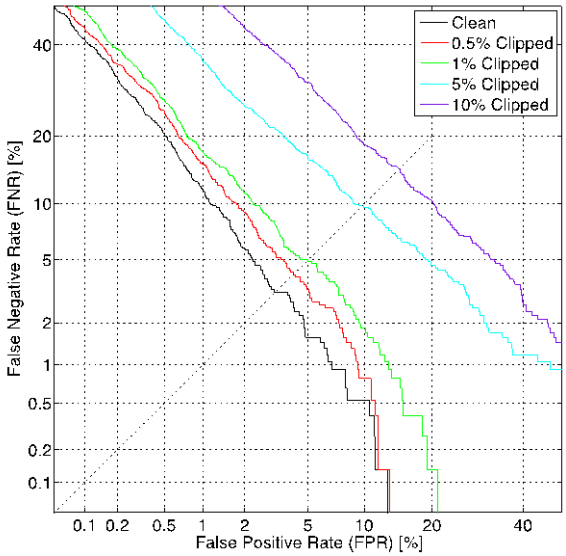
Table 39: *Comparison of speech quality measures for original clean speech, and various percentages (0.5%-10.0%) of waveform clipping for the same audio corpora. 6300 TIMIT sentences were processed for each entry in this table.*

	Quality Measure				
% Clipped Speech	NIST STNR	WADA SNR	BSS Eval (SAR)	PESQ	Human Assessment
Original	49.6	80.7	-	-	4.6
0.5%	48.8	83.8	17.9	3.9	-
1.0%	48.2	83.1	15.7	3.7	4.2
5.0%	44.6	77.7	10.4	3.1	4.0
10.0%	40.8	71.5	8.3	2.7	3.8

It is clear from the above results that clipping negatively impacts both automatic speech quality measures and human perception of speech. While human listeners were consistently more generous with their quality ratings than PESQ, which reports scores on almost the same scale, a similar downward trend is realized as the amount of clipping increases across all of the measures tested, further cementing the detrimental effect that clipping has on audio quality.

5.2.4: Peak Clipping – Impact on Speaker Recognition/Identification (SID)

To further investigate how clipping impacts speech audio signals, a number of speaker identification experiments were performed with the clipped-TIMIT corpus that was created. All trials were performed with an MFCC-based system using a GMM-UBM classifier with 256-mixture models. For each experiment, there were 380 target speakers, and a disjoint set of 250 speakers comprising the UBM. Two sentences per speaker were always used as the evaluation audio, with between two and eight sentences per speaker used as training data.



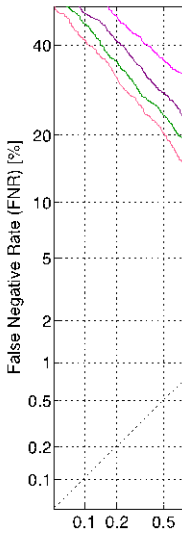
Test Data	EER (%)
Clean Original	3.2
0.5% Clipped	4.1
1% Clipped	5.0
5% Clipped	9.7
10% Clipped	14.6

Figure 89: *Clipping SID Evaluation: DET curve results: clean trained models; 5 test data configurations – clean, 0.5%, 1%, 5%, 10% clipped test data; Table summaries EER(%) rates.*

Figure 89 and the corresponding table show the results of this SID setup with models trained on eight clean TIMIT sentences, and tested against two sentences with varying amounts of clipping, as notated. It is clear that clipping very negatively impacts SID accuracy when it is present in the test audio, even when it is only present in amounts so low that humans report that it is barely perceived. This impact was much more severe than was expected, jumping from a 3.2% clean EER to 14.6% with 10% clipped test sentences.

The effect of clipping present in the audio used for training was also evaluated (see Figure 90). In this case, the test audio is clean/unclipped, and the training audio consists of eight sentences, of which between zero and six contain clipping at 10% clipping-to-speech ratio. As with the previous experiment, the performance impact is real and significant, more than doubling the EER when six of the eight training sentences contain clipping.

Train Data	EER (%)
8 Clean Sentences	3.2
6 Clean, 2 Clipped	3.9
4 Clean, 4 Clipped	5.0
2 Clean, 6 Clipped	7.6



6

7 FIGURE 90: CLIPPING SID EVALUATION: DET CURVE RESULT WITH FOUR LEVELS OF CLIPPED TRAIN DATA USED FOR SPEAKER ID ACOUSTIC MODEL CONSTRUCTION.

5.2.4: Peak Clipping – ClipDaT Algorithm and General Corpus Assessment

The final, major contribution from this project is the combination of the detection algorithm itself, as well as the creation of automation scripts for evaluating the clipping content of a corpus. These scripts process a directory and output both text labels detailing the location and amount of clipping in each audio file in the corpus, as well as a string of MATLAB commands that will create a graphical summary representation of this information. An example of this graphic is shown below, in this case detailing the clipping content of the Pan Arabic speech corpus, and consists of two parts. The upper portion, a pie chart, displays the total distribution between files that contain clipping, and those that are clean, along with the amount of actual samples within those files that are clipped. The lower portion of the graphic is a representation of *how* that clipping is distributed through the corpus. The image is created by representing all of the 'audio space' contained within the blue portion of the pie chart as white space, where progressively darker red pixels represent clipping events increasing in severity/length. The files are sorted from left to right from most to least clipping events, meaning that if only the left side of the graphic contains colored pixels, the majority of the clipping in the corpus is constrained to a small proportion of problem files, with just a small amount of clipping in the rest of the contaminated files, as is the case below. If instead, the colored pixels are spread more evenly, then the clipping is more equally distributed throughout the clipped files. Heavy banding indicates long strings of clipping within files, and should be investigated further.

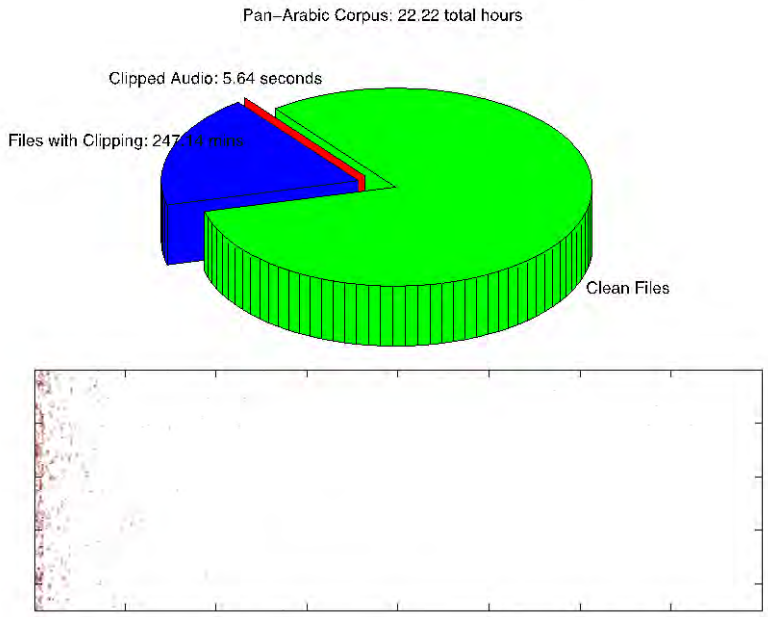


Figure 91: *Clipping SID Evaluation: Graphical representation of the amount and distribution of clipping within the Pan-Arabic speech corpus.*

5.3: Gender ID: robust speaker trait estimation.

In this task, the effort was performed based on a feedback request from USAF during a project update. The focus was to develop an effective gender ID for noisy/channel corrupted speech in various languages such as the DARPA RATS corpus.

5.3.1: Gender Identification Experiments on RATS data: Using Unsupervised Domain Adaptation to Improve Gender ID Accuracy

- An unsupervised domain adaptation strategy was implemented to overcome the lack of any labeled development data for RATS test utterances. An unsupervised convex max-margin clustering algorithm was first used to assign labels to unlabelled RATS utterances (about 500 utterances per channel), and the labels were subsequently used to adapt the PLDA parameters derived from the Fisher English (FE) training data. The adapted PLDA model was then used in Gender ID experiments on the RATS test utterances.
- On the RATS test utterances, the average weighted accuracy increased from 76.48% to 81.73% (after unsupervised domain adaptation). The relative gain in accuracy was +6.86%
- The average Equal Error Rate (EER) decreased from 20.89% to 17.80% as a result of the unsupervised domain adaptation. The relative decrease in EER stood at 14.75%.
- Table 40 shows results of the Gender ID experiments using 400-dimensional I-vectors and a PLDA backend. The results show the classification accuracy and EER before, and after carrying out the unsupervised domain adaptation strategy, with a channel-wise breakdown. The tests were carried out on complete (entire-duration) test utterances.
- The unsupervised domain strategy improved the Gender ID system's accuracy on all RATS channel test utterances except those from channels C, E and SRC (clean source).

Table 40: *Channel-wise classification accuracy, and EER of the i-Vector PLDA based Gender ID system. "Before" and "After" refer respectively to the results without, and with the adapted PLDA model (obtained using unsupervised domain adaptation).*

Channel	Accuracy-Before	Accuracy-After	EER-Before	EER-After
A	61.22	79.59	32.65	20.41
B	73.47	79.59	23.47	20.41
C	77.55	77.55	21.43	21.43
D	75.51	77.55	19.39	21.43
E	61.22	61.22	37.76	37.76
F	81.63	89.80	17.35	10.20
G	95.92	100	4.08	0.00
H	65.22	73.91	29.35	26.09
SRC	95.92	95.92	3.06	3.06
AVERAGE	**76.48**	**81.73**	**20.89**	**17.80**

5.3.2: Gender Identification System Development: Investigation of Duration Mismatch with very short Fisher English (FE) Test utterances

- To observe the effect of duration mismatch on the i-Vector based Gender ID system, the system trained on FE (complete) utterances was tested on shorter duration segments selected randomly from the complete FE test-set utterances. The results indicate severe degradation in both classification accuracy and EER.
- Table 41 shows the results of the Gender ID experiments using the system trained on whole utterances from the FE data, and tested respectively on the complete FE test segments, and test segments of shorter duration. A PLDA back-end was used for these experiments.
- To overcome the severe degradation observed in Table 42, the i-Vector based Gender ID systems were retrained using training utterances of the same length as the corresponding test utterances, by randomly selecting the desired length utterances from the complete FE training utterances. The dimensionality of the i-Vector was also adjusted to account for the shorter training and testing segments.
- Table 42 shows the results with the retrained I-vectors based Gender ID systems. The test segments were kept the same as in Table 41. Very significant improvements can be observed across all test-segments as compared to the results of Table 41. A PLDA back-end was used for these experiments.

Table 41: *Classification accuracy and EER obtained using the Gender ID system trained using complete FE utterances, on test-sets of different duration in testing*

Test Segment	Accuracy	EER
Complete	97.62	2.31
20s	82.12	17.85
15s	81.00	18.87
10s	77.77	22.10
3s	65.58	34.33

Table 42: *Gender Identification Results using test segments of various duration from the FE test-set, with the Gender ID system trained on same-length utterances as the test segments.*

Test Segment	Trained On	Accuracy	EER	i-Vector Dim
20s	20s	96.15	3.85	200
10s	10s	94.85	5.15	200
3s	3s	91.27	8.67	200

5.3.3: Gender Identification System Development: Comparison of i-Vector vs. GMM-UBM based Gender ID systems

- The i-Vector based Gender ID system was compared against a GMM-UBM based Gender ID approach for various duration test utterances from the FE test-set. Table 43 shows the results of the i-Vector based Gender ID system compared against the GMM-UBM gender ID approach.
- The GMM-UBM Gender ID systems were trained using MFCC-SDC features, similar to the i-Vector based systems.
- For the GMM-UBM system, Maximum-A-Posteriori (MAP) adaptation was used to obtain two gender-specific GMMs, which were then used to perform the Gender ID experiments on the test utterances from the FE test-set.
- The i-Vector PLDA based Gender ID system is shown to consistently outperform a GMM-UBM based approach to Gender ID across all the FE test-sets.
- Gender Separability in the i-Vector space was also investigated. Figure 92 shows the first two dimensions of Maximization of Mutual Information (MMI) based projection of 400-dimensional test i-Vectors form the FE corpus. Clearly, the i-Vectors from female (in black) and male (in red) test utterances are well separated.

Table 43: *Comparison of i-Vector based Gender ID system against a GMM-UBM based system on the FE test-set segments of duration 3s, 10s, 20s, and complete utterances. The training segments were of the same length as the corresponding test-segments.*

Test Duration	Accuracy		EER	
	i-Vec PLDA	GMM UBM	i-Vec PLDA	GMM UBM
3s	91.27	90.73	8.67	9.00
10s	94.85	93.65	5.15	6.23
20s	96.15	94.62	3.85	4.69
Complete	97.62	95.23	2.31	4.46

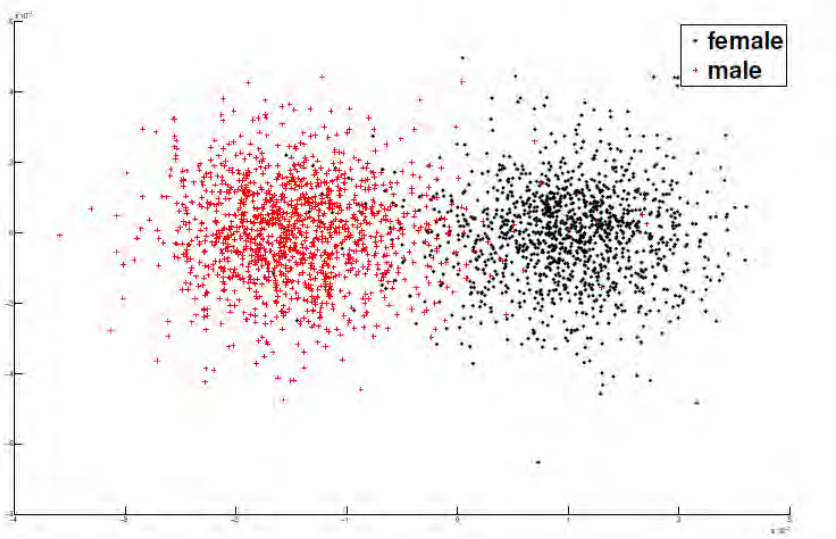


Figure 92: *First 2 dimensions of an MMI based projection of 400-dimensional i-Vectors of the Fisher English female (in black) and male (in red) test utterances. There were 2600 test utterances in total.*

5.3.4: Gender Identification: Publications

[1] S. Ranjan, G. Liu, J.H.L. Hansen, "i-Vector based Gender Identification for Severely Noisy and Multilingual DARPA RATS data," submitted to *ISCA-INTERSPEECH*, 2015.

5.4: Environmental & Speaker based Normalization:

In this area, the focus has been to explore signal processing strategies that can improve robustness for environmental factors, as well as speaker based variability. The effort included focused tasks on reverberation and vocal effort, with applications to large vocabulary speech recognition.

- **5.4.1: ASR on Reverberation and Increased Vocal Effort:** Channel, noise, and talking style mismatch are major sources of automatic speech recognition (ASR), and speaker identification (SID), and language and dialect identification (LID/DID) degradation. In this study [Boril12], it has been shown that a combination of cepstral distribution normalizations such as the quantile dynamics cepstral normalization (QCN) or cepstral gain normalization (CGN) with our modified low-pass RASTA filtering (RASTALP) is efficient in addressing acoustic mismatch conditions in ASR under reverberation and Lombard effect.

- **5.4.2: RASTALP Extension to Further Reduce Transient Effects:** CRSS-UTDallas recently proposed RASTALP filtering, which was shown to sufficiently alleviate a major portion of the transient distortions caused by the original RASTA filter. However, there is still a noticeable signal distortion due to the weighted *unit step* seen by the filter at the beginning of the processed signal. To address this, we have proposed a simple IIR filter pre-buffering scheme that completely eliminates transient distortions at the beginning of the processed signal (see Figure 93).

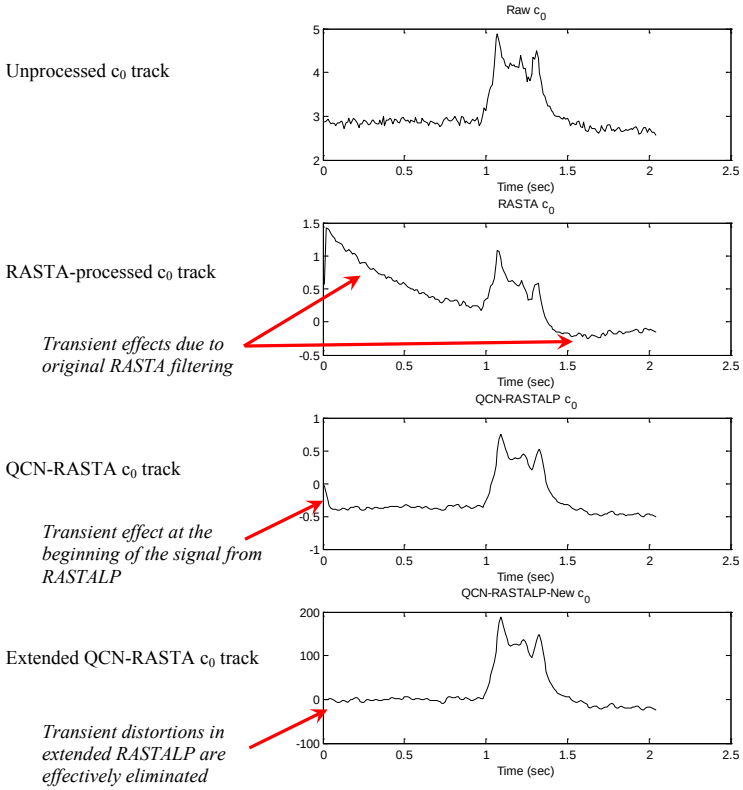


Figure 93: *Comparison of transient effects in traditional RASTA, and proposed RASTALP, and RASTALP with pre-buffering.*

- **5.4.3: Integration of CGN and QCN-RASTALP in Kaldi, Matalb, and Sphinx:** during this period, a software tool was developed that allows for effective integration of cepstral gain normalization (CGN), and QCN-RATSALP with the open source Kaldi toolkit as well as Sphinx and the Matlab environment.

- **5.4.4: Evaluation of normalizations in Kaldi tasks:** Experiments were performed to evaluate the now developed QCN-RASTALP, which includes pre-buffering on numerous ASR tasks including Switchboard/HUB5 task and CU-Move task with great success. Experimental results for the CU-Move corpus are detailed in Tables 44 and 45 [Hahm13]. The results are compared to the performance of speaker-based CMN and the best QCN-RASTALP setup (q = 10%). It can be seen that CGN outperform CMN in both setups with and without LDA+MLLT being applied in the feature front-end while QCN-RASTALP maintains the lowest WERs of all the three normalizations. The superior performance of QCN-RASTALP can be attributed to two factors – (i) more accurate alignment of the dynamic ranges of non-Gaussian cepstral distributions (low cepstral coefficients c0–c2 that reflect the energy and spectral slope of the speech signal tend to be multimodal and are very sensitive to the presence of environmental noise) and (ii) the employment of RASTALP filtering that alleviates the impact of non-stationary noises.

Table 44: *Performance of CMVN, CGN, and QCN-RASTALP on CU-Move task.*

	WER (%)		
Features	CMN	CGN	QCN-RASTALP
Plain	40.57	40.5	39.22
LDA+MLLT	38.61	37.39	37.24

Table 45: *Comparison of CMN and QCN-RASTA$_{LP}$ normalizations in the combination with model adaptation (LDA+MLLT front-end); word error rates - WER (%).*

Norm	Adapt	Baseline	Average amount of adaptation data (seconds)								
			5	8	15	36	69	135	331	647	848
CMN	MAP	38.6	38.8	38.9	38.8	39.0	39.1	39.1	38.9	38.6	38.3
	fMLLR		75.7	57.4	47.0	40.4	38.4	37.6	36.8	36.5	36.4
QCN-RASTA$_{LP}$	MAP	37.2	37.4	37.3	37.3	37.4	37.5	37.5	37.3	37.1	37.0
	fMLLR		58.3	43.1	38.5	36.8	36.3	36.0	35.7	35.5	35.4

VI. Conclusions:

This 36 month project has addressed five core research tasks, resulting in many scientific and engineering advancements, including code/algorithms, features, corpora, and best practices. Software and toolkits including MS-AcID and CLIP-DaT were developed and delivered.

Task 1 - Speaker ID (SID) Robustness: Speaker identification is known to be sensitive to mismatch in train/test conditions. Advancements here have addressed a range of issues relating to speaker, language, style, environment, and distortion mismatch which taken collectively have advanced the state of the art in speaker recognition. A total of 12 Journal Papers, 20 Conference Papers, and 5 PhD theses were produced from this effort – all of which have been delivered to the AFRL.

Task 2 - Open-Set Language ID (LID) / Dialect ID (DID): The focus here has been to develop new features, mismatch compensation schemes, and classification strategies for language identification. The outcomes from this effort include the toolkit: MS-AcID, which supports multiple backend i-Vector LID and SID, as well as data purification tools for harvesting multi-lingual audio in open-set LID. Advancements in the domain of open-set language rejection represent some of the first steps in the field in addressing the ability to more effectively turn away unwanted audio streams from languages which are not of interest. A total of 3 Journal Papers, 5 Conference Papers, and 2 PhD theses were produced from this effort – all of which have been delivered to the AFRL.

Task 3 – Co-Speaker Diarization/Environment (CoSpkrD): The ability to detect the presence of co-channel speech for usable speech detection as well as speaker-separation for subsequent speech systems is important for SID, LID, ASR, KWS, and other speech applications. In this area, the overlap speech detection has proven to be very challenging, yet algorithm advancements have resulted in two methods that have shown promise for detection and mitigation for improved speaker ID. In terms of diarization, numerous advancements have been made including the threshold optimized Combo-SAD (TO-Combo-SAD) which represents the most effective unsupervised SAD in the field today. Diarization advancements include detection of primary/secondary speaker, and effective word count estimation in unrestricted massive audio streams. A total of 1 Journal Paper, 12 Conference Papers, and 1 PhD thesis were produced from this effort – all of which have been delivered to the AFRL.

Task 4 – Automatic Speech Recognition/Keyword Spotting: (ASR/KWS): In this area, the formulation of next generation techniques for speech recognition have included articulatory, phonological, and prosody based detection solutions. Such knowledge has been integrated into systems to address problems in speech recognition, as well as keyword spotting (KWS) in new and emerging domains. English as well as other languages of interest (Arabic, Farsi) have been addressed, as well as robustness methods to improve performance in real audio data stream scenarios. While the emphasis here has been on KWS, these advancements have also been explored for unsupervised open-word-set based Speaker ID (SID) and Language/Dialect ID (LID/DID). A total of 2 Journal Papers, 11 Conference Papers, and 2 MS-EE/CS theses were produced from this effort – all of which have been delivered to the AFRL.

Task 5 - Speaker State Assessment/Environmental Sniffing (SSA/EnvS): The advancements here have focused on the development of new algorithms to assess unique knowledge regarding speakers within audio streams – sub-areas include physical speaker traits (i.e., height, gender), speaking style scenario/room (i.e., read, spontaneous, whisper, distant based speech), situational speaker state (i.e., physical task stress state, or general sentiment/emotional outlook). Automatic assessment of audio streams for nonlinear distortion was also considered. The primary outcome from that effort was the Clip-DaT toolkit, which allows for assessment of the degree of speech clipping in entire audio corpus collections. The resulting visualization methods also provide an effective and easy way to compare large corpora, as well as understand the impact of clipping on both training and test data for applications such as speaker ID. A total of 5 Journal Papers, 12 Conference Papers, 1 PhD thesis, and 4 MS-EE/CS theses were produced from this effort – all of which have been delivered to the AFRL.

VII. References:

Due to the wide scope of this five task effort, references corresponding to each of the task domains have been summarized within each Task description. What follows here is a summary of the publications which have been produced, as well as students supported and PhD/MS theses produced from this 36 month research effort.

PUBLICATIONS & STUDENTS COMPLETED: (2012-2015)
(Journal Papers, Conference Papers, PhD/MS Theses)

In total, the research stemming from the 3 year project covering 5 task topics have resulted in 19 peer-reviewed journal papers appearing, 60 peer-reviewed conference papers appearing, as well as the completion of 8 PhD's and 6 MS thesis students. Below is a summary of the publications and completed students, along with where they are presently employed in the United States.

Task 1 - Speaker ID (SID) Robustness:

- 12 Journal Papers; 20 Conference Papers; 5 PhD thesis; 0 MS-EE thesis

Task 2 - Open-Set Language ID (LID) / Dialect ID (DID):

- 3 Journal Paper; 5 Conference Papers; 2 PhD thesis; 0 MS-EE thesis

Task 3 – Co-Speaker Diarization/Environment (CoSpkrD):

- 1 Journal Paper; 12 Conference Papers; 1 PhD thesis; 1 MS-EE thesis

Task 4 – Automatic Speech Recognition/Keyword Spotting: (ASR/KWS):

- 2 Journal Papers; 11 Conference Papers; 0 PhD thesis; 2 MS-EE/CS thesis.

Task 5 - Speaker State Assessment/Environmental Sniffing (SSA/EnvS):

- 5 Journal Papers; 12 Conference Papers; 1 PhD thesis; 4 MS-EE/CS thesis

Journal Papers: [Based on this USAF support from 2012-2015] (17)

[1.] S.O. Sadjadi, J.H.L. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust Speaker and Language Identification," *Speech Communication*, accept April 2015. Vol. xxx, No. x, pp. xxxx-xxxx, Aug.. 2015

[2.] M. Mehrabani, J.H.L. Hansen, "Automatic Dialect Separation Assessment," *Inter. Journal of Speech Technology*, vol. 17, issue 4, pp. xxx-xxx, xxxx. DOI 10.1007/s10772-014-9268-y [ACCEPTED Nov. 2014] ISSN 1381-2416

[3.] C. Zhang, J.H.L. Hansen, "An Advanced Entropy-based Feature with Frame-Level Vocal Effort Likelihood Space Modeling for Distant Whisper-Island Detection," *Speech Communication*, vol. 66, pp. 107-117, 2015.

[4.] G. Liu, J.H.L. Hansen, "An Investigation into Back-End Advancements for Speaker Recognition in Multi-Session and Noisy Enrollment Scenarios," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1978-1992, Dec. 2014 [**Paper Selected as Highlight paper for Front – Cover of IEEE Trans. ASLP for this issue**]

[5.] J.H.L. Hansen, A. Kumar, P. Angkititrakul, "Environment Mismatch Compensation Using Average Eigenspace Based Methods For Robust Speech Recognition," *Inter. Journal of Speech Technology*, vol. 17, issue 4, pp. 353-365, 2014. DOI 10.1007/s10772-014-9233-9

[6.] S.O. Sadjadi, J.H.L. Hansen, "Blind Spectral Weighting for Robust Speaker Identification under Reverberation Mismatch Conditions," *IEEE Trans. Audio, Speech and Language Processing*, vol. 22, no. 5, pp. 935-943, May. 2014.

[7.] J.H.L. Hansen, J.-W. Suh, P. Angkititrakul, Y. Lei, "Effective Background Data Selection for SVM-Based Speaker Recognition for Unseen Test Environments: More is Not Always Better," *Inter. Journal Speech Technology*, vol. 17, issue 3, pp. 211-221, Sept. 2014.

[8.] T. Hasan, J.H.L. Hansen, "Maximum Likelihood Acoustic Factor Analysis Models for Robust Speaker Verification in Noise," *IEEE Trans. Audio, Speech and Language Processing*, vol. 22, no. 3, pp. 381-391, Feb. 2014 **[#2 top downloaded IEEE ASLP paper for Feb. 2014]**

[9.] T. Hasan, H. Boril, A. Sangwan, J.H.L. Hansen, "Multi-modal highlight generation for sports videos using an information-theoretic excitability measure," *EURASIP Journal of Advancements in Signal Processing*, vol. 173, (pg. 1-17), 2013.

[10.] Y. Chung, J.H.L. Hansen, "Compensation of SNR and noise type mismatch using an environmental sniffing based speech recognition solution," *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2013, No.doi:10.1186/1687-4722-2013-12, pp.1-14, June 2013

[11.] F. William, A. Sangwan, J.H.L. Hansen, "Automatic Accent Assessment Using Phonetic Mismatch and Human Perception," IEEE Trans. Audio, Speech and Language Processing, vol. 21, no. 9, pp. 1818-1828, Sept. 2013

[12.] J.H.L. Hansen, J.-W. Suh, M.R. Leonard, "In-set/out-of-set speaker recognition in sustained acoustic scenarios using sparse data," Speech Communication, vol. 55, pp. 769-781, April 2013.

[13.] O. Sadjadi, J.H.L. Hansen, "Unsupervised Speech Activity Detection using Voicing Measures and Perceptual Spectral Flux," IEEE Signal Processing Letters, vol. 20, no. 3, pp. 197-200, March 2013

[14.] T. Hasan, J.H.L. Hansen, "Acoustic Factor Analysis for Robust Speaker Verification," IEEE Trans. Audio, Speech and Language Processing, vol. 21, no. 4, pp. 842-853, April 2013

[15.] M. Mehrabani, J.H.L. Hansen, "Singing Speaker Clustering Based on Subspace Learning in the GMM Mean Supervector Space," Speech Communication, vol. 55, pp. 653-666, Feb. 2013.

[16.] X. Fan, J.H.L. Hansen, "Acoustic Analysis and Feature Transformation from Neutral to Whisper for Speaker Identification within Whispered Speech Audio Streams," Speech Communication, vol. 55, pp. 119-134, Jan. 2013

[17.] J.H.L. Hansen, E. Ruzanski, H. Boril, J.. Meyerhoff, "TEO-based Speaker Stress Assessment using Hybrid Classification and Tracking Schemes," Inter. Journal Speech Technology, vol. 15, issue 3, pp. 295-311,Sept. 2012.

[18.] J.-W. Suh, J.H.L. Hansen, "Acoustic Hole Filling for Sparse Enrollment Data using a Cohort Universal Corpus for Speaker Recognition," Journal of the Acoustical Society of America, vol. 131, no. 6, pp. 1515-1528, Feb. 2012.

[19.] A. Sangwan, J.H.L. Hansen, "Automatic Analysis of Mandarin Accented English using Phonological Features," Speech Communication, vol. 54, no. 1, pp. 40-54, Jan. 2012.

Conference Papers: **[Based on this USAF support from 2012-2015]** (60)

[These papers were all supported either fully or partially from USAF funding from 2012-2015. They are full 4-5 page papers submitted online, with 3-5 blind external reviewers completing reviews with authors submitting revised/updated papers if initially accepted].

[1.] S. Ghaffarzadegan, Hynek Boril**,** J.H.L. Hansen, "Generative Modeling of Pseudo-Target Domain Adaptation Samples for Whispered Speech Recognition," IEEE ICASSP-2015 *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, Paper#3612, pp. 5024-5028, Brisbane, Australia, April 19-24, 2015.

[2.] A. Ziaei, A. Sangwan, L. Kaushik, J.H.L. Hansen, "Prof-Life-Log: Analysis and Classification of Activities in Daily Audio Streams," IEEE ICASSP-2015 *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*,, Paper#3849, pp. 4719-4723, Brisbane, Australia, April 19-24, 2015

[3.] N. Shokouhi, A. Ziaei, A. Sangwan, J.H.L. Hansen, "Robust Overlapped Speech Detection and its Application in Word-Count Estimation for Prof-Life-Log Data," IEEE ICASSP-2015 *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*,, Paper#3895, pp. 4724-4728, Brisbane, Australia, April 19-24, 2015

[4.] M.M. Saleem, G. Liu, J.H.L. Hansen, "Weighted Training for Speech Under Lombard Effect for Speaker Recognition," IEEE ICASSP-2015 *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*,, Paper#3979, pp. 4350-4354, Brisbane, Australia, April 19-24, 2015

[5.] M.K.. Nandwana, A. Ziaei, J.H.L. Hansen, "Robust Unsupervised Detection of Human Screams in Noisy Acoustic Environments," IEEE ICASSP-2015 *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, Paper#4026, pp. 161-164, Brisbane, Australia, April 19-24, 2015

[6.] A. Misra, J.H.L. Hansen, "Spoken Language Mismatch in Speaker Verification: An Investigation with NIST-SRE and CRSS Bi-Ling Corpora," *IEEE SLT-2014: Spoken Language Technology Workshop*, paper PT3.2, pp. 372-377, Lake Tahoe, Dec. 7-10, 2014.

[7.] Q. Zhang, J.H.L. Hansen, "Training Candidate Selection for Effective Rejection in Open-Set Language Identification," *IEEE SLT-2014: Spoken Language Technology Workshop*, paper PT3.4, pp. 384-389, Lake Tahoe, Dec. 7-10, 2014.

[8.] G. Liu, C. Yu, N. Shokouhi, A. Misra, H. Xing, J.H.L. Hansen, "Utilization of Unlabeled Development Data for Speaker Verification," *IEEE SLT-2014: Spoken Language Technology Workshop*, paper PT3.10, pp. 419-423, Lake Tahoe, Dec. 7-10, 2014.

[9.] S.M. Mirsamadi, J.H.L. Hansen, "Multichannel Feature Enhancement in Distributed Microphone Arrays for Robust Distant Speech Recognition in Smart Rooms," *IEEE SLT-2014: Spoken Language Technology Workshop*, paper PW1.4, pp. 507-512, Lake Tahoe, Dec. 7-10, 2014. [**SLT-2014 Best Paper Award**]

[10.] S.M. Mirsamadi, J.H.L. Hansen, "Multichannel Speech Dereverberation Based on Convolutive Nonnegative Tensor Factorization for ASR Applications," *ISCA Interspeech-2014*, Paper #414, Singapore, Sept. 14-18, 2014.

[11.] N. Shokouhi, S.O. Sadjadi, J.H.L. Hansen, "Co-channel Speech Detection via Spectral Analysis of Frequency Modulated Sub-bands," *ISCA Interspeech-2014*, Paper #437, Singapore, Sept. 14-18, 2014.

[12.] M.K. Nandwana, J.H.L. Hansen, "Analysis and Identification of Human Scream: Implications for Speaker Recognition," *ISCA Interspeech-2014*, Paper #974, Singapore, Sept. 14-18, 2014.

[13.] A. Ziaei, L. Kaushik, A. Sangwan, J.H.L. Hansen, "Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions," *ISCA Interspeech-2014*, Paper #994, Singapore, Sept. 14-18, 2014.
[14.] A. Ziaei, A. Sangwan, J.H.L. Hansen, "A Speech System for Estimating Daily Word Counts," *ISCA Interspeech-2014*, Paper #1028, Singapore, Sept. 14-18, 2014.
[15.] C. Yu, G. Liu, J.H.L. Hansen, "Acoustic Feature Transformation using Unsupervised LDA for Speaker Recognition," *ISCA Interspeech-2014*, Paper #1288, Singapore, Sept. 14-18, 2014.
[16.] S. Ghaffarzadegan, H. Boril, J.H.L. Hansen, "Model and Feature Based Compensation for Whispered Speech Recognition," *ISCA Interspeech-2014*, Paper #1383, Singapore, Sept. 14-18, 2014.
[17.] G. Liu, C. Yu, A. Misra, N. Shokouhi, J.H.L. Hansen, "Investigating State-of-the-Art Speaker Verification in the case of Unlabeled Development Data," *ISCA Odyssey-2014 Workshop on Speaker and Language Recognition*, Joensuu, Finland, June 16-29, 2014.
[18.] Q. Zheng, G. Liu, J.H.L. Hansen, "Robust Language Recognition Based on Diverse Features," *ISCA Odyssey-2014 Workshop on Speaker and Language Recognition*, Joensuu, Finland, June 16-29, 2014.
[19.] G. Liu, J.H.L. Hansen, "Supra-Segmental Feature Based Speaker Trait Detection," *ISCA Odyssey-2014 Workshop on Speaker and Language Recognition*, Joensuu, Finland, June 16-29, 2014.
[20.] C. Yu, G. Liu, S.-J. Hahm, J.H.L. Hansen, "Uncertainty Propagation in Front End Factor Analysis For Noise Robust Speaker Recognition," *IEEE ICASSP-2014, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, pp. 4050-4054, (paper #1569862409), Florence, Italy, May 4-9, 2014
[21.] S. Ghaffarzadegan, H. Boril, J.H.L. Hansen, "UT-VOCAL EFFORT II: Analysis and Constrained-Lexicon Recognition of Whispered Speech," *IEEE ICASSP-2014, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, pp. 2563-2567, (paper #1569862901), Florence, Italy, May 4-9, 2014
[22.] A. Sangwan, L. Kaushik, C. Yu, J.H.L. Hansen, D.W. Oard, "'Houston, We have a solution' : Using NASA Apollo Program to advance Speech and Language Technology," ISCA INTERSPEECH-2013, pp. 1135-1139, Lyon, France, August 25-29, 2013 [Paper 1434]
[23.] T. Hasan, J.H.L. Hansen, "Acoustic Factor Analysis based Universal Background Model for Robust Speaker Verification in Noise," ISCA INTERSPEECH-2013, pp. 3127-3131, Lyon, France, August 25-29, 2013 [Paper 1019].
[24.] H. Boril, Q. Zhang, P. Angkititrakul, J.H.L. Hansen, D. Xu, J. Gilkerson, J.A. Richards, "A Preliminary Study of Child Vocalization on a Parallel Corpus of US and Shanghainese Toddlers," ISCA INTERSPEECH-2013, pp. 2405-2409, Lyon, France, August 25-29, 2013 [Paper 1445].
[25.] K. Godin, S.O. Sadjadi, J.H.L. Hansen, "Impact of Noise Reduction and Spectrum Estimation on Noise Robust Speaker Identification," ISCA INTERSPEECH-2013, pp. 3656-3660, Lyon, France, August 25-29, 2013 [Paper 512].
[26.] M. Mehrabani, J.H.L. Hansen, "Dimensionality Analysis of Singing Speech Based on Locality Preserving Projections," ISCA INTERSPEECH-2013, Lyon, France, August 25-29, 2013 [Paper 489].
[27.] W. Kim, J.H.L. Hansen, "An Advanced Feature Compensation Method Employing Acoustic Model With Phonetically Constrained Structure,"IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 7083-7086, (paper #4352), Vancouver, Canada, May 26-31, 2013.
[28.] O. Sadjadi, J.H.L. Hansen, "Robust Front-End Processing For Speaker Identification Over Extremely Degraded Communication Channels," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 7214-7218, (paper #5002), Vancouver, Canada, May 26-31, 2013.
[29.] N. Shokouhi, A. Sathyanarayana, O. Sadjadi, J.H.L. Hansen, "Overlapped-Speech Detection with Applications to Driver Assessment for In-Vehicle Active Safety Systems," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 2834-2838, (paper #5132), Vancouver, Canada, May 26-31, 2013.
[30.] K.A. Williams, J.H.L. Hansen, "Speaker Height Estimation Combining GMM and Linear Regression Subsystems," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 7552-7556, (paper #5154), Vancouver, Canada, May 26-31, 2013.
[31.] T. Hasan, O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, J.H.L. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 6783-6787, (paper #5335), Vancouver, Canada, May 26-31, 2013.

[32.] G. Liu, T. Hasan, H. Boril, J.H.L. Hansen, "An Investigation on Back-End for Speaker Recognition in Multi-Session Enrollment,," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 7755-7759, (paper #5488), Vancouver, Canada, May 26-31, 2013.
[33.] A. Ziaei, A. Sangwan, J.H.L. Hansen, "PROF-LIFE-LOG: Personal Interaction Analysis for Naturalistic Audio Streams," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 7770-7774, (paper #5517), Vancouver, Canada, May 26-31, 2013 (pdf [1.01MB]).
[34.] L.N. Kaushik, A. Sangwan, J.H.L. Hansen, "Sentiment Extraction From Natural Audio Streams," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 8485-8489, (paper #5538), Vancouver, Canada, May 26-31, 2013.
[35.] Q. Zhang, H. Boril, J.H.L. Hansen, "Supervector Pre-Processing for PRVM-Based Chinese and Arabic Dialect Identification," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 7363-7367, (paper #5548), Vancouver, Canada, May 26-31, 2013.
[36.] C. Yu, K. Wojcicki, P. Loizou, J.H.L. Hansen, "A New Mask-Based Objective Measure for Predicting the Intelligibility of Binary Masked Speech," IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 7030-7033, (paper #5002), Vancouver, Canada, May 26-31, 2013.
[37.] T. Hasan, R. Saeidi, J.H.L. Hansen, D. Van Leeuwen, "Duration Mismatch Compensation for i-Vector based Speaker Recognition," IEEE ICASSP-2013, IEEE ICASSP-2013, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 7663-7667, (paper #5355), Vancouver, Canada, May 26-31, 2013.
[38.] H. Boril, A. Ziaei, J.H.L. Hansen, "Prof-Life-Log: Production of Conversational Speech as a Function of Varying Environment," LENA Foundation Conference, Denver, CO, April 28-30, 2013.
[39.] A. Ziaei, A. Sangwan, J.H.L. Hansen, "Prof-Life-Log: Robust Audio Environment Detection for Naturalistic Audio Streams Using the LENA Device," LENA Foundation Conference, Denver, CO, April 28-30, 2013
[40.] H. Boril, A. Sangwan, J.H.L. Hansen, "Arabic Dialect Identification - 'Is the Secret in the Silence?' and Other Observations," ISCA Interspeech-2012, Mon-O1b-01, pg. 1-4,, Portland, OR, Sept. 9-13, 2012
[41.] S.O. Sadjadi, J.H.L. Hansen, "Mean Hilbert Envelope Coefficients (MHEC) for Robust Speaker Recognition," ISCA Interspeech-2012, Wed-O7b-05, pg. 1-4, Portland, OR, Sept. 9-13, 2012
[42.] K. Godin, T. Hasan, J.H.L. Hansen, "Physical Task Stress Detection using Acoustic Features Inspired by Glottal Waveform Analysis," ISCA Interspeech-2012, Wed-SS6-15, pg. 1-4, Portland, OR, Sept. 9-13, 2012.
[43.] T. Hasan, J.H.L. Hansen, "Integrated Feature Normalization and Enhancement for robust Speaker Recognition using Acoustic Factor Analysis," ISCA Interspeech-2012, Wed-P6d-04, pg. 1-4, Portland, OR, Sept. 9-13, 2012.
[44.] W. Kim, J.H.L. Hansen, "Gaussian Map based Acoustic Model Adaptation using Untranscribed Data for Speech Recognition in Severely Adverse Environments," ISCA Interspeech-2012,Wed-P7b-02, pg. 1-4,, Portland, OR, Sept. 9-13, 2012.
[45.] T. Hasan, J.H.L. Hansen, "Front-end Channel Compensation using Mixture-dependent Feature Transformation for i-Vector Speaker Recognition," ISCA Interspeech-2012, Tue-O5b-02, pg. 1-4, Portland, OR, Sept. 9-13, 2012.
[46.] M. Mehrabani, J.H.L. Hansen, "Speaker Clustering for a Mixture of Singing and Reading," ISCA Interspeech-2012, Thu-P9b-01, pg. 1-4, Portland, OR, Sept. 9-13, 2012.
[47.] S. Uluskan, J.H.L. Hansen, "Phoneme Class Based Adaptation for Distant Noisy Speech," ISCA Interspeech-2012, Wed-P7b-06, pg. 1-4, Portland, OR, Sept. 9-13, 2012.
[48.] A. Ziaei, A. Sangwan, J.H.L. Hansen, "Prof-Life-Log: Real life audio event matching using content based information," ISCA Interspeech-2012, Thu-O10d-04 pg. 1-4, Portland, OR, Sept. 9-13, 2012
[49.] G. Liu, C. Zhang, J.H.L. Hansen, "A Linguistic Data Acquisition Front-End for Language Recognition Evaluation," ISCA Odyssey-2012 (Speaker & Language Recognition Workshop), pp. 224-228, Singapore, June 25-28, 2012
[50.] T. Hasan, J.H.L. Hansen, "Factor Analysis of Acoustic Features using a Mixture of Probabilistic Principal Component Analyzers for robust Speaker Verification," ISCA Odyssey-2012 (Speaker & Language Recognition Workshop), pp. 243-247, Singapore, June 25-28, 2012

[51.] S.O. Sadjadi, J.H.L. Hansen, “Blind Reverberation Mitigation for Robust Speaker Identification,” IEEE ICASSP-2012, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 4225-4228, (paper #1767), Kyoto, Japan, March 25-30, 2012

[52.] G. Liu, J.-W. Suh, J.H.L. Hansen, “A Fast Speaker Verification with Universal Background Support Data Selection,” IEEE ICASSP-2012, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 4121-4124, (paper #3854), Kyoto, Japan, March 25-30, 2012

[53.] T. Hasan, H. Boril, A. Sangwan, J.H.L. Hansen, “A Multi-Modal Highlight Extraction scheme for Sports Videos using an Information-Theoretic Excitability Measure,” IEEE ICASSP-2012, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 2381-2384, (paper #3159), Kyoto, Japan, March 25-30, 2012

[54.] W. Kim, J.H.L. Hansen, “Feature Compensation employing Online GMM Adaptation for Speech Recognition in Unknown Severely Adverse Environments,” IEEE ICASSP-2012, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 4121-4124, (paper #2003), Kyoto, Japan, March 25-30, 2012

[55.] G. Liu, Y. Lei, J.H.L. Hansen, “Robust Feature Front-End for Speaker Identification,” IEEE ICASSP-2012, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 4233-4236, (paper #3247), Kyoto, Japan, March 25-30, 2012

[56.] A. Sangwan, A. Ziaei, J.H.L. Hansen, “ProfLifeLog: Environmental Analysis and Keyword Recognition for Naturalistic Daily Audio Streams,” IEEE ICASSP-2012, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 4941-4944, (paper #3725), Kyoto, Japan, March 25-30, 2012

[57.] S.O. Sadjadi, H. Boril, J.H.L. Hansen, “A Comparison of Front-End Compensation Strategies for Robust LVCSR under Room Reverberation and Increased Vocal Effort,” IEEE ICASSP-2012, IEEE Inter. Conf. on Acoustics, Speech and Signal Processing, pp. 4701-4704, (paper #3395), Kyoto, Japan, March 25-30, 2012

[58.] A. Singh, A. Sangwan, J.H.L. Hansen, "Improved Parcel Sorting by combining Automatic Speech and Character Recognition," IEEE ESPA-2012, IEEE Inter. Conf. on Emerging Signal Processing Applications, Las Vegas, Nevada, Jan. 12-14, 2012

[59.] T. Hasan, G. Liu, S.O. Sadjadi, N. Shokouhi, H. Boril, A. Misra, K.W. Godin, J.H.L. Hansen, “CRSS Systems for 2012 NIST Speaker Recognition Evaluation," NIST SRE-2012 Meeting, Orlando, FL, Dec. 11-12, 2012.

[60.] H. Boril, S.O. Sadjadi, J.H.L. Hansen, "A Study on Combined Effects of Reverberation and Increased Vocal Effort on ASR," LISTA-12: Listening Talker Workshop, Edinburgh, Scotland, May 2-3, 2012. http://listening-talker.org/workshop

8 LIST OF ACRONYMS:

- AFA — Acoustic Factor Analysis
- AWV — Average Weighted Variance
- ActDCF — Actual Detection Cost Function
- ASR — Automatic Speech Recognition
- BSW — Blind Spectral Weighting
- CMS — Cepstral Mean Subtraction
- CMVN — Cepstral Mean and Variance Normalization
- CTS — Conversational Telephone Speech
- DCT — Discrete Cosine Transform
- DCF — Detection Cost Function
- DET — Detection Error Trade-off
- DRR — Direct to Reverberant Ratio
- DID — Dialect Identification
- EER — Equal Error Rate
- EM — Expectation Maximization
- FA — Factor Analysis
- FAR — False Alarm/Accept Rate
- FRR — False Reject Rate
- GCDS — Gaussianized Cosine Distance Scoring
- GMM — Gaussian Mixture Model
- HMM — Hidden Markov Model
- HPS — Harmonic Product Spectrum
- JFA — Joint Factor Analysis
- k-NN — k-Nearest Neighbor
- KLD — Kullback-Leibler Divergence
- KWS — Keyword Spotting
- L2LR — L2-Regularized Logistic Regression
- LDA — Linear Discriminative Analysis
- LDC — Linguistic Data Consortium
- LID — Language Identification
- LLR — Log-Likelihood Ratio
- LPC — Linear Prediction Coefficients
- LPP — Locality Preserving Projections
- LR — Logistic Regression
- LRE — Language Recognition Evaluation
- LSA — Latent Semantic Analysis
- MAP — Maximum A Posteriori
- MFCC — Mel Frequency Cepstral coefficients
- MHEC — Mean Hilbert Envelope Coefficients
- MLP — Multi-Layer Perceptron
- MinDCF — Minimum Detection Cost Function
- MOLRT — Multiple-Observation Likelihood Ratio Test
- MVDR — Minimum Variance Distortion-less Response
- NAP — Nuisance Attribute Projection
- NIST — National Institute of Standards and Technology

- NLP — Natural Language Processing
- NMF — Nonnegative Matrix Factorization
- PCA — Principal Component Analysis
- PF — Phonological Features
- PLDA — Probabilistic Linear Discriminative Analysis
- PLP — Perceptual Linear Prediction
- PPRLM — Parallel Phone Recognition followed by Language Modeling
- PMVDR — Perceptual Minimum Variance Distortion-less Response Coefficients
- PNCC — Power Normalized Cepstral Coefficients
- RATS — Robust Automatic Transcription of Speech A
- RIR — Room Impulse Response
- ROC — Receiver Operating Characteristic
- SAD — Speech Activity Detection
- SDC — Shifted Delta Cepstrum
- SF — Spectral Flux
- SID — Speaker Identification
- SIR — Signal to Interference Ratio
- SNR — Signal to Noise Ratio
- SOLRT — Single-Observation Likelihood Ratio Test
- SPL — Sound Pressure Level
- SRE — Speaker Recognition Evaluation
- SVM — Support Vector Machine
- TF-IDF — Term Frequency–Inverse Document Frequency
- UBM — Universal Background Model
- UBS — Universal Background Support data selection
- UBSSVM — UBS based SVM
- VAD — Voice Activity Detection
- VOA — Voice of America
- VSF — Variance of Spectrum Flux
- WCCN — Within Class Covariance Normalization