



AFRL-RI-RS-TR-2020-075

ADVANCEMENTS IN ROBUST SPEECH PROCESSING

UNIVERSITY OF TEXAS AT DALLAS

APRIL 2020

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

**AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE**

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 nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (<http://www.dtic.mil>).

AFRL-RI-RS-TR-2020-075 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /

WAYNE N. BRAY
Work Unit Manager

/ S /

JAMES S. PERRETTA
Deputy Chief, Information
Exploitation & 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	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) APRIL 2020		2. REPORT TYPE FINAL TECHNICAL REPORT		3. DATES COVERED (From - To) AUG 2015 – MAY 2019	
4. TITLE AND SUBTITLE ADVANCEMENTS IN ROBUST SPEECH PROCESSING				5a. CONTRACT NUMBER N/A	
				5b. GRANT NUMBER FA8750-15-1-0205	
				5c. PROGRAM ELEMENT NUMBER 63260F / 34260F / 62788F	
6. AUTHOR(S) John H.L. Hansen and Wayne Bray				5d. PROJECT NUMBER G2AU	
				5e. TASK NUMBER WU	
				5f. WORK UNIT NUMBER TD	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Texas at Dallas – Center for Robust Speech Systems 800 W. Campbell Ave. Richardson, TX 75080-1407				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-2020-075	
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 The Advancements in Robust Speech Processing effort has investigated and developed novel approaches for multiple speech processing technologies to support work being conducted by the Air Force Research Laboratory. Specific areas of research addressed speaker ID with specific emphasis on speaker variability, Language and Dialect ID (LID-DID), speaker diarization, robust automatic speech recognition and Noise and Speaker Modeling.					
15. SUBJECT TERMS Speech Processing, Language Identification, Speaker Identification, Keyword Spotting, Automatic Speech Recognition, Speech Enhancement					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 79	19a. NAME OF RESPONSIBLE PERSON WAYNE BRAY
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (Include area code) N/A

Table of Contents

1	SUMMARY	1
2	INTRODUCTION	2
3	METHODS, ASSUMPTIONS, AND PROCEDURES	3
3.1	Speaker ID - Speaker Variability (SID-SV):	3
3.1.1.	Longitudinal Speaker variability.....	3
3.1.2.	Lombard Effect	4
3.2	Language ID (LID) / Dialect ID (DID):	4
3.3	Audio Diarization in Naturalistic Environments (DiarNAT)	5
3.3.1.	Environment Robustness for Speech Systems.....	5
3.3.2.	Robust ASR	5
3.3.3.	ASR Robustness in Noise/Reverberation/Distance Audio Capture.....	6
3.4	Noise & Speaker Modeling for Knowledge Assessment.....	6
4	RESULTS AND DISCUSSION	7
4.1	Speaker Identification	7
4.1.1.	Flexible Toolkit (MS-AcID)	7
4.1.2.	Curriculum Learning based Approaches for Robust Speaker Recognition.....	7
4.1.3.	Language Mismatch in SID	13
4.1.4.	Domain Mismatch Normalization.....	17
4.1.5.	Longitudinal Variability/MARP	19
4.1.6.	Lombard Effect & Vocal Effort.....	26
4.2	Open-Set Language ID (LID) / Dialect ID (DID)	29
4.3	Audio Diarization in Naturalistic Environments (DiarNAT)	35
4.4	Automatic Speech Recognition/Keyword Spotting.....	39
4.4.1.	Accented Automatic Speech Recognition	39
4.5	Noise & Speaker Modeling for Knowledge Assessment.....	45
4.5.1.	Speaker Recognition with Nonlinear Distortion: Clipping Analysis and Impact.....	45
4.5.2.	Automatic Sentiment Detection in Naturalistic Audio	50
5	CONCLUSION	59
6	References.....	60
7	ACRONYMS.....	67

LIST OF FIGURES

Figure 1 Overview of the Domain Adversarial Training (DAT) framework. Adversarial domain classifier has a gradient reversal layer. Speaker classifier and domain classifier both take input from the joint feature extractor, are optimized to excel in their own tasks.	14
Figure 2 Overview of the proposed Adversarial Discriminative Domain Adaptation (ADDA) approach. Source DNN encoder is fixed during the adversarial adaptation.	14
Figure 3 Overview of the i-Vector/PLDA system. Train data used for modeling UBM, total variability (TV) matrix, LDA (or SVDA) and PLDA, enrollment and test data used for evaluation of the system [23].	17
Figure 4 The number of target trials N_t for each of the 25 female speakers across absolute session difference (ASD) with an all-vs-all evaluation protocol.	21
Figure 5 Normalized offset parameters w_0 for individual ASDs and ASD ranges.	23
Figure 6 The UBNF feature extraction diagram. Sigmoid nonlinearity is used with softmax normalization for the output layer of DNN.	30
Figure 7 A conceptual semi-supervised learning framework with GANs. The “feature matching” trick is also employed to construct the generator loss, as proposed in [35].	31
Figure 8 Proposed SincNet-VTL approach for extracting speaker embedding from time-domain speech. (a) In Stage 1, SincNet is trained for frame-level speaker identification using out-domain data. (b) In Stage 2, we adopt the trained SincNet as feature extractor for in-domain data. We max() or avg() pooled frame-level features to get utterance-level embedding.	36
Figure 9 The architecture of waveform SincNet [40]. Sinc-Layer performs time-domain convolutions on raw speech. Next, two 1D convolutional layers and three fully connected layers filter the input. Final soft-max layer perform speaker classification.	36
Figure 10 PLTL data: Comparing i-vector with average pooled F1, F2 and F3 embeddings. w/o PCA means without PCA based dimension reduction.	38
Figure 11 AMI: 6 meetings. F2-avg with PCA (51 dim.) shows significant improvements over i-vector with PCA (51 dim.).	38
Figure 12 The proposed student-teacher learning to advance a multi-accent model: using accent-specific teachers (a), using multi-accent teachers (b).	41
Figure 13 Comparison of original clean and clipped speech that includes waveform, spectrogram, and averaged MFCC feature vectors for three phonemes (/ei/, /n/, /sh/).	46

Figure 14 DET curve results using i-Vector with PLDA approach on NIST SRE-08 corpus: performance for 0%, 0.5%, 1%, 5%, and 10% clipping distortion.....	49
Figure 15 Fig. 1. Baseline audio sentiment detection system uses Automatic Speech Recognition (ASR) to generate transcripts which are then processed by the sentiment classifier. Proposed sentiment detection system uses Keyword Spotting (KWS) to detect sentiment-bearing terms in the audio data. An iterative feature pruning technique is proposed to generate the sentiment-bearing terms for KWS. Offline development of customized text based sentiment models.	51
Figure 16 A plot of the variation in sentiment detection accuracy vs. simulated word error rate (WER). The trend suggests that only a small fraction of words can effectively determine sentiment, and most words are irrelevant from a detection perspective. This implies that KWS may be a feasible and superior solution for sentiment detection over ASR.	53
Figure 17 Accuracy distribution curve for proposed hybrid ME-KWS audio sentiment system. This gives information regarding the most optimal keyword list to be used.....	56

LIST OF TABLES

Table 1 EER (%) OBTAINED BY INCLUDING THE DIFFERENT CHANNELS IN CL-PLDA FOR THE 30S-30S DEV TRIAL LIST.....	9
Table 2 EER (%) OBTAINED BY VARYING THE NUMBER OF ITERATION (n) AFTER INCLUDING DATA FROM THE LAST CHANNEL (H) IN CL-PLDA USING THE 30S-30S DEV LIST	10
Table 3 EER, MISS @ 3% FA,MISS @ 4% FA,MISS @5 % FA, FA AT 3% MISS, AND FA AT 10% MISS FOR THE DARPA RATS SID TASK FOR ENROLLMENT-TEST SETS.....	11
Table 4 EER AND minDCF10 FOR OUR PROPOSED CL-PLDA AND ACL-PLDA BASED SID SYSTEMS COMPARED AGAINST THE TRADITIONAL MULTI-SESSION PLDA FOR THE NOISY VERSION OF NIST SRE 2010 C5 TRIALS	12
Table 5 Speaker verification results using different models with a PLDA Back-end.....	16
Table 6 EER/minDCF results comparing LDA and SVDA. The dimension of i-Vectors is reduced from 600 to 400.....	18
Table 7 Speaker recognition results comparing LDA and SVDA in terms of EER/minDCF without dimension reduction.....	19
Table 8 EER% for subsets of trials grouped according to absolute session difference (ASD). N_t and N_{nt} denote the number of target and non-target trials respectively.	22
Table 9 Proposed quality measure functions (QMFs) for incorporating aging information, i.e. ASD, as an additional parameter in conventional score calibration. λ was set at 0.1 in our experiments.....	23
Table 10 Performance of score-only calibration approaches: Full (F), Mismatched (MM), and Matched (M) and proposed score-aging calibration approaches Q1–Q3.....	24
Table 11 10-fold cross validation performance of score-only calibration, Full (F), and a score-aging calibration proposal, Q3. The mean values across the 10 folds are presented.....	25
Table 12 Cross-gender calibration, Full (F), and score-aging, Q1.	25
Table 13 EERs (%) for clean-Lombard trials of read speech. Noise levels L1, L2 and L3 represent 70, 80 and 90 dB-SPL for Crd and Hwy, and for Pnk, 65, 75, 85 dB-SPL. The ALL condition denotes the case where trials for all noise types and/or levels are pooled.....	27
Table 14 Cross-validation calibration of scores from clean-Lombard mismatched trials: performance metrics across Lombard noise types for different calibration approaches. The best performing metric(s) in each case are in bold.....	29

Table 15 Statistics of MGB-3 ADI Corpus. Utt stands for number of utterances, Dur. stands for duration in hours.	32
Table 16 Semi-supervised GANs architecture for i-vector features.	32
Table 17 Dev results of 3 baseline and 2 proposed systems.	33
Table 18 Eval results of 3 baseline and 2 proposed systems.	33
Table 19 Eval results of 3 baseline and 2 proposed systems–data augmented.	34
Table 20 Comparison with the recent results on MGB-3 ADI task.	35
Table 21 CRSS-PLTL data: Effect of PCA (51 components) on DER (%) for i-vector, F2-avg and F2-max features.	38
Table 22 Percentage of CSO between the multi-accent model (Teacher model) and different US-specific models.	42
Table 23 CER% of adapting a multi-accent trained model to Indian accent using tempered KL-divergence with three soft- outputs: Outputs of the multi-accent model (MA), outputs of accent-specific model with no teacher (No-Teacher) and an accent-specific student model with $\lambda=0.9$ and $T=4$ (Student). The baseline performance is shown in the first row.	43
Table 24 CER% of multi-accent without teacher model (MA NT), Multi-accent with a single teacher mode (MA ST), multi-accent with accent specific teacher (MA MT) and accent specific models on US English, Hispanic English (HIS) and Indian English (IND).	44
Table 25 Table 2: Comparison of four speech quality measures for original clean speech, and different levels (0.5%-10.0%) of waveform clipping on TIMIT data.	48
Table 26 EER and Minimum DCF results on NIST SRE-08 corpus: performance for clean trained speaker models and clean, 0.5%, 1%, 5%, 10% clipped test data.	50
Table 27 results for the proposed hybrid ME-KWS and ASR.	57
Table 28 Performance across various parameters in ut-opinion database	58

1 SUMMARY

This three-year effort has focused on addressing five research tasks:

- (i) Speaker ID with specific emphasis on speaker variability (SID-SV)
- (ii) Language and Dialect ID (LID-DID) to improve LID/DID with low or non-uniform resources including Open-Set LID/DID language rejection
- (iii) Diarization in Naturalistic Environments - algorithm formulation for daily audio streams which could contain speaker-overlap, noise/distortion, long periods of non-speech or silence. Automatic speech recognition/keyword spotting in noise employing deep learning concepts with articulatory/phonological speech modeling
- (iv) Noise and Speaker Modeling for knowledge extraction/assessment.

All these tasks share a common theme of developing novel and robust approaches to automated speech processing technology.

2 INTRODUCTION

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

- i. **Speaker ID - Speaker Variability (SID-SV):** to address SID based on speaker variability. This effort will include: (i) longitudinal time/session variability, (ii) vocal effort (iii) speech in noise due to Lombard effect.
- ii. **Language ID / Dialect ID (LID-DID):** to improve LID/DID with low or non-uniform resources including Open-Set LID/DID language rejection. Robustness to mismatch in both acoustics as well as speaking structure. Advancements will concentrate on novel design of effective fusion of spectral and prosodic front-ends, discriminative dimensionality reduction, score normalization, improved data purification for acoustic model training/testing.
- iii. **Audio Diarization in Naturalistic Environments (DiarNAT):** the ability to perform automatic assessment of continuous “free running” input speech for speaker identity, conversational structure, keywords or topic content, all represents challenges in both broadcast data as well as naturalistic daily speech. Most diarization focuses on broadcast content, while the effort here will focus on advancing the state of the art in algorithm formulation for daily audio streams which could contain speaker-overlap, noise/distortion, long periods of non-speech or silence, or multiple individuals working together to solve collaborative problems.
- iv. **Speech Recognition/Keyword Spotting:** deep learning techniques including convolutional networks show great progress on controlled audio tasks. The focus here will be on development of speech recognition using articulatory and phonological based strategies, including bottleneck features as an alternative to traditional speech spectrogram strategies.
- v. **Noise & Speaker Modeling for Knowledge Assessment (N+S-MOD-KA):** noise and environmental distortion causes significant loss in speech/language system performance. When audio data is “damaged”, it often becomes difficult to employ it for other domains. In this task, the focus is to develop algorithms to quantify noise/distortion in audio streams, as well as assessment of speaker structure for knowledge extraction.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Speaker ID - Speaker Variability (SID-SV):

The focus in this area is to address SID for speaker variability – which include longitudinal variation, session variation and vocal effort variation including flavors of Lombard effect (LE). The proposed research in this domain will concentrate on addressing speaker variability as it relates to impacting speaker recognition/verification system performance.

3.1.1. Longitudinal Speaker variability

Longitudinal speaker variability will be considered across four different time-scales: (1) Within-session; (2) between-session, short-term: separated by several days-weeks; (3) between-session, medium term: separated by several months; (4) between-session, long-term: separated by several years. Dominant sources of variability expected to be present at each time-scale include: (a) Conversational and communication factors; (b) Everyday changes in the voice, due to speaker emotional state and physical health; (c) Combination of vocal aging and everyday variability, (d) vocal aging in general. There are clearly challenges to studying long-term variability in speech due to access of relevant speech data.

Limited research has been conducted in the area of longitudinal speaker analysis. However, it is proposed to explore the tracking of longitudinal changes in the voice at the: Low level: spectral envelope, conventional short-term features (Mel-filtered Cepstral Coefficients- MFCC, Perceptual Linear Prediction - PLP), alternative features (Perceptually Motivated Minimum Variance Distortionless Response - PMVDR [1], Teager Energy Operator - TEO [2], phase-based); Medium level: trajectory of vocal measures, including average F0, F1-F3, local jitter and shimmer, speaking rate; High level: prosodic information, F0 contours, idiolect, word count [3]. Given this background, we will use this information to formulate new aging-invariant SID features, and to provide warning flags to SID system when significant aging variability is present. An ultimate goal for speaker aging analysis would be to develop a life-cycle model of the voice, from childhood, through adolescence and adulthood, and into old age.

We will consider advancements based on evaluating longitudinal corpora with several state-of-the-art in-house SID systems, based around an i-vector PLDA framework [4] [5]. We will assess the impact on speaker models (GMM supervectors and i-vectors), output scores, error metrics (Equal Error Rate - EER, Decision Cost Function - minDCF, log likelihood ratio calibration cost - C_{lir}), and in the forensic case, “strength of evidence” likelihood ratios (LRs). We will first develop aging invariant SID solutions at the score level, via calibration [6] or side-information approaches [7] [8]. The “calibration of SID output scores is a necessary procedure for setting reliable decision thresholds, and for evaluating application-independent performance [9]. Scores are typically calibrated by applying a linear transformation learned on a set of development scores. This transformation can be extended to include additional “quality” information to improve discrimination performance (e.g. EER or minDCF). Quality, in this instance, is any information about the sample, which has a relationship with the system output score, e.g. SNR. In SID, duration information has been successfully incorporated in calibration as an additional parameter [6] [10]. We propose to apply this idea to ageing in SID. Relationships between SID score and quality information including speaker age, gender, and the time-lapse between train and test samples will

be used to develop linear and non-linear score transformations that result in improved discrimination performance.

3.1.2. Lombard Effect

The phenomenon whereby speakers alter their speech in noisy environments, in order to maintain intelligible communication, is referred to as the Lombard Effect (LE). In severe noise, the change in speech production due to LE severely degrades the performance of speech recognition systems using neutral speech as a reference [11]. The changes in speech characteristics under LE have been shown to be dependent on the noise type (e.g. white, pink, car), and the noise level [12], resulting in different “flavors” of LE. An initial study on SID under LE [12] suggests that the acoustic mismatch between SID models and LE speech can be reduced by adapting the neutral speaker models towards speaker-specific LE speech. We aim to extend this approach by developing a general transformation of neutral to LE speech, removing the need for speaker-specific LE speech. We will approach this challenge via acoustic subspace compensation: speaker models will be trained for each sample in a database of LE speech (such UT-SCOPE [13]). Applying PCA (principal components analysis) to the set of mean-normalized models will reveal the main “direction” of LE-related variability in the acoustic space. Knowledge of this subspace can then be applied to unseen data to “project away” LE variability in an SVM or GMM-UBM framework, via NAP (nuisance attribute projection) and Eigenchannel approaches respectively [14]. It will be established if this approach can be generalized to unseen LE flavors. We will subsequently investigate approaches to incorporate this compensation in an i-vector PLDA framework. In addition, we will investigate if front-end techniques previously developed for ASR under LE [11], including blind cepstral normalizations, maximum likelihood spectral transformations, and temporal filtering, can be applied to SID.

3.2 Language ID (LID) / Dialect ID (DID):

The task of language identification (LID) or dialect identification (DID) is to identify a spoken language or dialect captured in an audio stream. Automated systems provide potential benefits over human-based LID/DID as they can be trained for an extensive number of languages or dialects in a short period of time and be deployed in a distributed fashion to process multiple channels simultaneously, that all at considerably lower costs compared to human experts. Automated LID/DID systems can benefit numerous applications in the area of spoken language processing. They can directly aid humans in retrieving a spoken language in a multi-language/dialect settings. Automated LID/DID may provide a valuable input in security, defense, and forensic applications, as well as in speech-to-speech translation technologies. Their input may be crucial for emergency calls or field operation communications as they can help route the interlingual content to online expert interpreters. Another broad area of application is in automated customer service centers, information retrieval systems, and in diarization of live broadcasting or indexing of large offline multimedia archives. Languages and dialects can be distinguished by their unique acoustic-phonetic inventories, rules governing the possible sequences of phones to form words (phonotactics), prosodic patterns, and language/dialect-specific vocabulary and grammar. Currently, mostly acoustic and phonotactic models have been popular for automated LID/DID. Phonotactic approaches usually utilize phone recognizers and statistical modeling of output phone streams to distinguish unique components of each language.

The focus in this area is to develop new advancements to address both language and dialect identification. These will include novel features, mismatch compensation schemes, and training/classification strategies for language identification. These methods will consider existing audio data provided by USAF, DARPA RATS – LID corpus, and data available from NIST LRE. Our main focus will be on Open-Set LID/DID where the task is to assign an incoming sample to a particular language from a specific group of target languages (inset languages) or reject it as an unknown - an out-of-set language.

3.3 Audio Diarization in Naturalistic Environments (DiarNAT)

The focus in this research task is to formulate effective algorithms for diarization (i.e., who said what, when, where, how, and to whom?) for extensive long naturalistic audio streams. Research in environment tracking has focused on segmentation, clustering and classification. Most studies focus on classification and tend to assume that segmentation is available. This assumption tends to fail for practical naturalistic audio recordings. Another assumption frequently seen in environment classification studies is that pure environment samples are available for training and classification (i.e., audio samples that are not corrupted by other noise sources or speech). In practice, corruption of environment samples with speech is very common. Finally, it is also noted that certain temporal constraints are inherent to naturalistic recordings. For example, abrupt decision changes in environment classification are highly unlikely, i.e., it is nearly impossible to move from office to car environment and then back to office within minutes. Most existing noise/environment corpora contain short noise samples drawn randomly from time and do not provide time-continuity which is necessary for researchers to study this phenomenon or exploit it in solution design. The mentioned shortcomings make it harder to obtain real advancements from environment estimation.

3.3.1. Environment Robustness for Speech Systems

We have been collecting large volumes of naturalistic data (such as Prof-Life-Log, NASA Apollo Space Missions, and Peer Led Team Learning). We propose to explore solutions which exploit these large data resources. The biggest challenge in building solutions for compensation is to be able to handle previously unseen environments. In order to address this issue, we propose to build a universal background model (UBM) for environment (borrowing from a popular concept in speaker identification). UBM based approaches such as supervector or i-Vector would help us to better characterize environments. In a parallel manner, noise from our collection can be used to corrupt speech data and produce noisy features (and thereby provide a parallel corpus of clean and noisy features). Now, a compensation system can be trained using noise characterization vector and noisy speech vectors as input and clean speech vectors as output. The UBM can be upgraded as we continue to collect more data (and this process is bound to make the whole system more comprehensive).

3.3.2. Robust ASR

A study of the current literature on deep neural networks suggests that the neural networks are adept at learning complex non-linear relationships. The relationship between noisy and clean speech, stressed and neutral speech, accented and neutral speech, etc. are highly complex and non-

linear. It is suggested that for at least the proportion of speech production that represents a consistent speech production shift, that these relationships could be learned using deep learning techniques. We propose to explore such mismatch using deep learning based modeling. We will explore the following areas in the context of DNN: (i) DNN adaptation for environmental mismatch; (ii) alternative network architectures; (iii) multichannel data in DNN. Our study in the area of DNN adaptation for environmental mismatch will focus on developing effective discriminative adaptation strategies that will require only a small number of adaptation utterances from the target environment. In the second domain, the use of alternative network architectures such as convolutional neural networks (CNN) and recurrent neural networks (RNN) will be explored. RNNs in particular are attractive choices for reverberant ASR since they explicitly model the inter-frame correlations. In the third domain, effective use of multichannel data for a DNN-based ASR system will be considered. 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.

3.3.3. ASR Robustness in Noise/Reverberation/Distance Audio Capture

This task will focus on advancements in speech recognition using phonological speech modeling for advancements in KWS. In addition, further advancements in robust speech recognition to address noise, reverberation, vocal effort and other sources of variability will be explored. Recently, deep learning techniques such as deep neural networks and convolutional neural networks have allowed researchers to make rapid advancements in speech recognition [Hinton12]. However, traditional robustness problems related to production (accent/dialect), environment (noise/channel), and mismatch in general, continue to negatively impact speech recognition performance. At some level, there are some problems where simply using more training data with traditional machine learning strategies cannot solve the core scientific questions. However, given the benefits of deep learning (i.e., better performance but in structured data rich environments), versus limitations (i.e., extensive training data needed as well as extensive computing requirements), it would be appropriate to explore how deep learning methods could offer new tools to address these traditional mismatch problems. At CRSS, we have been investigating speech recognition and keyword spotting for naturalistic corpora such as Prof-Life-Log (PLL), Apollo Space Missions, UT-Opinion, and Peer Led Team Learning (PLTL). The tasks are very challenging because the data contains a variety of real-world conditions such as dynamic noise, variable signal-to-noise ratio, various topics, significant out of vocabulary words, distant speech capture, changing levels of vocal effort, etc.

3.4 Noise & Speaker Modeling for Knowledge Assessment

This last task will focus on the development of algorithms to assess new knowledge regarding noise/distortion and speaker modeling knowledge. This research will result in new algorithms to estimate various speaker traits, as well as special conditions related to the noise/environment. We will continue to develop superior techniques for height estimation and also explore new physical traits such as weight, age, health, gender, etc. Additionally, we will expand the research to include new traits such as (i) sentiment/emotion/stress, (ii) social signal detection (laughter, fillers, etc.), and (iii) communication metrics (such as word count). Extracting knowledge of this nature can help improve speech, speaker, and language recognition systems.

4 RESULTS AND DISCUSSION

4.1 Speaker Identification

This area of research has focused on robust automated identification of speakers based on speech and has focused on application of curriculum learning and a variety of mismatch conditions that can significantly degrade automated system performance.

4.1.1. Flexible Toolkit (MS-AcID)

From previous funding, CRSS-UTDallas produced a *Flexible Speaker Recognition Toolkit* entitled “MS- AcID”, which supported multi-session (MS) data, as well as five different back-end acoustic ID solutions which were fused for an overall state-of-the-art speaker recognition system. Officially, CRSS-UTDallas team received the IEEE ICASSP-2013 best student paper award (financially supported by IBM) for their paper on robust speaker recognition, stemming from their contributions to the NIST SRE-2012 competition.

Unofficially, in the NIST i-Vector Speaker ID challenge, CRSS-UTDallas finished second amongst more than 60 teams worldwide (a company in Russia placed first). This platform was delivered to AFRL, and will serve as the basis for continuous tool improvement.

4.1.2. Curriculum Learning based Approaches for Robust Speaker Recognition

Performance of speaker identification (SID) systems is known to degrade rapidly in the presence of mismatch such as noise and channel degradations. This study introduces a novel class of curriculum learning (CL) based algorithms for noise robust speaker recognition. We introduce CL-based approaches at two stages within a state-of-the-art speaker verification system: at the i-Vector extractor estimation and at the probabilistic linear discriminant (PLDA) back-end. Our proposed CL-based approaches operate by categorizing the available training data into progressively more challenging subsets using a suitable difficulty criterion. Next, the corresponding training algorithms are initialized with a subset that is closest to a clean noise-free set, and progressively moving to subsets that are more challenging for training as the algorithms progress. We evaluate the performance of our proposed approaches on the noisy and severely degraded data from the DARPA RATS SID task, and show consistent and significant improvement across multiple test sets over a baseline SID framework with a standard i-Vector extractor and multisession PLDA-based back-end. We also construct a very challenging evaluation set by adding noise to the NIST SRE 2010 C5 extended condition trials, where our proposed CL-based PLDA is shown to offer significant improvements over a traditional PLDA based back-end.

In this study, we propose a novel model based approach to improve SID performance in the presence of severe noise and channel distortions that is inspired by the CL based Machine Learning algorithms. CL based algorithms operate by categorizing the training data into subsets based on a suitable *difficulty* criterion. We introduce CL based training at two distinct stages of a state-of-the-art i-Vector PLDA based SID system: at the PLDA back-end level, and at the i-Vector extractor matrix estimation level. First, we present the CL based PLDA (CL-PLDA) algorithm that operates by initializing the PLDA training EM algorithm with easy examples, and gradually adding difficult examples as the algorithm progresses. Compared to CL-PLDA, the traditional multi-condition

PLDA operates by training on all the available data in each iteration of the training EM algorithm. We measure the performance of CL-PLDA on noisy and severely distorted utterances of the DARPA RATS SID data where we use the ratio of the intended signal to non-signal components of the radio channels as the CL difficulty criterion [15]. We show that CL-PLDA consistently outperforms a multi-condition PLDA for SID in noisy conditions with enrollment/test sets of varying duration. We hypothesize that CL-PLDA is able to find an overall better local minima during the training that results in improved estimation of parameters than the traditional PLDA, resulting in better overall SID performance in noisy and degraded conditions.

We note that we had proposed a preliminary version of CL-PLDA in [16], which we develop further significantly in this study by:

- 1) Training separate duration-matched CL-PLDA back-ends to mitigate the performance degradation when testing on enrollment-test sets of very short duration for the DARPA RATS SID task.
- 2) Present a detailed treatment for implementing CL-PLDA with all the necessary equations.
- 3) Compare the SID performances in terms of several DARPA metrics in addition to EER and Detection Error Trade-off (DET) plots.
- 4) Develop a new version of CL based PLDA that gives comparable performance to CL-PLDA when SNR is used as the CL difficulty criterion. This was performed for the very challenging SID trial list we created by adding noise from the PRISM set to the NIST SRE 2010 C5 condition trials [17] [18].

This new version of CL-PLDA called Anti-CL-PLDA (ACLPLDA) differs from the original version in the manner in which the training data is used, in the sense that the training is initiated with the most difficult examples before including comparatively easier examples in training.

Here, we introduce two CL based PLDA training approaches: CL-PLDA and ACL-PLDA. The key step in both approaches is to divide the training data into subsets based on a difficulty criterion, and introduce more data sequentially to the training algorithm in order of increasing difficulty for CL-PLDA; and decreasing difficulty for ACL-PLDA. Our CL based techniques involve training on all the available data, and the learning algorithm is allowed access to the training data arranged sequentially according to some difficulty criterion that results in better estimation of the model parameters. The key necessity of any CL based learning algorithm is a metric for assigning the difficulty of the training examples, which is introduced next.

We hypothesize that for a labeled factor analysis scheme such as PLDA, the SNR values can indeed be a suitable difficulty criterion, since easy examples with high SNRs can be more informative for estimation of the latent variables. For experiments on the noisy version of the NIST SRE 2010 C5 condition trials reported in this study, we have used SNR values as the

difficulty assignment criterion. To elaborate further, an utterance with a high SNR value is deemed to be an easier example compared to an utterance with low SNR. For the DARPA RATS SID task data used in this study, a clean source utterance was retransmitted through eight radio channels with different transmitter and receiver equipment. The retransmitted utterances are severely noisy and highly degraded with shifts in pitch, carrier modulation/demodulation, non-linear intermittent fading, long time scale amplitude variation, ring modulation among others [15].

For the CL-PLDA based SID system, the entire dataset for training the back-ends was arranged channel-wise according to the SND-ND values from [20, Table I]. Specifically, the order in which the data was included in CL-PLDA was SRC (source) utterances, followed by the utterances from channels B, E, A, F, C, and D, G and H respectively. We used the 30s-30s dev trial list to determine the optimal number of iterations per channel for our proposed CL-PLDA based approach. We have observed that a reasonable choice for iterations per channel is to keep it close to the number of iterations used to train the baseline multi-session PLDA system, since this allows for adequate number of iterations per new channel for the CL-PLDA training. We also investigated the effect of including training data from the various channels to our CL-PLDA approach. To this end, Table 1. shows the EER obtained after including data from the clean source (SRC) and the different radio channels in CL-PLDA. Five iterations are used after including the data of each new channel in CL-PLDA training.

Table 1 EER (%) OBTAINED BY INCLUDING THE DIFFERENT CHANNELS IN CL-PLDA FOR THE 30S-30S DEV TRIAL LIST

Channels included in CL-PLDA	EER (%)
SRC	38.99
+B	29.25
+E	16.82
+A	14.14
+F	13.75
+C	13.67
+D	13.28
+G	13.28
+H	13.31

Table 2 EER (%) OBTAINED BY VARYING THE NUMBER OF ITERATION (n) AFTER INCLUDING DATA FROM THE LAST CHANNEL (H) IN CL-PLDA USING THE 30S-30S DEV LIST

No. of iterations for last channel (H)	1	2	3	4	5
EER	13.28	13.35	13.36	13.31	13.31

It can be observed that in general, the SID performance improves after including each new channel in CL-PLDA training with the important exception of the last channel (H). After including data from channel H, and using five iterations, a slight degradation in EER can be observed. We therefore carried additional tuning after including data from the last channel. Table 2 shows the effect of varying the number of iterations after including data from the last channel (H) for CLPLDA training using the 30s-30s dev trials. As can be observed, using values of n higher than 1 leads to a slight increase in EER. Therefore, after adding the data from the last channel (H), only a single iteration of the EM training algorithm is used to obtain the final CL-PLDA model parameters. It can also be observed that using only a single iteration after including data from the last channel does not offer any additional improvements compared to the CL-PLDA model obtained using the previous channel (G). However, we chose to include data from the last channel in our final CL-PLDA model since it could possibly help, and was not causing any degradation if only a single iteration was used. A possible reason for only a single iteration needed after adding the data from the last channel could be that CL-PLDA has already trained gradually on most of the training data, and so only a single update after adding more noisier and degraded data at the end is sufficient. The same set of i-Vectors as used for the baseline PLDA based SID systems were used for the CL-PLDA based SID back-ends. No optimization was performed for the other duration trials, and the same configuration as determined by the 30s-30s dev trial list was used in all experiments for the DARPA RATS SID task.

For training the i-Vector extractor matrix using our proposed CL-IE approach, the data for training the T matrix was divided into subsets in the exact same manner as for the CL-PLDA back-end according to the SND-ND values of the channels. To train the CL-IE based i-Vector extractor, 2 iterations ($n = 2$, from Algorithm 3) were carried out with every new channel's addition for all the channels, except the last channel. After adding the data from the last channel, four iterations of the EM algorithm were used to obtain the final i-Vector extractor T matrix. For the CL-IE based SID systems, traditional multi-session PLDA back-ends were used. We also perform SID experiments with the combined CL-IE and CL-PLDA based approaches. For these CL-IE + CL-PLDA experiments, i-Vectors were first extracted using the CL-IE based method. Next, the CL-PLDA approach was used for training the SID back-ends. The CL-IE + CL-PLDA based approach, therefore, introduces CL at two stages of the i-Vector PLDA based state-of-the-art SID framework.

Table 3 EER, MISS @ 3% FA, MISS @ 4% FA, MISS @ 5% FA, FA AT 3% MISS, AND FA AT 10% MISS FOR THE DARPA RATS SID TASK FOR ENROLLMENT-TEST SETS

Evaluation Condition	SID Approach	EER (%)	Miss @3% FA	Miss @4% FA	Miss @5% FA	FA @3% Miss	FA @10% Miss
120s-120s	PLDA (5 iterations)	7.88	17.71	14.23	11.78	20.25	5.97
	PLDA (100 iterations)	7.32	15.45	11.99	9.98	17.34	4.99
	CL-PLDA ($n = 5$)	6.66	14.10	10.90	9.07	13.46	4.37
	CL-IE (PLDA: 5 iterations)	7.45	16.33	12.98	10.61	18.97	5.39
	CL-IE (PLDA: 100 iterations)	6.89	14.20	10.78	9.06	16.67	4.41
	CL-IE + CL-PLDA ($n = 5$)	6.30	12.96	10.06	8.03	12.99	4.02
30s-30s	PLDA (5 iterations)	11.34	28.25	23.90	20.65	35.36	13.39
	PLDA (100 iterations)	10.41	24.71	20.90	17.90	31.69	10.94
	CL-PLDA ($n = 5$)	9.64	22.53	18.80	16.08	27.31	9.33
	CL-IE (PLDA: 5 iterations)	10.89	26.47	22.07	19.58	34.98	12.24
	CL-IE (PLDA: 100 iterations)	10.14	23.17	19.56	17.09	31.51	10.38
	CL-IE + CL-PLDA ($n = 5$)	9.33	21.67	17.73	15.09	27.01	8.61
10s-10s	PLDA (5 iterations)	17.82	51.42	45.97	41.12	56.52	30.76
	PLDA (100 iterations)	17.08	46.38	41.38	36.95	56.17	29.16
	CL-PLDA ($n = 5$)	16.55	45.77	40.43	36.17	54.85	27.93
	CL-IE (PLDA: 5 iterations)	17.57	49.80	44.67	39.98	58.22	29.64
	CL-IE (PLDA: 100 iterations)	16.80	45.07	40.08	36.34	56.74	27.86
	CL-IE + CL-PLDA ($n = 5$)	16.46	44.31	39.36	35.38	56.13	26.63
3s-3s	PLDA (5 iterations)	26.78	79.55	75.16	71.05	68.52	48.32
	PLDA (100 iterations)	26.00	77.11	72.34	68.41	69.86	48.58
	CL-PLDA ($n = 5$)	25.80	77.07	72.68	68.66	67.50	46.66
	CL-IE (PLDA: 5 iterations)	26.88	79.76	75.57	71.76	69.52	49.31
	CL-IE (PLDA: 100 iterations)	26.67	77.46	72.67	68.81	71.89	50.08
	CL-IE + CL-PLDA ($n = 5$)	26.28	77.48	72.71	68.78	69.30	48.45

For both CL-PLDA and CL-IE + PLDA, 5 iterations ($n = 5$) per new channel (except the last channel) were used.

Table 3 shows the performance of our proposed CL-PLDA, CL-IE, and CL-IE + CL-PLDA based SID approaches compared against a traditional multi-session PLDA for the 120s-120s, 30s-30s, 10s-10s and 3s-3s enrollment test-sets. Table 3 also shows Miss @ 3% FA, Miss @ 4% FA, Miss @ 5% FA, FA at 3% Miss, and FA at 10% Miss. Since we found the traditional multi-session PLDA to be sensitive to the number of iterations, we have included two sets of baseline systems trained with 5 and 100 iterations respectively. To do a similar comparison for the traditional PLDA back-end trained with the i-Vectors extracted using the CL-IE algorithm, we have also included results with 5 and 100 iterations for the CL-IE based systems. Both CL-PLDA and CL-IE + CL-PLDA based approaches were trained with five iterations ($n = 5$ from Algorithm 1) for all channels except the last channel as mentioned previously.

Based on the results from Table 3, our proposed CL-based SID approaches outperform traditional multi-session PLDA in terms of the additional DARPA RATS performance metrics significantly. For the 120s-120s evaluation condition, our CL-IE + CL-PLDA based approach reduced the Miss @ 3% FA, Miss @ 4% FA, Miss @ 5% FA, FA at 3% Miss, and FA at 10% Miss by 26.82%, 29.30%, 31.83%, 35.85% and 32.66% (all values are relative) respectively compared to the multi-session PLDA based SID system trained with 5 iterations. Very significant reduction in these metrics can also be observed for 30s-30s and 10s-10s trials. Similar to what was observed for EER, CLPLDA based approach also outperforms CL-IE+CL-PLDA for the 3s-3s condition.

For ACL-PLDA experiments, we use the same three groups of i-Vectors for utterances with SNRs 8 dB, 15 dB and 20 dB, but the training proceeded in a reverse fashion starting with the 8 dB i-Vectors, followed by 15 dB and 20 dB i-Vectors respectively. For ACL-PLDA system trained with augmented training set that included clean data as well, the training involved an extra data subset corresponding to the clean i-Vectors at the end. In [19], to extract the i-Vectors for the noisy sets, the SAD decisions from the clean version of the speech utterances were used. Since we did not want SAD errors from noisy data to color the performance comparison for all systems, we also performed a separate set of experiments where we used the clean SAD decisions for extracting the corresponding noisy i-Vectors. This can be viewed as making use of a very efficient SAD approach that operates reasonably well on noisy data. We report the performance of the baseline multi-session PLDA, the CL-PLDA, and the ACL-PLDA SID systems for these *clean SAD* experiments. All three SID systems were trained with the augmented training set that also include i-Vectors for the clean data in addition to the i-Vectors for the noisy data.

Table 4 EER AND minDCF10 FOR OUR PROPOSED CL-PLDA AND ACL-PLDA BASED SID SYSTEMS COMPARED AGAINST THE TRADITIONAL MULTI-SESSION PLDA FOR THE NOISY VERSION OF NIST SRE 2010 C5 TRIALS

SID System	Back-end Trained On	Tested on	EER (%)	minDCF10
PLDA	Clean	Clean	2.41	0.048
PLDA	Clean	Noisy	9.02	0.087
PLDA	Noisy	Clean	6.75	0.081
PLDA	Noisy + Clean	Clean	4.50	0.068
PLDA	Noisy	Noisy	6.40	0.082
CL-PLDA	Noisy	Noisy	6.30	0.078
ACL-PLDA	Noisy	Noisy	6.26	0.079
PLDA	Noisy + Clean	Noisy	6.06	0.079
CL-PLDA	Noisy + Clean	Noisy	5.59	0.074
ACL-PLDA	Noisy + Clean	Noisy	5.54	0.075

Two sets of experiments are reported: SID back-ends trained with only noisy data; SID back-ends trained with the augmented training set with noisy + clean data. Results for the clean version of the trials, and the noisy version of the trials trained using the PLDA trained on only clean utterances are also shown.

Table 4 shows results of the baseline multi-session PLDA system compared against our proposed CL-PLDA, and ACLPLDA based SID approaches for the noisy version of NIST SRE 2010 C5 extended condition trials. We also show results for the original NIST SRE2010 C5 clean trials using a PLDA back-end, and results for noisy trials using the clean data trained back-end. For noisy trials, we have shown two sets of results corresponding to the original noisy training set, and the augmented training set that also includes i-Vectors of the clean utterances for all the three types of SID approaches. We show the results in terms of EER and minDCF10 for all the investigated approaches. For the SID system trained on clean data only, a very significant degradation in EER can be observed when switching from clean to noisy trials marked by an increase in EER from 2.41(%) to 9.02(%). As can be inferred from the results, both our CL based PLDA approaches outperform traditional multi-session PLDA backend. It can also be observed that including clean

data for training the ACL-PLDA and CL-PLDA based approaches results in significant reductions in EER and minDCF10.

4.1.3. Language Mismatch in SID

Speaker verification systems often degrade significantly when there is a language mismatch between training and testing data. Being able to improve cross-lingual speaker verification system using unlabeled data can greatly increase the robustness of the system and reduce human labeling costs. In this study, we introduce an unsupervised Adversarial Discriminative Domain Adaptation (ADDA) method to effectively learn an asymmetric mapping that adapts the target domain encoder to the source domain, where the target domain and source domain are speech data from different languages. ADDA, together with a popular Domain Adversarial Training (DAT) approach, are evaluated on a cross-lingual speaker verification task: the training data is in English from NIST SRE04-08, Mixer 6 and Switchboard, and the test data is in Chinese from AISHELL-I.

We use a recently proposed successful speaker model called X-vector [20], to extract speaker representations, and a Probabilistic Linear Discriminant Analysis (PLDA) back-end to compare pairs of enrollment and test speaker embeddings. The X-vector model is based on a Time-Delay Deep Neural Network (TDNN) architecture that computes speaker embeddings from variable-length acoustic segments. The network consists of layers that operate on speech frames, a statistics pooling layer that aggregates over the frame-level representations, additional layers that operate at the segment-level, and finally a softmax output layer. The embeddings are extracted after the statistics pooling layers.

In order to address the cross-lingual speaker verification problem, we first implement a Domain Adversarial Neural Network (DANN) [21] using DAT [22] to transfer speaker information from labeled English data to another language where only unlabeled data exists, for example, Chinese. DANN in Figure 1 is a Y-shaped network with two discriminative branches: a speaker recognizer and an adversarial language classifier. Both branches take input from a shared feature extractor that aims to learn hidden representations that capture the underlying information of the speaker and are independent of languages. We can implement the language independent speaker verification system assuming that DANN can learn features that perform well on speaker classification for the source and target language data, are independent with respect to the shift in language. This can be done by minimizing the speaker classification loss and maximizing the domain classification loss with a gradient reversal layer.

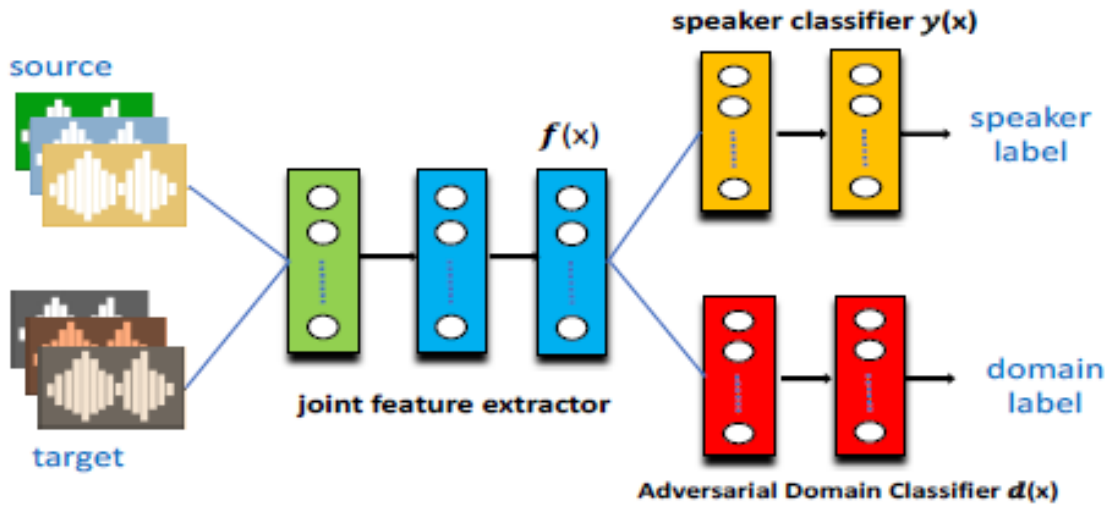


Figure 1 Overview of the Domain Adversarial Training (DAT) framework. Adversarial domain classifier has a gradient reversal layer. Speaker classifier and domain classifier both take input from the joint feature extractor, are optimized to excel in their own tasks.

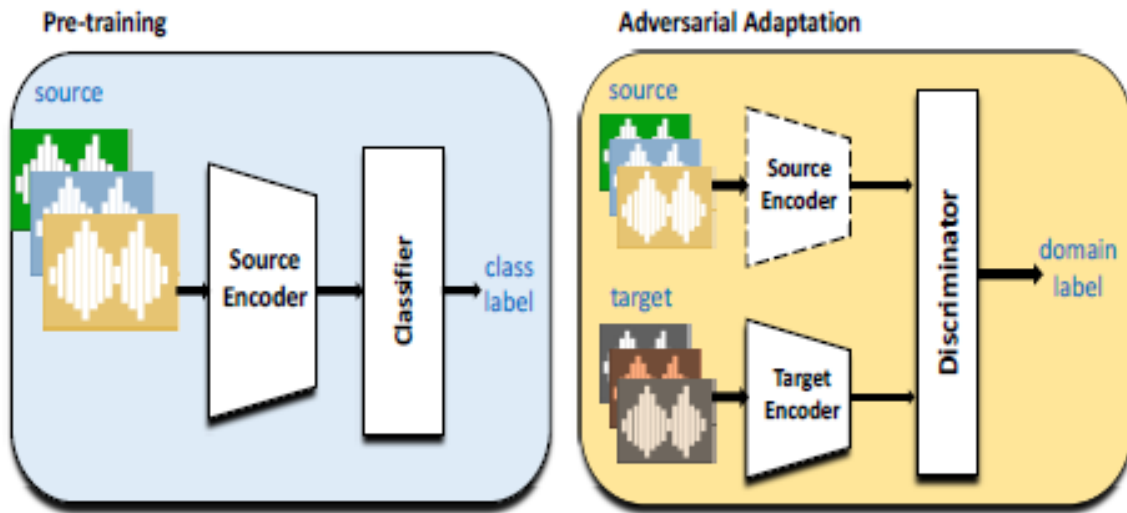


Figure 2 Overview of the proposed Adversarial Discriminative Domain Adaptation (ADDA) approach. Source DNN encoder is fixed during the adversarial adaptation.

DANN mainly has two components: 1) a speaker recognizer y for the source data; 2) an adversarial language classifier d that predicts a scalar indicating whether the input speech is from the source language or the target language. The two classifiers take input from the shared feature extractor f , which operates on the average of the speaker embeddings. The loss function of DANN is a multi-task loss that combines the loss of the speaker classifier and the domain classifier with a weight λ . Training DANN consists in optimizing,

$$\begin{aligned} \mathbb{E}(\theta_f, \theta_y, \theta_d) = & \frac{1}{n} \sum_{i=1}^n \mathcal{L}_y^i(\theta_f, \theta_y) - \lambda \left[\frac{1}{n} \sum_{i=1}^n \mathcal{L}_d^i(\theta_f, \theta_d) \right. \\ & \left. + \frac{1}{n'} \sum_{i=n+1}^N \mathcal{L}_d^i(\theta_f, \theta_d) \right], \end{aligned} \quad (1)$$

where $\theta_f, \theta_y, \theta_d$ are parameters of the joint feature extractor and two classifiers, and $\mathcal{L}_y, \mathcal{L}_d$ are the prediction and the domain loss functions. n and n' are the number of samples of the source and target domain data respectively. We can optimize this loss function using stochastic gradient descent to get the parameters. Using this DAT approach, we are able to minimize the divergence between the source and target feature distributions. Therefore, the learned embeddings are less dependent on the shift in language.

Different from the DAT method that applies a gradient reversal layer to confuse the domain classifier, we apply the Adversarial Discriminative Domain Adaptation (ADDA) approach to directly learn an asymmetric mapping, in which we modify the target model in order to match the source distribution. A summary of this entire training process is provided in Figure 2. Unlike the DAT method which uses a shared feature encoder, our proposed ADDA approach uses separate encoders for the source and target domain data. When there is a significant domain shift, the DAT method may not work well since it inherently assumes that source and target domain data has a similar class distribution.

We define input samples $\mathbf{x} \in \mathbf{X}$ with data labels $\mathbf{y} \in \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ are input space and output space, respectively. In our speaker verification experiments, $\mathbf{x}$ and $\mathbf{y}$ are x-vectors and speaker labels. The probabilistic distribution $D(\mathbf{x}; \mathbf{y})$, however, might be different between training and evaluation dataset due to various domain mismatch such as language mismatch. We denote $S(\mathbf{x}; \mathbf{y})$ and $T(\mathbf{x}; \mathbf{y})$ as source domain and target domain distribution respectively. Our goal is to minimize the distance between the empirical source and target mapping distributions. We firstly learn a source mapping M_s , along with a source classifier C , and then learn to map the target domain encoder to the source domain. Using this approach has demonstrated promising results on cross-lingual text-independent speaker verification tasks using ADDA.

We use SRE04-08, Mixer6 and Switchboard data to train the TDNN based x-vector model. We follow the Kaldi SRE16 recipe to augment the training data by adding noises and reverberations. We use an energy based VAD and the raw feature to train. The model are 23-dimensional MFCCs. Having established the x-vector system using English data, we now try to address the challenge of evaluation enrollment and test speakers for a mismatched language, Chinese. To accomplish this, a set of unlabeled data for the new language is needed. We use the target domain AISHELL

unlabeled training data. We extract x-vectors on source domain SRE and SWBD data and target domain AISHELL unlabeled data to train the adaptation network. We train the Adversarial Domain Adaptation Network (ADAN) in two steps. First, we train a DNN encoder and classifier on SRE and SWBD x-vectors. Next, we use the pre-trained source model as an initialization for the target DNN encoder and perform adversarial adaptation to learn a target domain mapping on the AISHELL unlabeled x-vectors.

For this experiment, our base architecture is a three-layer Deep Neural Network that is fine-tuned on the source domain for 100 epochs using a batch size of 128. When training ADDA, the adversarial discriminator consists of three additional fully connected layers: two hidden layers and an adversarial discriminator output. With the exception of the output, these additionally fully connected layers use a ReLU activation function. ADDA target encoder training then proceeds for another 100 epochs with a batch size of 128. For the DAT training, the shared feature encoder is a three-layer DNN. We use an Adam optimizer with a learning rate 10^{-4} . The speaker classifier and the language classifier are two-layer DNNs. To confuse the language domain classifier, the language classifier has a gradient reversal layer. We use a multi-task loss with equal weights to combine the two cross entropy losses.

Table 5 Speaker verification results using different models with a PLDA Back-end

	EER(%)	MinDCF
x-vector	9.331	0.7755
x-vector + DAT	8.741	0.7475
ADDA embedding	7.645	0.7257
x-vector + PLDA adaptation	9.162	0.7095
x-vector + DAT + PLDA adaptation	7.799	0.6989
ADDA embedding + PLDA adaptation	7.504	0.7062

In this section, we show experimental results using x-vector, x-vector with DAT and x-vector with ADDA training with and without PLDA adaptations in Table 5. We use Linear Discriminant Analysis (LDA) to reduce all three embeddings to 256 dimension for comparison. Also, we concatenate the DAT embedding with the x-vector since we find it always performs better than a single DAT embedding. From Table 5, we observe that our proposed method, ADDA, greatly improves Equal Error Rate (EER) on AISHELL test trials. After ADDA adaptation, EER of the x-vector system decreases from 9.331% to 7.645%, relatively 18.07%. The ADDA approach also achieves relatively 12.54% improvement compared with the concatenated x-vector and DAT embedding. The major reason that ADDA works better might be that it uses an adversarial discriminator to adapt the target encoder to the source domain. Also, by initializing the target representation space with the pre-trained source model, we can effectively learn the asymmetric mapping function.

4.1.4. Domain Mismatch Normalization

i-Vector feature representation with probabilistic linear discriminant analysis (PLDA) scoring in speaker recognition system has recently achieved effective permanence even on channel mismatch conditions. In general, experiments carried out using this combined strategy employ linear discriminant analysis (LDA) after the i-Vector extraction phase to suppress irrelevant directions, such as those introduced by noise or channel distortions. However, speaker-related and – nonrelated variability present in the data may prevent LDA from finding the best projection matrix. In this study, we exclusively use support vectors of each class to find the optimum linear transformation. Post-processing of the i-Vectors by discriminant analysis via support vectors (SVDA) and traditional LDA is evaluated on NIST2010 speaker recognition evaluation (SRE) core and extended core (coreext) conditions.

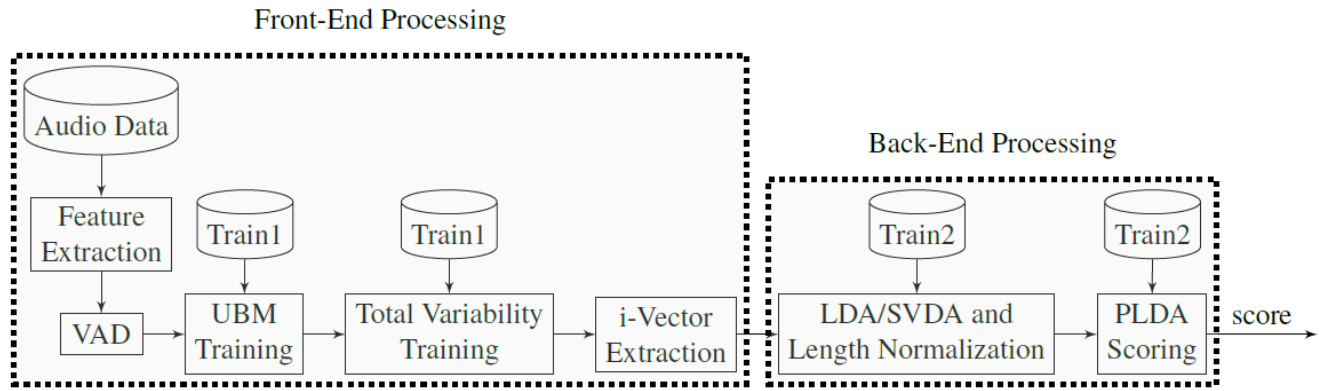


Figure 3 Overview of the i-Vector/PLDA system. Train data used for modeling UBM, total variability (TV) matrix, LDA (or SVDA) and PLDA, enrollment and test data used for evaluation of the system [23].

The overall block-diagram of i-Vector/PLDA speaker recognition used in this study is depicted in Fig. 1. The speaker and channel-dependent GMM supervector in the i-Vector configuration is factorized as:

$$M = m + Tw \quad (2)$$

where m is the UBM speaker- and channel-independent supervector, T is the low rank total variability matrix (TV) that maps the high dimensional GMM supervector into lower dimensional i-Vector representation w . Using training data, the UBM and TV matrix will be modeled by expectation maximization (EM) method. In the E-step, w is considered as a latent variable with normal prior distribution $N(0; I)$. The class separation measure for SVDA is similar to traditional LDA; however, only distinct support vectors will be used to calculate the within class and between class covariance matrices.

From the aspect of classification, [24] shows SVM performs better than LDA. For multi-class problems, Fisher criterion in LDA finds the subspace that gives well-separated classes more

importance than those are closer. Therefore, the classes that are already distinct will move away further from each other; however, the closer classes will not be treated the same. In contrast, SVM focuses more on hardly-separable classes. From this viewpoint, we expect that the transformation find by the SVDA will work better for closer classes. Moreover, the within class and between class covariance matrices calculated by SVDA only uses support vectors instead of using all the training samples. Obviously, SVDA finds the discriminatory directions using the boundary structure of classes; and also SVM is a well-known method for small sample size problem [24]. On the other hand, while solving SVM problem we can adjust the tolerance of classification error; therefore, generalization can be controlled conveniently in SVDA rather than LDA.

The extracted feature vectors contain 19 Mel-frequency features as well as the frame energy appended with delta and delta-delta coefficients. The window length and shift size are 25-ms and 10-ms, respectively. In addition, 3-s sliding window cepstral mean normalization has been applied on feature vectors. Non-speech frames are also discarded using energy based voice activity detection (VAD). 2048-mixture full covariance UBM and total variability matrix have been trained using both male and female data collected from SRE2004, 2005, 2006, 2008 and Switchboard II phase 2, 3 and Switchboard Cellular Part1 and Part2. Next, 600-dimensional i-Vectors have been extracted. The dimension of i-Vectors then reduced to 400 using LDA/SVDA technique. Data used for training LDA, SVDA and PLDA has been restricted to the male speakers (for the sake of tractability) from SRE2004, SRE2005, SRE2006, and 2008. To evaluate the system, we used male trials of core and extended core conditions of SRE2010. All the experiments are carried out on telephony condition (condition 5) of NIST SRE2010. The enrollment/test condition combinations used in the experiments and the statistics of training and enrollment data, as well as trials are provided in Table 6. In addition, to evaluate the performance of the system on short duration test segments, after applying VDA, the first 3, 5, 10, 20 and 40s of extended core test data have been extracted. For training SVM the publically available LIBSVM toolkit has been used.

Table 6 EER/minDCF results comparing LDA and SVDA. The dimension of i-Vectors is reduced from 600 to 400.

Enrollment/Test	LDA	SVDA		
		traditional 1-vs-1	weighted 1-vs-1	1-vs-rest
Core/Core	1.66 / .037	1.13 / .0399	1.25 / .0364	1.42 / .0368
Coreext/Coreext	1.5 / .0297	1.35 / .0308	1.3 / .0287	1.39 / .029
Coreext/Coreext3sec	14.5 / .0984	14.22 / .0974	14.23 / .0974	14.2 / .0975
Coreext/Coreext5sec	9.71 / .0924	9.64 / .0915	9.55 / .0909	9.81 / .092
Coreext/Coreext10sec	5.61 / .0759	5.58 / .0749	5.60 / .0737	5.72 / .076
Coreext/Coreext20sec	3.17 / .0585	3.12 / .0574	3.17 / .0573	3.35 / .0593
Coreext/Coreext40sec	2.48 / .0448	2.4 / .0423	2.37 / .0407	2.42 / .0411

Table 6 summarizes the performance of i-Vector/PLDA speaker recognition comparing SVDA against LDA. To calculate the between and within class covariance matrices using SVDA three strategies have been considered: traditional one-versus-one, weighted one-versus-one and one-versus-rest. In one-versus-one strategy, the SVM will be applied to just two classes, therefore we need to model $C(C - 1)/2$ SVMs; in contrast to the one-versus-rest approach that each class will be classified against all data from all other speakers (need to train C SVM classifiers).

Table 7 Speaker recognition results comparing LDA and SVDA in terms of EER/minDCF without dimension reduction.

Enrollment/Test	LDA	SVDA
		traditional weighted
Core/Core	1.58 / .039	1.45 / .038 1.46 / .04
Coreext/Coreext	1.46 / .0302	1.37 / .0301 1.36 / .0302

Table 7 reports results comparing LDA and SVDA without dimension reduction for core and extended core conditions. In terms of EER, SVDA outperforms LDA significantly; however, in terms of minDCF there is just marginal improvement. In summary, with regard to the number of hyper-parameters, computation time, and the performance (approximate 32% relative improvement in EER) SVDA works really well.

In this study, the effectiveness of SVDA in the i-Vector/PLDA speaker recognition has been studied. The EER and minDCF scores achieved from the experiments carried out on NIST SRE2010 task prove that SVDA consistently works better than LDA. In contrast to LDA that limits the discriminatory information to the centroid of classes, SVDA captures the boundary structure of them.

4.1.5. Longitudinal Variability/MARP

A speaker verification evaluation is presented on the Multisession Audio Research Project (MARF) corpus, for which speakers were recorded at regular intervals, in consistent conditions, over a period of three years. It is observed that the performance of an i-vector system with probabilistic linear discriminant analysis (PLDA) modelling decreases progressively, in terms of both discrimination and calibration, as the time intervals between train and test sessions increase.

A standard i-vector [25] speaker verification system with probabilistic linear discriminant analysis (PLDA) modeling [26] was used for the experiments in this study. At the front-end, 13-dimensional MFCCs were extracted over 20 ms windows at 10 ms intervals, and appended with first and second derivatives. Mean and variance normalization and RASTA filtering were subsequently applied. Speech activity detection (SAD) was applied via Combo-SAD [27], which leverages a combination of speech voicing and spectral flux features. Gender-dependent UBMs of 1024 components were trained with a subset of NIST SREs 2008 and 2010 consisting of microphone speech from speakers of US English. A maximum of 30 seconds of speech post-SAD

was included from each session. Approximately 30 hours of data were allocated for training male and female UBMs respectively. An i-vector extractor matrix T of rank 400 was estimated from the same recordings used to train the UBM. Linear discriminant analysis (LDA) was applied, reducing the i-vector dimensionality to 200. Finally, the i-vectors were mean and length normalized, and whitened. The speaker- and session-dependent i-vector distribution was modelled with PLDA, again using the UBM development data.

From the full set of MARP speakers, a reduced set of 35 males and 25 females was formed by discarding those speakers with less than five sessions of data. In the resulting set, the average number of sessions per-speaker was 14. Since each session was a conversation side of 10 minutes duration, the quantity of active speech for a given speaker was typically less than five minutes. Each session was therefore divided into ‘chunks’ of exactly 60 second duration post-SAD. Fixing the length of these chunks removes the effect of duration [18, 23] from the subsequent experiments.

Between one and four 60-second chunks were extracted from each session (dependent on the quantity of active speech). Chunks were extracted such that they were maximally dispersed throughout the session. An ‘all-vs-all’ evaluation protocol was adopted, whereby all chunks were considered, in turn, as training data, with the remainder of same-gender (different-session) chunks serving as testing data. Performance was then evaluated according to the absolute difference (in time) between the sessions from which the training and testing chunks were drawn. We define the absolute session difference (ASD), for a given trial, as the absolute difference between the indices (2–21) of the corresponding train and test sessions. Each ASD unit therefore represents a time interval of 1–2 months.

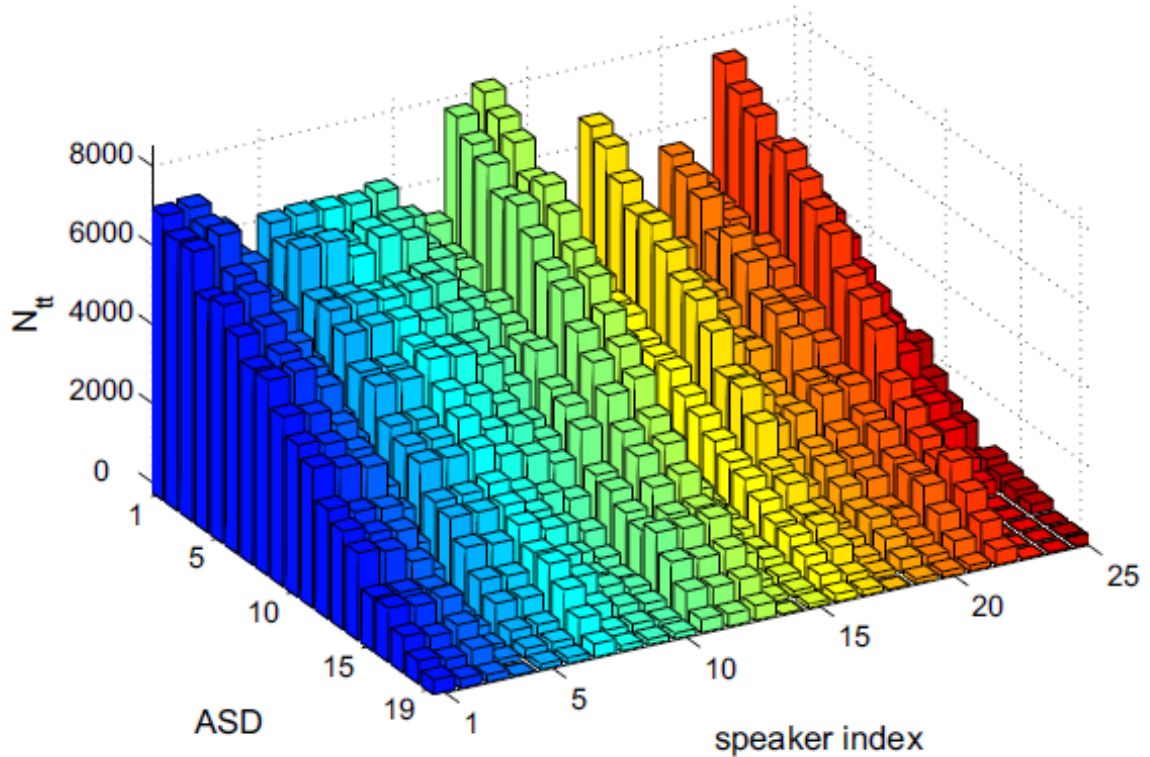


Figure 4 The number of target trials N_t for each of the 25 female speakers across absolute session difference (ASD) with an all-vs-all evaluation protocol.

Figure 1 displays the number of target trials for all female speakers in the all-vs-all protocol, across all ASDs. The distribution of target trials across ASD is similar for the 35 male speakers. It is clear that the number of target trials varies between speakers, and decreases progressively with ASD. The imbalance in the number of trials per-speaker and per-ASD is accounted for at the performance analysis stage via a trial weighting scheme [28] [29], whereby the trials corresponding to a particular speaker-ASD combination are weighted by the inverse of the total number of trials of that combination.

A breakdown of performance over different ASD ranges is shown in Table 8. For males, the EER increases progressively with ASD, from 2.44% to 4.35%. Aside from the EERs at ASD ranges 8–11 and 12–15 being close in value, this increase is approximately linear. The rate of EER increase is similar in the female case, and aside from ASD range 8–11, is also approximately linear. The difference in EER between the first and last ASD ranges is significant: a 78% and 56% relative increase for males and females respectively.

Table 8 EER% for subsets of trials grouped according to absolute session difference (ASD). N_t and N_{nt} denote the number of target and non-target trials respectively.

	ASD range	1-19	1-3	4-7	8-11	12-15	16-19
male	EER%	3.36	2.44	3.02	3.47	3.50	4.35
	$N_{tt} \times 10^4$	7.6	2.3	2.4	1.7	0.8	0.4
	$N_{nt} \times 10^6$	2.5	0.7	0.8	0.6	0.3	0.1
female	EER%	5.16	4.36	4.78	4.55	6.21	6.87
	$N_{tt} \times 10^4$	5.9	1.9	1.9	1.3	0.6	0.2
	$N_{nt} \times 10^6$	1.3	0.4	0.4	0.3	0.1	0.1

To convert raw (uncalibrated) scores s output from the speaker verification system into linearly calibrated likelihood ratios x , the following function was applied:

$$x = w_0 + w_1 s \quad (3)$$

where w_0 and w_1 are offset and scaling parameters respectively. Given a set of development data, w_0 and w_1 were optimized via logistic regression. This optimization relies on the assumption that the score distributions of the development and test datasets are similar. Due to the unique (longitudinal) nature of the MARP corpus, there are no other readily available databases suitable for this optimization. Thus, for the first set of calibration experiments in this section, offset and scaling parameters are optimized on the test scores directly (self-calibration). In a follow-up experiment, independence between development and test datasets is introduced by adopting both cross-validation and cross-gender approaches. To assess the impact of longitudinal speaker variability on conventional score calibration, Equation 1, three scenarios are considered:

- Full (F): one set of calibration parameters is optimized on the full set of scores, i.e. from all ASDs.
- Mismatched (MM): one set of calibration parameters is optimized on scores in the ASD range 1–3.
- Matched (M): five sets of calibration parameters are optimized: one set from the scores of each ASD range considered in Table 8.

Since a relationship was observed between ASD and speaker verification performance, in terms of EER, Table 8, here we propose to extend the score calibration given in Equation 1 by incorporating aging information (i.e. ASD) as an additional term. We adopt the approach introduced in [18], which incorporated recording duration information in calibration via ‘Quality

Measure Functions’ (QMFs). The extended score calibration is given by:

$$x = w_0 + w_1 s + Q(w_2, ASD, \dots) \quad (4)$$

Where Q denotes a QMF defining the way in which ASD (and any additional parameters) are incorporated into the calibration. w_2 is a new calibration parameter to be optimized on the development set. To inspire suitable functions Q for aging calibration, the evaluation scores were first calibrated according to the ‘Matched’ scenario. However, in optimizing the parameters at each ASD range, the scaling parameter, w_1 , was fixed, and the offset parameter w_0 was allowed to vary. This process was similar to the ‘shared scaling’ experiments presented in [18]. The resulting offset parameters for each ASD range, and each individual ASD, are plotted in

Figure 5. Three candidate QMFs are proposed in

Table 9 to approximate the relationship between offset parameter and ASD in

Figure 5.

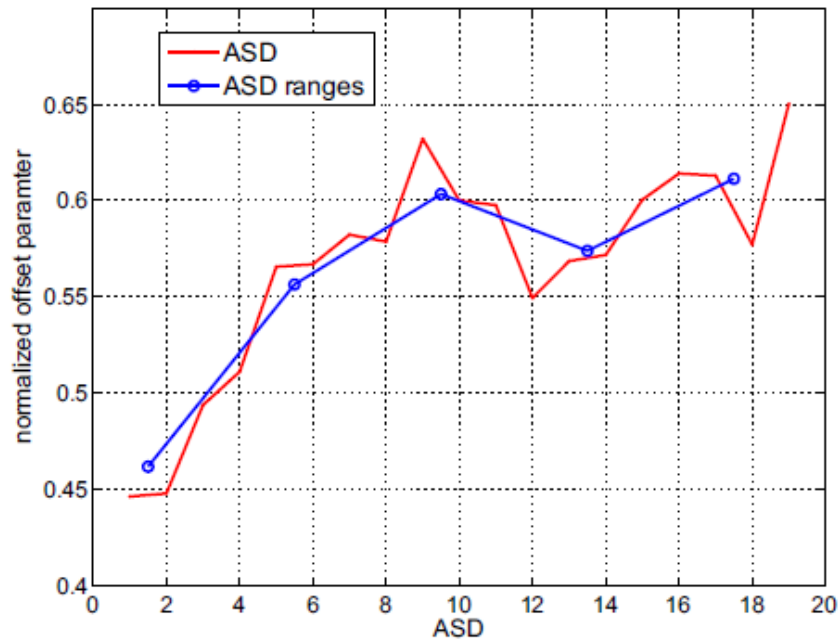


Figure 5 Normalized offset parameters w_0 for individual ASDs and ASD ranges.

Table 9 Proposed quality measure functions (QMFs) for incorporating aging information, i.e. ASD, as an additional parameter in conventional score calibration. λ was set at 0.1 in our experiments.

QMF: $Q_n(w_2, ASD, \dots)$	additional parameters
$Q_1 = w_2 ASD$	w_2
$Q_2 = w_2 \log (ASD)$	w_2
$Q_3 = w_2 (1 - e^{-\lambda \cdot ASD})$	w_2, λ

The evaluation scores of male speakers were self-calibrated according to the three score calibration scenarios Full, Matched and Mismatched, via Equation 1, and according to the three aging calibration proposals,

Table 9. Performance in terms of C_{llr} and $R_{mc}\%$ is given in Table 3 for each of the five ASD ranges. Expectations for the score calibration proposals are for the ‘Mismatched’ (MM) to perform worst, as it uses a limited range of ASDs for parameter estimation, and for ‘Matched’ (M) to perform best, as it represents an ideal (but unrealistic) case where data is available to optimize separate parameters at each ASD range. The performance of ‘Full’ (F) is expected to fall somewhere in between, as it uses the full range of ASDs to optimize one set of parameters. It can be seen from the first three rows of Table 10, that this expected behavior is observed. In all cases, discrimination error and miscalibration increase with ASD range.

The bottom three rows of Table 10 indicate the performance of the three QMF proposals. At all but the first ASD range, all decrease error relative to the ‘Mismatched’ and ‘Full’ cases, and come close to the ‘Matched’ performance. The relative improvement with each QMF increases with ASD: at the 16-19 range, a 37% reduction in R_{mc} is achieved with Q_1 relative to the ‘Full’ case. Although Q_1 performs slightly better than the other QMFs, there is little difference between their performance.

Table 10 Performance of score-only calibration approaches: Full (F), Mismatched (MM), and Matched (M) and proposed score-aging calibration approaches Q_1 – Q_3 .

	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

To evaluate score-aging calibration with parameters optimized on independent data, two approaches were taken. The first was a 10-fold cross-validation (CV) scheme, where 21 male speakers were used for parameter optimization and the remaining 14 males were used for testing. The second was a cross-gender (CG) scheme, where the full set of scores from female speakers were used for parameter optimization, and the full set of male scores for testing. The results of both approaches, and best performing QMFs, are presented in Table 11 and Table 12.

The absolute values of C_{llr} and R_{mc} are much larger than in Table 10. In the CV case, this is due to the differences between score and ASD distributions across speakers. In the CG case, the R_{mc} values are an order of magnitude greater than those in Table 10. Along with the difference in score and ASD distributions between genders, this is due to the separate development data used for male and female systems. Although neither scenario provides ideal calibration data, relative reductions in R_{mc} of 12% and 8% at the 16-19 ASD range are obtained in CV and CG cases respectively, by including aging information. Q_1 performs slightly better than Q_2 and Q_3 in the CG case, suggesting it may generalize best across datasets.

Table 11 10-fold cross validation performance of score-only calibration, Full (F), and a score-aging calibration proposal, Q_3 . The mean values across the 10 folds are presented.

ASD range		1-3	4-7	8-11	12-15	16-19
F	C_{llr}	.105	.123	.155	.151	.229
	$R_{mc}\%$	22.09	13.89	18.32	23.77	51.66
Q_3	C_{llr}	.101	.123	.154	.149	.220
	$R_{mc}\%$	17.44	13.89	17.56	22.13	45.70

Table 12 Cross-gender calibration, Full (F), and score-aging, Q_1 .

ASD range		1-3	4-7	8-11	12-15	16-19
F	C_{llr}	.282	.434	.579	.611	.843
	$R_{mc}\% \times 10^1$	20.65	28.41	33.86	35.94	40.18
Q_1	C_{llr}	.292	.433	.568	.584	.786
	$R_{mc}\% \times 10^1$	21.74	28.32	33.30	33.91	36.79

This study presented a speaker verification evaluation of controlled three-year longitudinal data. Discrimination and calibration rates were observed to increase progressively over this time interval. An EER increase over an interval as short as 6 months was observed, indicating that this phenomenon is an important consideration for speaker recognition systems in practice. An extension to conventional score calibration, incorporating aging as side-information, was shown to reduce discrimination and calibration error, particularly at larger ASDs.

4.1.6. Lombard Effect & Vocal Effort

The Lombard effect is the involuntary tendency of speakers to increase their vocal effort in noisy environments in order to maintain intelligible communication. This study assesses the impact of the Lombard effect on the performance of a current speaker verification system. Lombard speech produced in the presence of several noise types and noise levels is drawn from the UT-Scope corpus. The performance of an i-vector PLDA (Probabilistic Linear Discriminant Analysis) system is observed to degrade significantly with Lombard speech. The resulting error rates are found to be dependent on the noise type and noise level. A score calibration scheme based on Quality Measure Functions (QMFs) is adopted, allowing noise information to be incorporated into calibration.

The UT-Scope corpus [12] consists of speech produced under cognitive and physical stress, along with speech produced under the Lombard effect. A subset of the Lombard portion of UT-Scope, consisting of recordings of 30 native US English speakers (24 female, 6 male), was used in this study. The subjects were recorded in several noisy conditions and in a clean condition. Noise was presented through open-air headphones; therefore, all recordings are clean. Three background noise types were considered: large crowd noise, highway noise (recorded inside a car traveling at 65 mph on a highway with the windows half-open), and pink noise. Additionally, each noise type was presented at three levels: crowd and highway at 70, 80 and 90 dB-SPL, and pink at 65, 75 and 85 dB-SPL. Prior to recording, subjects underwent a puretone hearing test in the range of 100 Hz – 8 kHz, according to ASHA standards, to rule out those with hearing loss. A soundproof recording booth was used to collect recordings over a close-talk microphone. For each subject, 20 TIMIT sentences, five repetitions of 10 digits and approximately one minute of spontaneous speech were recorded at each noise type and noise level. An additional 80 TIMIT sentences were collected in the clean condition.

A standard i-vector framework [25] with PLDA [30] modeling was used for the speaker verification experiments. At the front end, MFCCs of 13 dimensions were extracted over 20 ms windows at 10-ms intervals, and appended with first and second order derivatives. Frame-based mean and variance normalization were subsequently applied. Speech activity detection (SAD) was applied via Combo-SAD [27]. A gender-*independent* Universal Background Model (UBM) of 1024 components was trained with a subset of NIST SRE 2008 and 2010 data, consisting of approximately 60 hours of gender-balanced microphone US-English speech. A 400-dimensional total variability matrix was estimated from the same recordings used to train the UBM. The i-vectors were post-processed by mean and length normalization, and whitening [31]. Their dimensionality was reduced to 200 via Linear Discriminant Analysis (LDA), and the within and between-speaker i-vector variability was modeled with (simplified) PLDA [30], again using the UBM development data.

An evaluation was designed to assess the sensitivity of speaker verification performance to the variants of Lombard speech represented in the test corpus. To consider all permutations of clean (i.e., neutral speech) and Lombard speech, an all-vs-all speaker verification evaluation was adopted, in which each recording was treated in turn as a training sample, with all other recordings serving as testing samples.

Given the trials from the all-vs-all approach, there are many combinations of train and test recording noise type and level that could be analyzed. In this study, the focus is on mismatched trials, i.e., those involving the comparison of one clean recording with one recording of Lombard speech. To establish baseline performance for the system, trials involving only clean speech are also assessed. A detailed analysis of the read speech trials is provided in Table 13, where the performance at each noise type and level is provided. There is a clear relationship between the noise presentation level in dB-SPL and the resulting EER; for crowd and highway noise, there is an approximately linear increase in EER with increasing noise level. With pink noise, the EER increase is greater going from L1 to L2 (65 to 75 dB-SPL) than from L2 to L3 (75 to 85 dB-SPL). Considering the pooled noise condition (ALL), the increase in EER with increasing noise level is approximated well by a linear relationship.

Table 13 EERs (%) for clean-Lombard trials of read speech. Noise levels L1, L2 and L3 represent 70, 80 and 90 dB-SPL for Crd and Hwy, and for Pnk, 65, 75, 85 dB-SPL. The ALL condition denotes the case where trials for all noise types and/or levels are pooled.

	L1	L2	L3	ALL
Cln-Crd	6.09	7.93	10.13	8.54
Cln-Hwy	4.09	6.64	9.57	7.38
Cln-Pnk	3.03	6.04	6.78	6.27
ALL	5.25	7.26	9.60	7.62

We consider a linear calibration transformation [19], in which raw scores s are transformed into calibrated scores s' given offset and scaling parameters w_0 and w_1 :

$$s' = w_0 + w_1 s, \quad (5)$$

where w_0 and w_1 are obtained by logistic regression optimization [32] on development data. Given the varied effects of noise type and noise level on the EER, Table 1, it would be expected that different optimal linear calibration parameters exist for each noise condition. Thus, the ‘conventional’ score calibration defined in Equation 1 will not provide equally effective calibration across all possible conditions.

One solution is to obtain separate linear calibration parameters for each noise condition. Matched calibration can be expected to provide better calibration than conventional calibration, with the drawback of requiring optimization of multiple discrete sets of parameters.

A more flexible solution than matched calibration is to incorporate variability into the calibration transformation. One means of achieving this is through the use of Quality Measure Functions (QMFs) [12] to incorporate ‘quality’ information in a manner that is useful for calibration. In the present case, the noise type and level used to elicit Lombard speech can be considered quality information. In the QMF approach, linear calibration is extended by including an additional term:

$$s' = w_0 + w_1 s + w_2 Q(N) \quad (6)$$

where Q is a QMF depending on some information I associated with the train and or test samples corresponding to the scores s . The w_2 parameter governs the influence of the additional term and is obtained by optimization on a development set in the same manner as Equation 5. This approach can be expected to outperform conventional calibration, but potentially fall short of matched calibration performance.

To propose suitable functions Q for incorporating Lombard noise information into calibration, the relationship between conventional calibration parameters and noise conditions was assessed with a ‘shared scaling’ [12] approach. In this analysis, the scores for different noise conditions are calibrated via Equation 6, with the scaling term w_1 held constant across conditions and the offset term w_0 allowed to vary. The change in the offset term across noise conditions can then inform suitable QMF functions. This procedure was applied to scores at each noise level, for each noise type independently, and for all noise types combined.

In addition to the EER, which represents the point on the DET curve at which the miss and false alarm probabilities are equal [33], several other error metrics are considered in this study:

- $C_{\ell\ell r}$, cost of log-likelihood ratio [17]: this measures discrimination and calibration over all possible operating points on the DET curve.
- $C_{\ell\ell r}^{\min}$, minimum cost of log-likelihood ratio: the minimum bound on $C_{\ell\ell r}$, representing the case of no calibration error. It is therefore a measure of pure discrimination.
- C_{mc} , miscalibration cost: since $C_{\ell\ell r}^{\min}$ contains no calibration loss, a measure of pure calibration can be defined as:

$$C_{mc} = C_{\ell\ell r} - C_{\ell\ell r}^{\min} \quad (7)$$

A cross-validation approach was used to evaluate several calibration approaches on independent sets of development and evaluation speakers:

1. Mismatched: conventional calibration parameters (Equation 5) are optimized on all clean-clean trials.
2. Pooled: conventional calibration parameters are optimized on all clean-Lombard trials.
3. Matched: conventional calibration parameters are optimized on clean-Lombard trials for each noise level.

4. Q_{linear} : both conventional and Lombard calibration parameters (Equation 6) are optimized on all clean-Lombard trials, with a linear QMF.

Calibration parameters were optimized on trials from a random set of 20 speakers and used to calibrate the scores of the remaining 10 speakers. This was repeated 10 times, with error metrics calculated at each iteration. The resulting average metrics are provided in Table 14.

Table 14 Cross-validation calibration of scores from clean-Lombard mismatched trials: performance metrics across Lombard noise types for different calibration approaches. The best performing metric(s) in each case are in bold.

	Crowd				Highway				Pink				ALL			
Calibration	EER%	C_{llr}	C_{llr}^{min}	C_{mc}	EER%	C_{llr}	C_{llr}^{min}	C_{mc}	EER%	C_{llr}	C_{llr}^{min}	C_{mc}	EER%	C_{llr}	C_{llr}^{min}	C_{mc}
uncalibrated	7.55	7.22	0.22	7.00	7.49	6.90	0.21	6.69	6.86	7.36	0.19	7.17	7.61	7.19	0.22	6.97
mismatched	7.55	1.05	0.22	0.83	7.49	1.07	0.21	0.86	6.86	0.80	0.19	0.61	7.61	0.99	0.22	0.77
pooled	7.55	0.34	0.22	0.12	7.49	0.40	0.21	0.19	6.86	0.38	0.19	0.19	7.61	0.37	0.22	0.15
matched	7.52	0.34	0.21	0.13	7.31	0.41	0.20	0.21	6.36	0.36	0.17	0.19	7.34	0.36	0.20	0.16
Q_{linear}	7.31	0.34	0.21	0.13	7.12	0.40	0.20	0.20	6.49	0.37	0.18	0.19	7.33	0.36	0.21	0.15

It is clear that all forms of calibration dramatically reduce those metrics measuring calibration performance, i.e., C_{llr} and C_{mc} . Among the four calibration approaches, mismatched calibration performs poorly; given that mismatched calibration is trained with only clean trials, this would be expected. The remaining approaches are close in performance, with the proposed linear QMF approach (Lombard calibration) achieving the lowest calibration and discrimination error in the majority of cases.

4.2 Open-Set Language ID (LID) / Dialect ID (DID)

As a special case of Language Identification (LID), Dialect Identification (DID) requires us to identify different dialects within the same language class. Because of the remarkable similarity between dialects, DID is considered to be a more challenging problem than general LID. For an i-vector based LID/DID, prior studies have shown advancements with deep neural networks (DNNs) over Gaussian Mixture Models (GMMs) in acoustic modeling. In this study, a novel i-vector representation which is based on unsupervised bottleneck features is examined as the feature to identify dialects from Arabic broadcast speech. To utilize the unlabeled training data, semi-supervised learning with generative adversarial networks (GANs) are incorporated in the back-end classifier development. Experiments with the proposed method in the third release version of the Multi-Genre Broadcast (MGB-3) Challenge yields the best single system performance among all submitted systems.

Here we present a systematic study for DID task with unsupervised/semi-supervised learning concepts in both front-end i-vector modeling and back-end classifier sides. The primary contribution of this study is integrating recent advancements in front-end features, unsupervised bottleneck features (UBNF) for i-vector extraction, and a new semi-supervised learning

framework with GANs for identifying Arabic dialects. The concept of UBNF is similar to traditional bottleneck features, but without the need for text information. In the UBNF paradigm, a universal background model (UBM) is trained to align acoustic features with mixture indexes, which are assumed to represent phonetic like information. As shown in Figure 6, a UBM is trained using MFCCs with Shifted Delta cepstral (SDC) features in the training set. Subsequently, frame level phonetic labels are estimated according to highest GMM posteriors. With the estimated phonetic labels, a deep bottleneck network is trained on the filter-bank features. The DNN architecture is illustrated in Figure 6, a 40-D normalized activation vector is extracted from the bottleneck layer.

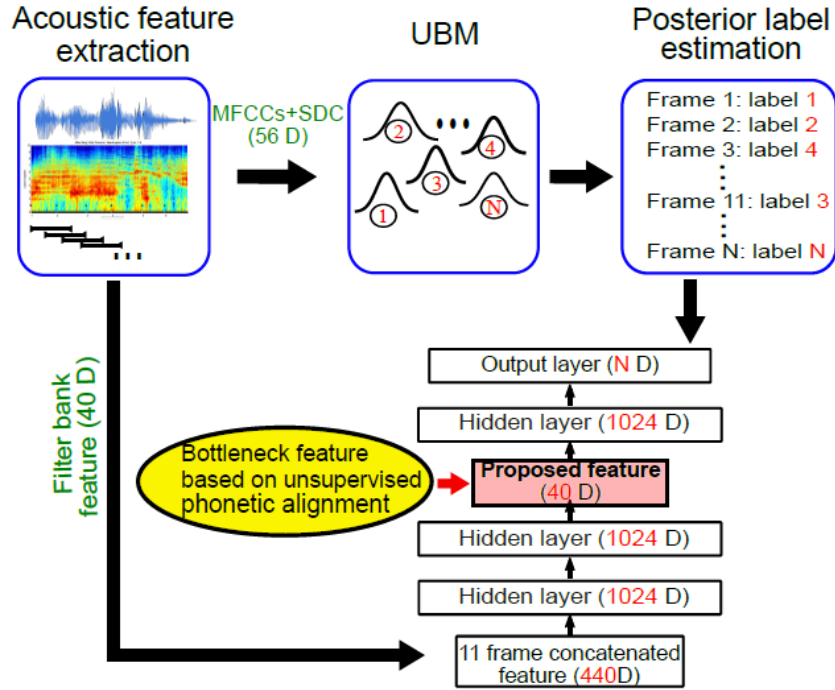


Figure 6 The UBNF feature extraction diagram. Sigmoid nonlinearity is used with softmax normalization for the output layer of DNN.

The proposed UBNF then goes through the standard i-vector modeling process, which includes UBM training, UBM posterior calculation, Baum-Welch statistics extraction, total variability matrix training etc. We then incorporate the UBNF and the concept of semi-supervised adversarial training for dialect identification, given the unlabeled data for the MGB-3ADI challenge.

In the original GANs design, there is a generative network $G(z; \theta_G)$ that produces samples from noise to fool a discriminator network $D(x)$, where the discriminator tries to identify the generated samples. By jointly optimize both G and D , a powerful unsupervised generative model is learned that can produce samples close to the real data [34]. In this study, the GANs discriminator network is not simply for detecting whether the sample is real or “generated”. We wish to ensure that the discriminator network is also a classifier that distinguishes features from different dialects, like a multi-task within GANs framework [35] [36]. To do so, we add the samples from the generator

G to the real K classes data set, labeling them with a new “generated” class $y = K + 1$, so that the classifier is expanded to $K + 1$ classifier becomes:

$$L = -\mathbb{E}_{\mathbf{x}, y \sim p_d(\mathbf{x}, y)} [\log p_m(y|\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim G} [\log p_m(y = K + 1|\mathbf{x})], \quad (8)$$

where $\mathbf{x}$ is the i-vector, y is the corresponding label, $p_d(\mathbf{x}, y)$ is the real data distribution, and $p_m(\mathbf{x}, y)$ is the distribution modeled by the discriminator. We formulate two losses from the cross-entropy loss of the classifier:

$$L_{sup} = -\mathbb{E}_{\mathbf{x}, y \sim p_d(\mathbf{x}, y)} [\log p_m(y|\mathbf{x}, y < K + 1)], \quad (9)$$

$$L_{unsup} = -\{\mathbb{E}_{\mathbf{x} \sim p_d(\mathbf{x})} \log[1 - p_m(y = K + 1|\mathbf{x})] + \mathbb{E}_{\mathbf{x} \sim G} \log[p_m(y = K + 1|\mathbf{x})]\}, \quad (10)$$

Thus, the discriminator D is optimized with all labeled data using the supervised loss L_{sup} ; at the same time, the discriminator also maximizes the probability of the real data to K real data classes with the unsupervised GANs objective L_{unsup} . In this way, unlabeled data is utilized with a semi-supervised framework. A conceptual semi-supervised learning framework with GANs is depicted as

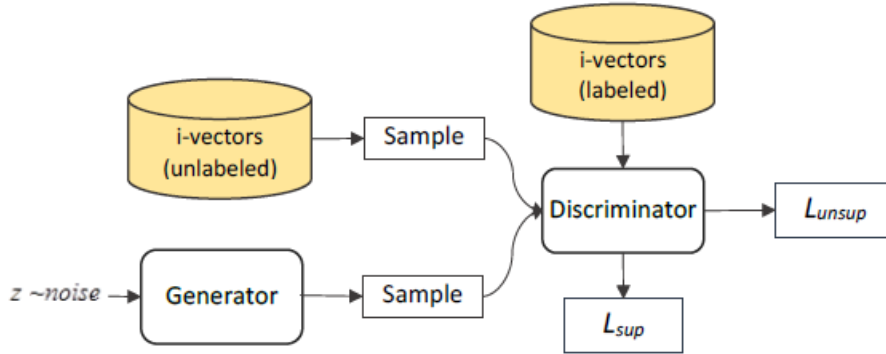


Figure 7 A conceptual semi-supervised learning framework with GANs. The “feature matching” trick is also employed to construct the generator loss, as proposed in [35].

The Arabic Dialect Identification (ADI) training dataset was collected from the Broadcast News domain from 5 Arabic dialects including Egyptian (EGY), North African or Maghrebi (NOR), Gulf or Arabian Peninsula (GLF), Levantine (LAV), and Modern Standard Arabic (MSA). Data recordings were digitized at 16 KHz. The recordings were segmented into utterances with random durations. The test and the development datasets came from the same broadcast domain, but the recording setup was different from the training data. It is noted that unlabeled MBG-2 data is also provided for system development, which is useful in UBM, T-matrix as well as semi-supervised GANs classifier training. The corpus statistics are listed in **Error! Reference source not found.**Table 15

Table 15 Statistics of MGB-3 ADI Corpus. Utt stands for number of utterances, Dur. stands for duration in hours.

Dialect	Training		Development		Eval	
	Utt #	Dur.	Utt #	Dur.	Utt #	Dur.
EGY	3093	12.4	298	2.0	302	2.0
GLF	2744	10.0	264	2.0	250	2.1
LAV	2851	10.3	330	2.0	334	2.0
MSA	2183	10.4	281	2.0	262	1.9
NOR	2954	10.5	351	2.0	344	2.1
Total	13825	53.6	1524	10.0	1492	10.1

Two ADI systems from the MGB-3 challenge organizers are provided as baselines. An additional one MFCC-SDC/i-vector + GB system is also developed by here for cross-system comparisons. One lexical and one BNF/i-vector system are provided by the ADI challenge organizers [37]; both text and acoustic features are followed by a multi-class SVM classifier to find the probabilities of Arabic dialects. We refer to these systems as I_{bs-t} and II_{bs-a} , respectively.

The third baseline is a standard UBM/i-vector system, with 56-D MFCC-SDC (7 static cepstra appended to 49 shifted delta cepstra), energy based speech activity detection and a 1024-mixture UBM (with both MGB-2 and MGB-3 training data). The dimension of i-vector is 600, with a Gaussian back-end implemented to identify each dialect. The system is labeled with III_{bs-gb} .

As illustrated in Figure 6, the same 56-D MFCC-SDC features with an energy based SAD is provided for unsupervised acoustic partitioning; a 2048-mixture GMM is only trained using MGB-3 training set. The dimension of output layer of the deep BNF network is also 2048. The network architecture is summarized as 440-1024-1024- 40-1024-2048, where the input layer is the concatenation of 11- frame filter-bank features. 40-D BNF features extracted from MGB- 2 and MGB-3 training set are the input of standard GMM/i-vector pipeline, and 600-D i-vector used for the final feature for back-end classifiers.

Table 16 Semi-supervised GANs architecture for i-vector features.

Generator architecture			Discriminator architecture		
Input	2×hidden	Output	Input	3×(hidden & dropout)	Output
100	500	601	601	1024 & 0.5	5

The Semi-supervised GANs architecture is shown in Table 16. The generator G is a feed-forward dense network with two hidden layers. A Gaussian noise vector z is the input to G , and the output is the “generated i-vector”. The discriminator D is also the supervised classifier for ADI. It has three hidden layers, where 50% of each hidden layer parameters is randomly dropped out to address overfitting problem [38].

In our submissions to MGB-3 ADI Challenge, duration information is also appended to the 600-D i-vector space as auxiliary features for classifier training. The intuition behind this operation is to add uncertainty to the estimated i-vector before back-end modeling, a performance gain is expected with this simple duration calibration [39]. Our experiments on the MGB-3 ADI Dev set confirms a relative 3% accuracy improvement achieved for different front-end features. Therefore, 601-D i-vector+duration solution is used for all our ADI systems.

Throughout the experiments on the ADI Dev and Eval sets, three metrics (e.g., Accuracy (ACC), Recall (RCL), and Precision (PRC)) are used to report system performance. The single system performance is presented in this section. In addition to three baselines, the UBNF/i-vector + GB and UBNF/i-vector + GANs proposed solutions are also reported in Table 17. These two systems are noted as IV_{u-gb} and IV_{u-gans} , respectively.

Table 17 Dev results of 3 baseline and 2 proposed systems.

System	ACC	RCL	PRC
I _{bs-t}	48.26%	50.33%	49.13%
II _{bs-a}	58.09%	61.37%	58.83%
III _{bs-gb}	63.37%	63.82%	64.29%
IV _{u-gb}	65.88%	66.49%	65.31%
IV _{u-gans}	69.42%	69.83%	69.02%

As shown in from Table 17, the baseline III_{bs-gb} with MFCC-SDC/i-vector and a Gaussian back-end outperforms two MGB- 3 ADI challenge baselines. Compared with other features, the UBNF/i-vector generally has a better characterization for the ADI task when we fix other conditions (i.e., the same Gaussian back-end is applied on III_{bs-gb} and IV_{u-gb}). The GANs based semi-supervised learning outperforms all the systems by a substantial margin, which shows the advantages of our proposed method.

We apply classifiers which are trained with only training data for the Eval set. Table 18 lists performance of the 5 single systems. In terms of ACC, overall performance is slightly lower than the Dev set. This maybe attribute to relatively more short utterances in the Eval set, which is also a motivation to perform duration calibration in the i-vector space described above. The Eval set result does confirm that the UBNF feature and Semi-supervised GANs classifier have consistent advantages over traditional methods.

Table 18 Eval results of 3 baseline and 2 proposed systems.

System	ACC	RCL	PRC
I _{bs-t}	47.64%	48.33%	47.23%
II _{bs-a}	57.33%	59.57%	58.83%
III _{bs-gb}	63.23%	63.62%	63.91%
IV _{u-gb}	65.45%	66.37%	66.89%
IV _{u-gans}	69.16%	70.27%	69.76%

Although the training, Dev and Eval sets are from the same broadcast domain, there is still differences between training and the Dev/Eval set. Here, we note that the Dev and Eval sets are more similar to each other. So, introducing Dev set data for classifier training will benefit more on Eval set. In this experiment, we randomly introduce 2/3's of the Dev data (around 1000 utterances, we leave the other 1/3 of Dev data for score level fusion) to the training set. The performance after Dev data augmentation is detailed in Table 19. We see a 6% relative improvement achieved by simple data augmentation, which shows the importance of in-domain data for the ADI task. An impressive accuracy of 73.86% is achieved with only a single system IV_{u-gans} .

In order to predict final scores combining our multiple single systems. We build a fused model by training a logistic regression model for fusion. Let $\{x = x_1, x_2, \dots, x_n\}$ be the score features by concatenating each single system output. In the logistic regression model, the target binary variable y is a Bernoulli random variable of which the probability of occurrence is dependent on the prediction

Table 19 Eval results of 3 baseline and 2 proposed systems—data augmented.

System	ACC	RCL	PRC
I_{bs-t}	52.61%	53.63%	52.23%
II_{bs-a}	59.78%	62.07%	60.73%
III_{bs-gb}	65.81%	66.62%	66.91%
IV_{u-gb}	69.64%	70.12%	69.89%
IV_{u-gans}	73.86%	74.67%	73.52%

Given below. Regression coefficients ω are estimated using the maximum likelihood estimation. Scores from each single system are combined with the estimated coefficients to get the fusion score $\hat{y}$.

$$p(y = 1|x, \omega) = \frac{1}{1 + \exp(-\omega^T x)}$$

$$\hat{y} = \omega^T x$$
(11)

The system fusion together with the latest updates on this task is reported in Table 6. Among all results, the “UBNF/i-vector+GANs” single system achieves slightly better results compared with current state-of-the-art single system solution [20], which again shows the advancement of UBNF feature for capturing accurate acoustic unit and semi-supervised GANs based classifier for utilizing unlabeled data. Finally, our fused system gives 79.86% accuracy on ADI task with a relative improvement over the MGB-3 ADI baseline by .4%.

Table 20 Comparison with the recent results on MGB-3 ADI task.

System	ACC	RCL	PRC
Shon et.al (fusion) [20]	81.36 %	/	/
Our system fusion	79.86%	80.27%	79.87%
Shon et.al (fusion) [27]	75.0%	75.1%	75.5%
IV_{a-gans}-single	73.86%	74.67%	73.52%
Shon et.al [20]	73.39 %	/	/

The main focus of this study is to develop a single system which can incorporate both discriminative and generative abilities from deep neural networks. GANs based semi-supervised learning has been employed as a back-end classifier for a DID task. The “UBNF/i-vector+GANs” single system improves performance over the MGB-3 ADI baseline with a relative +28.8% accuracy gain, which is state-of-the-art performance reported. Furthermore, system fusion further boosts the performance to a $\sim 80\%$ for ADI task.

4.3 Audio Diarization in Naturalistic Environments (DiarNAT)

Speaker diarization tells who spoke and when in an audio stream. SincNet is a recently developed novel convolutional neural network (CNN) architecture where the first layer consists of parameterized sinc filters. Unlike conventional CNNs, SincNet take raw speech waveform as input. This paper leverages SincNet in vanilla transfer learning (VTL) setup. Out-domain data is used for training SincNet-VTL to perform frame-level speaker classification. Trained SincNet-VTL is later utilized as feature extractor for in-domain data. We investigated pooling (max, avg) strategies for deriving utterance-level embedding using frame-level features extracted from trained network. These utterance/segment level embedding are adopted as speaker models during clustering stage in diarization pipeline. We compared the proposed SincNet-VTL embedding with baseline i-vector features. We evaluated our approaches on two corpora, CRSS-PLTL and AMI.

We investigate SincNet for speaker diarization where the first layer consists of sinc filters. Sinc-Layer learns compact band-pass filters suitable for speaker modeling. It is parameterized by cut-off frequencies of these bandpass filters. The gain of sinc filters is learned by later (convolutional and fully connected) layers in SincNet architecture (see Figure 9). SincNet was developed for speaker recognition in practical scenario where small training data (few seconds/speaker) was available while the test utterances were very short [40]. We leverage efficient SincNet in a vanilla transfer learning (VTL) setup where the SincNet was trained for frame-level speaker recognition on out-domain data and later trained SincNet-VTL was used for extracting speaker embedding from in-domain data (see Figure 8). We investigated several possibilities for extracting features, namely F1, F2 and F3 that were later pooled to fetch segment-level speaker models. We employed length-normalized SincNet-VTL embedding in a diarization pipeline that uses ground-truth speaker segmentation and spherical K-means clustering.

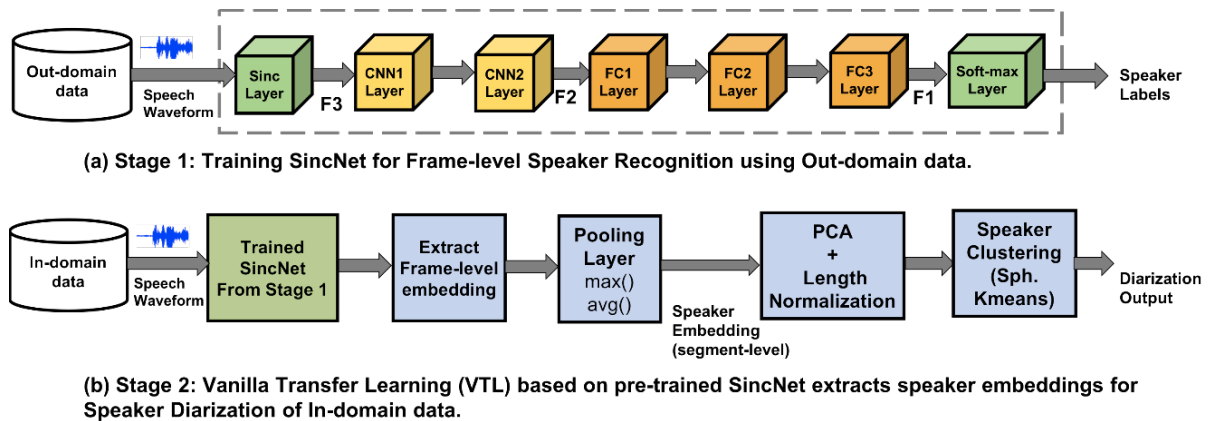


Figure 8 Proposed SincNet-VTL approach for extracting speaker embedding from time-domain speech. (a) In Stage 1, SincNet is trained for frame-level speaker identification using out-domain data. (b) In Stage 2, we adopt the trained SincNet as feature extractor for in-domain data. We max() or avg() pooled frame-level features to get utterance-level embedding.

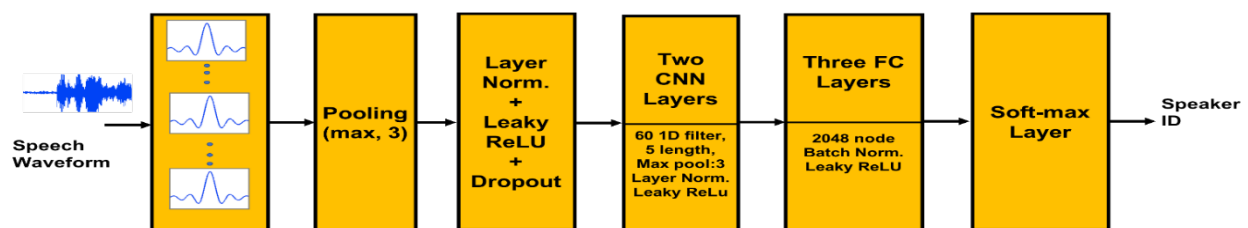


Figure 9 The architecture of waveform SincNet [40]. Sinc-Layer performs time-domain convolutions on raw speech. Next, two 1D convolutional layers and three fully connected layers filter the input. Final soft-max layer perform speaker classification.

Speaker embedding extracted from trained neural networks are emerging alternatives to i-vectors for speaker modeling. SincNet is a recently developed novel architecture designed for efficient processing of raw waveform [40]. Researchers found SincNet superior to CNN for speaker recognition and verification tasks [40]. We used SincNet trained on out-domain data for vanilla transfer learning (VTL). We propose to leverage out-domain data in speaker diarization through SincNet-VTL approach (see Figure 8). We used TIMIT corpus [41] as out-domain data for training SincNet. We ensured text-independent speaker modeling by not including utterances with same text for all speakers, in the training data. Non-speech at the start and end of each utterance was discarded for SincNet training. Time-domain speech signal was divided into 200ms frames with 10ms skip-rate. SincNet was trained using raw speech waveform for frame-level speaker recognition. Sinc-layer parameters were initialized with Mel-scale cutoff frequencies while rest of the network was initialized with Glorot scheme [42]. Final soft-max layer implements frame-level speaker classification. The complete network was trained jointly using RMSprop optimizer with learning rate 0.001. We trained it for 360 epochs with batch size of 64. Trained SincNet-VTL has 462 nodes in output layer corresponding to speakers in training data. We tuned network hyper-parameters on TIMIT corpus. During embedding extraction on CRSS-PLTL corpus there were some segments that lasts for less than 200ms. We repeated those segments until it becomes

a segment of 1s for getting speaker embedding. Since SincNet-VTL was trained on 200ms windows with 10ms skip-rate, we needed at-least 200ms for doing a forward pass on trained SincNet-VTL. We propose pooling frame-level features extracted using trained SincNet-VTL for getting segment-level embedding (see Figure 8).

The CRSS-Peer Led Team Learning (PLTL) corpus contains multi-stream audio recordings collected from five PLTL teams over 11 weeks each, thus leading to 55 sessions. These five teams were chosen from an undergraduate chemistry course. Each PLTL session lasted for approximately 80 minutes and constitute discussions between 6-8 students plus a peer-leader. Peer leader guides the group to arrive at correct solutions without explicitly telling the solution.

During PLTL sessions, each participant wore a LENA device (with not-so-close microphone) for collecting naturalistic audio [43] [44] [45]. In this manner, we collected multi-stream audio for each session (number of streams was same as total participants). The salient features of this data are: (i) many segments with overlapped-speech; (ii) short conversational-turns; (iii) multiple noise-sources; and (iv) reverberation. These factors made PLTL speaker diarization challenging. In this paper, we choose the channel corresponding to PLTL leader for single-channel diarization evaluation. This evaluation set has 8 speakers and lasted for about 80 minutes. It is important to note that many speaker turns lasted for less than or equal to 1 second.

Augmented Multi-party Interaction (AMI) corpus provides speaker annotated multi-modal data from meeting scenarios. In this study, we choose 6 meetings from AMI corpus as evaluation set. These six meetings have four speakers each. We used mixed headset audio for experiments reported in this paper. Our AMI evaluation set consists of sessions: IS1006d (31 min.), IS1003d (36 min.), IS1001a (16 min.), IS1000a (27min.), IS1003b (27 min.) and IS1008d (25 min.).

Diarization error rate (DER) was used for scoring the systems with respect to ground-truth annotations. It was introduced in the NIST Rich Transcription Spring 2003 evaluation (RT-03S). It is defined as the total percentage of reference time that is not correctly attributed to a speaker. Mathematically, DER is given as:

$$\text{DER} = \frac{\Phi_{fa} + \Phi_{miss} + \Phi_{spk}}{\Phi_{total}}, \quad (12)$$

where Φ_{total} is the total time of all reference segments, Φ_{fa} is the system speaker-time not attributed to the reference speaker, Φ_{miss} is the total reference speaker-time not attributed to a system speaker, and Φ_{spk} is the total reference speaker-time attributed to a wrong speaker.

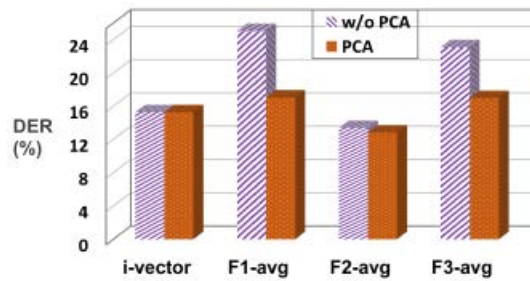


Figure 10 PLTL data: Comparing i-vector with average pooled F1, F2 and F3 embeddings. w/o PCA means without PCA based dimension reduction.

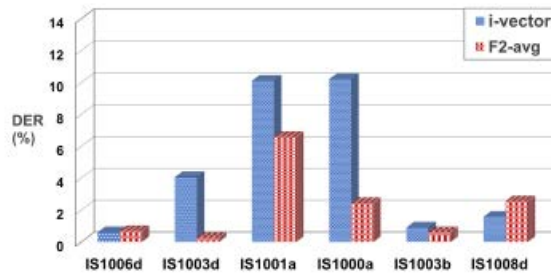


Figure 11 AMI: 6 meetings. F2-avg with PCA (51 dim.) shows significant improvements over i-vector with PCA (51 dim.).

We kept all audio data at 16 kHz for experiments reported in this paper. We trained an i-vector extractor on TIMIT using ground-truth SAD. Since main focus of this paper is to develop a speaker model for diarization, we used ground-truth speaker segmentation information. In this paper, we adopted 75-dimensional (dim.) i-vector as many segments in PLTL were approximately 1s duration (or shorter). SincNet speaker embeddings has dimensions: F1 (462), F2 (2048), F3 (6420). For all our experiments reported here, we perform length-normalization of i-vectors/embeddings followed by spherical K-means clustering with cosine similarity. For some experiments, we employed principal component analysis (PCA) for dimension reduction to 51. Table 21 shows the effect of PCA (51 dimension) on DER (%) for three features:

Table 21 CRSS-PLTL data: Effect of PCA (51 components) on DER (%) for i-vector, F2-avg and F2-max features.

	i-vector	F2-avg	F2-max
w/o PCA	15.26	13.37	43.55
PCA	15.26	12.81	14.36

Looking at Table 21, we see that i-vector did not get DER improvements from PCA as i-vectors were already lower dimensional. We see F2-max embedding have benefited the most with PCA. Even if F2-avg has got relatively small reduction in DER with PCA as compared to F2-max, we get the best DER using PCA on F2-avg embedding. After this point, we stick to average pooling as it was better than max pooling for all the three embedding. Figure 10 shows that F2-avg is best feature for speaker diarization leading to absolute and relative DER improvements of 2.45% and 19.12%, respectively with respect to i-vector baseline. These comparisons were done on PLTL evaluation set as it is our target domain. Figure 11 shows comparison of F2-avg with i-vector features for AMI data. Proposed F2-avg embedding gave significant DER (%) improvement as compared to i-vector baseline on AMI data. On average, F2-avg leads to absolute and relative DER improvements of 2.39 % and 52.06%, respectively over i-vectors.

4.4 Automatic Speech Recognition/Keyword Spotting

Current successful ASR systems employ deep neural network (DNN) models as acoustic model combined with a hidden Markov model (HMM) [46] [47], or use them within an end- to-end configuration [48] [49]. Such systems perform well if they are evaluated in the same condition with their training data. However, in real scenarios, speech typically exhibits wide variation due to the differences in room acoustics and reverberation, speakers and accents, and also environment or recording/channel distortions. For many of these situations, it is possible to simulate (augment) or collect more data to generalize the final ASR system. Given training data of multiple conditions, we need to exploit the data efficiently to train an improved multi-condition (domain) ASR model [50] [51] [52] [53].

4.4.1. Accented Automatic Speech Recognition

The general problem of accent within speech technology is a challenging problem, since non-native speech causes loss in performance for speech recognition and diarization systems. In this study, we investigate employing student-teacher learning to advance a multi-accent model. In student-teacher learning, rather than training a model directly on hard targets, training is carried out in two steps [54] [55]. First, we train several complex, distinct, complementary teacher acoustic models. Next, we train a student model constraining it to mimic the soft-outputs (posteriors or logits) of the original trained teachers. This technique has been successfully applied in the ASR domain: distant-talking ASR [56], multilingual [57] [58], domain adaptation [59], and others.

In this study, we investigate a novel approach of knowledge transference to achieve an advanced multi-accent model. In our proposed scenario, teacher models are accent-specific Recurrent Neural Network – Connectionist Temporal Classification (RNN-CTC) models which are only trained with the corresponding accent English data. However, since the outputs of different trained CTC models are not necessarily aligned, we cannot simply use the ensemble of CTC teacher models. To align the outputs of the accent-specific model, we propose to train them under the guidance of a single multi-accent model. In addition to aligning the CTC trained models, this approach also achieves better generalization. Having the aligned accent-specific models, we train an improved multi-accent model under the supervision of both the teacher models and CTC cost of the target transcription. The resulting multi-accent model significantly outperforms the baseline

multi-accent model trained with no teacher model. Adapting the best multi-accent model to each accent, but constraining it to mimic the soft-outputs of the corresponding accent-specific model generalizes the adaptation process. The proposed approach of adaptation outperforms the knowledge distillation version of KL-divergence [59].

Long Short-Term Memory (LSTM) networks have proven to be effective in many sequential tasks [21]. Having gated memory cells to store information within the network enables them to exploit long-range context and produces related outputs with an arbitrary delay. In the speech recognition domain, an LSTM neural network can outperform conventional RNNs [22]. Alex et al., [23] introduced a new architecture for LSTM networks that receive input features in both forward and backward directions and showed improved performance for acoustic models compared to a unidirectional trained model. In this study, we also employ bidirectional LSTM cells in the recurrent layers to address the problem of accent/non-native speech recognition.

Hinton et al., [16], introduced the term "knowledge distillation", suggesting a new temperature parameter to soften the softmax outputs before being used to guide the training of another neural network. The main idea for student-teacher learning comes from the fact that the distribution of outputs produced by a trained neural network contain underlying relations between output labels. Training another network (student) to output such soft labels which are easier to achieve than hard labels, regularizes the student trained model. As a general setting, having a trained complex model or ensemble of neural networks (teacher models), it is possible to achieve an improved single (smaller) student by constraining it to mimic the soft outputs produced by the teacher(s) [15, 16]. A general framework for teacher-student learning is obtained through the cross-entropy (CE) between outputs of the teacher and student.

Transferring knowledge from a model trained with CTC is challenging [25] and has not been adequately explored. Outputs of a CTC trained model are spiky [26], implying that the model tends to give very sharp posterior probabilities. While the probability of a single class may be close to 1, the rest of the classes are typically closer to 0. In addition, for our setting which uses 'blank', since the model just needs one spike of each character to output the desired transcription, most output spikes are 'blank' which does not have an explicit phonetic similarity with other speech activities, however, there should be some other underlying relation with the neighboring characters. Finally, because of the fact that CTC does an arbitrary alignment between labels and network outputs, as well as having a model with memory which remembers the acoustic states and outputs spikes at any time, the timing of the probability spikes is different from the true frame-character based alignments [27].

Student-teacher learning would be more efficient if it could distill the knowledge from an ensemble of trained complementary models into a single system. It is suggested to construct these complementary models with: training models with different training data, employing alternate architectures of neural network (e.g., convolutional neural network or LSTM), or initializing with different approaches among others. In our case, we consider multi-accent data with the aim to achieve an advanced multi-accent (MA) English model. Each accent has its own underlying similarities between speech acoustic units. For example, in an Indian accent, phoneme /t/, in many words, is pronounced more like a voiced sound making it closer to /d/. However, in native English, /t/ is an unvoiced phoneme that has the same vocal tract configuration but alternate excitation to

phoneme /d/. In addition, for end-to-end ASR that also models the grammar of languages/accents, alternate accents might possess alternate grammar structure or word choice that could influence the relations between acoustic units.

We hypothesize, accent specific teacher models, which are just trained with data from one accent, benefits the student model more than a multi-accent teacher model. However, as is discussed, training multiple LSTM-CTC using different accent speech data results in models with their outputs not aligned. To address this issue, we suggest a novel architecture employing student-teacher model: first we train a multi-accent model with training data from all accents, then we train multiple accent specific models from scratch under the guidance of the multi-accent model. Being trained with the same teacher but seeing only the data of one accent results in aligned accent specific models. The last step is to train a new multi-accent model with these individual teachers from which we obtain soft-outputs of each accent from the corresponding model Figure 12.

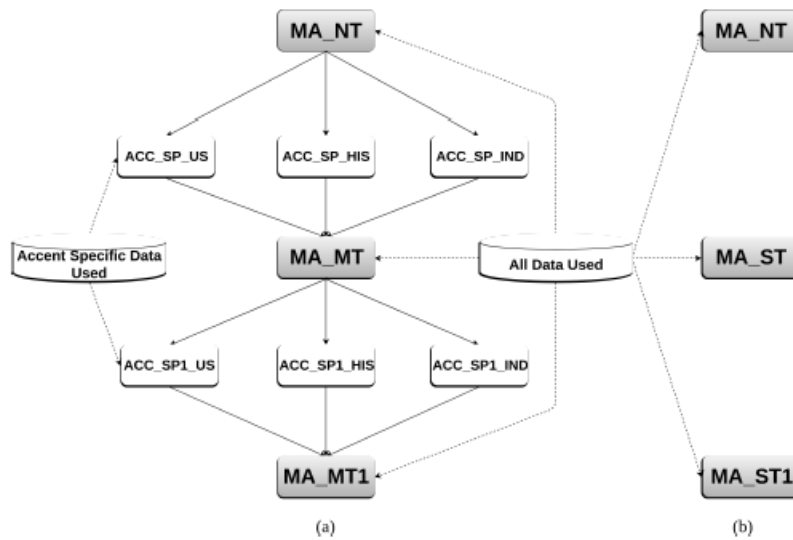


Figure 12 The proposed student-teacher learning to advance a multi-accent model: using accent-specific teachers (a), using multi-accent teachers (b).

Having student models that perform better than the teacher provides a new space to explore system advancements. Could these advanced student models teach another generation of students with the aim to achieve further improvement? The knowledge that comes from an improved model is more accurate and probably more close to actual similarities between labels, providing an easier point for students to achieve and generalize the learning. To this end, we apply our proposed student teacher learning one step further as shown in Figure 12. We employ the advanced multi-accent model from the previous step (MA MT) to train three new accent-specific models (i.e., ACC SP1 US, ACC SP1 IND, and ACC SP1 HIS). Next, we exploit these accent-specific models to teach a further superior multi-accent model (MA MT1).

For training and evaluating the model across accents, the UT- CRSS-4EnglishAccent corpus is used [60]. This corpus of 420 speakers was collected at CRSS-UTDallas and consists of four

major English accents: US (native), Hispanic, Indian and Australian. The data for each accent consists of about 100 speakers balanced for gender and age, with session content that consists of read and spontaneous speech. In our study, we use US, Hispanic and Indian parts of the corpus to train both multi-accent and accent-specific models. In this corpus, for each accent, there is about 28h of training data, 5h of development and 5h of evaluation data.

We extract 26 dim Mel filterbank coefficients for each 25ms frame with a skip rate of 10ms. We expand each frame by stacking 4 frames to each side, then the frames are decimated by skipping 2 frames per each frame for processing. We use the skip process as described in [61].

The neural network architecture starts with two feed forward layers each of 500 neurons, where their outputs go through two bidirectional LSTM layers with 300 neurons in each direction. The LSTM layers are followed by two forward layers each containing 500 neurons. We use Adam Optimizer with an initial learning rate of 0.001 to train the model. Gradients are computed from Mini-batches of 30 utterances. We employ early-stopping by monitoring the performance on a held-out validation set during training epochs. In the evaluation step, we employ a beam search decoding [48] with a beam width of 100 with no language model or lexicon information.

In this section, we examine our proposed approach to address the problem of aligning the CTC trained models. The base- line model for this section is trained with the multi-accent data (i.e., US, Indian and Hispanic English) which is used as a teacher model. Next, we train a US-specific (i.e., only trained with US English part of the corpus) student model in three settings: 1) trained from scratch with no guidance from the teacher (No-Teacher), 2) trained student model with $\lambda = 0.5$, 3) trained student model with $\lambda = 0.9$. We examine two settings of student models to investigate how much the supervision of the teacher influences the student alignments, however, a setting with $\lambda = 1.0$ is not reported because this setting does not result in better ASR performance. In all experiments, default values of λ and T are 0.9 and 4, respectively (unless otherwise specified).

Table 22 Percentage of CSO between the multi-accent model (Teacher model) and different US-specific models.

	No_Teacher		Teacher_ λ :0.5		Teacher_ λ :0.9	
	Train	Test	Train	Test	Train	Test
Teacher model	81%	78%	89%	82%	95%	87%

To examine the overlap between spikes of two CTC trained models, we obtain the index of maximum character per frame for each model, resulting in two sequences of character indexes for each utterance. Next, we calculate an average of overlap between these two sequences for all utterances of the data, referred to as the "characters' spikes overlap" (CSO). Table1 shows the percentages of CSO for training and test utterances of US English data between the baseline and the three accent-specific models. The base- line and No Teacher model have approximately 80% CSO, showing the difference between CTC trained models' spikes.

However, training the student model under the guidance of the baseline model increases CSO to 95% for the train data. The proposed approach could increase CSO to an acceptable point for knowledge distillation where we only need to have aligned models for training data. However,

despite the CSO increase for the test data, there exists some room for improvement in scenarios of ensembling the CTC models in the evaluation step.

Table 23 CER% of adapting a multi-accent trained model to Indian accent using tempered KL-divergence with three soft- outputs: Outputs of the multi-accent model (MA), outputs of accent-specific model with no teacher (No-Teacher) and an accent-specific student model with $\lambda=0.9$ and $T=4$ (Student). The baseline performance is shown in the first row.

	MA (baseline: 14.2) + Adaptation		
	MA [20]	No-Teacher	Student
MA adapted	12.4	12.2	11.7

To investigate whether having high CSO influences knowledge transferring between two models, we adapt the baseline model to Indian accent using tempered KL-divergence [59] with three different soft-outputs (Table 23). Adapting the baseline model using outputs of the model itself as soft- outputs [59] results in a +12.7% relative CER improvement compared to the baseline. However, adapting using the student model outperforms the former setting, demonstrating that accent-specific models provide better soft-outputs which represent the underlying similarities between characters of that accent. Using outputs of No-Teacher model as soft out- puts of the adaptation process performs better than using the baseline’s soft outputs, demonstrating the efficacy of accent specific models even with low CSO.

To examine our proposed multi-step knowledge transferring (Figure 12), first, we train a multi-accent model with all accents pooled together without any teacher information (MA NT). To this end, we train a student model with the same architecture to examine how much the single teacher model can regularize another multi-accent model (Figure 12 b). Having fixed soft character alignments from the teacher model not only generalizes the student model (MA ST), it makes the training more stable compared to CTC which changes the alignments dynamically during the training steps. As shown in Table 24, the resulting MA ST model achieves improved performance compared to MA NT model. To consider the proposed two steps of knowledge transferring in the single teacher case, we exploit MA ST to teach an even more improved student model (MA ST1). This MA ST1 achieves a greater decrease in CER for all accents, demonstrating the efficacy of knowledge distillation over multiple teacher-student generations.

Table 24 CER% of multi-accent without teacher model (MA NT), Multi-accent with a single teacher mode (MA ST), multi-accent with accent specific teacher (MA MT) and accent specific models on US English, Hispanic English (HIS) and Indian English (IND).

Model	Teacher Model	US	HIS	IND	Ave
MA_NT	None	14.1	13.5	14.2	13.9
ACC_SP0	None	17.3	18.2	17.6	17.7
ACC_SP	MA_NT	15.2	14.5	13.2	14.3
MA_ST	MA_NT	11.9	11.6	12.0	11.8
MA_ST1	MA_ST	11.5	10.8	11.7	11.3
MA_MT	Acc_Sp	11.3	10.8	11.4	11.2
Acc_Sp1	MA_MT	14.2	14.0	12.7	13.6
MA_MT1	Acc_Sp1	11.2	10.8	11.3	11.1
MA_MT1_Adpt	MA_MT1	11.2	10.8	10.9	11
MA_MT1_Adpt1	Acc_Sp1	11.2	10.7	10.5	10.8

In this section, we investigate the effectiveness of having multiple accent-specific teacher models to train a multi-accent model. We train three accent specific models from scratch as well as under the guidance of MA NT. Each model is only trained with the corresponding accent speech data (ACC SPs models). These resulting accent specific models perform better than accent specific models which are trained with no teacher (ACC SP vs. ACC SP0 in Table 24). These improvements show that using soft-outputs of the teacher model not only aligns the student models (Table 22), it regularizes them as well.

We leverage the three accent-specific aligned models (i.e., ACC SP US, ACC SP HIS, and ACC SP IND) to train an overall improved multi-accent multi-teacher model (MA MT). As shown in Figure 12, soft-outputs of each accent data are produced by the corresponding ACC SP model. The resulting model outperforms both former improved multi-accent models (MA ST and MA ST1), demonstrating that accent specific models provide better underlying relations between output characters, resulting in a more generalized student model. Comparing the average CER of ACC SP models (14.3%) with MA ST (11.8%) demonstrates that for our multi-accent data, accent-specific teachers are more effective than a superior multi-accent teacher model. Following the diagram of Figure 12 to transfer knowledge one step further, we achieved the best multi-accent model (MA MT1) which outperforms all other multi-accent models yielding a relative CER gain of +20.1% vs. baseline (MA NT). Adapting the best multi-accent model to each accent using both Acc Sp1 (see Figure 12) and MA MT1 to regularize the adaptation process, again supports the idea that accent specific model outputs perform better than multi-accent model outputs (MA MT1 Adpt vs. MA MT1 Adpt in Table 24).

We proposed to train multiple CTC accent-specific models under the guidance of a single multi-accent teacher model to align their outputs. This approach not only aligned the CTC trained model, but also generalized the resulting student models. To achieve an advanced multi-accent model, we also proposed a novel approach of knowledge transfer, where we trained a multi-accent model (student model) under the supervision of accent-specific aligned models. The proposed method

was shown to significantly outperform the base- line multi-accent model trained without any teacher model. Having this advanced multi-accent model leads to further improvement, by training new accent-specific models from which we guide a new multi-accent model. This second step of knowledge transfer yields the best multi-accent model providing a +20.1% CER gain over the original baseline multi-accent model. Accent-specific models not only led to the best multi-accent model, their soft-outputs regularize the adapting process of the multi-accent model to each accent. Finally, constraining the adapted model to imitate the accent specific models' outputs results in a more generalized adapted model compared to the method employing outputs of the multi-accent model itself.

4.5 Noise & Speaker Modeling for Knowledge Assessment

Noise and speaker modeling for knowledge assessment research has focused on identifying and addressing noise and distortion that will impact automated processing such as clipping as well as identifying supplementary speech information such as sentiment.

4.5.1. Speaker Recognition with Nonlinear Distortion: Clipping Analysis and Impact

The formulation of advanced speech and language technology in the past has historically relied on carefully organized data collection, typically in controlled laboratory conditions. However, today a vast amount of audio data appears daily on the web, data servers for call centers, and personal devices such as smart phones. Due to the cost in collecting organized and focused speech/language corpora, researchers have turned to found naturalistic audio material in order to reduce both cost and collection time, as well as increase the diversity of speakers, languages, and topic content. While there is a plethora of audio material available, non-uniformity and recording impurities could raise great concerns regarding the viability of resulting algorithms.

Audio peak clipping occurs when the volume of the recorded audio signal exceeds the input range of the micro- phone's pre-amplifier, or the audio data is recorded without an appropriate input automatic gain control (AGC). Portions of the signal above the maximum voltage would be clamped to the maximum value of the signal when it passes through the analog to digital (A/D) converter. The loss of high amplitude samples introduces a non-linear distortion in the form of odd harmonics in high frequencies, resulting in audible artifacts in the recorded audio.

Given the ever increasing availability of speech and audio data in the field, the speech community needs improved speech tools to better characterize and understand issues in found data. In this study, we restrict the potential acoustic issue to wave- form peak clipping, which occurs when audio data is recorded at an improperly adjusted gain setting and without an input AGC, or when the AGC cannot respond quickly enough to suppress impulsive audio events. Clipping detection can be even more challenging when clipped data is transformed to different for- mats (e.g., wav, mp3, etc.), since peak clipped values can be re-assigned to new clipped max values. Speech peak clipping occurs when the volume of the speech 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, and all portions of the signal above the maximum

voltage of the pre-amplifier will be clamped to the maximum value as it passes through the A/D converter. The manifestation of this loss of data, and the introduction of this plateau shape comes in the form of non-linear distortion, especially odd harmonic distortion in higher frequencies, resulting in audible artifacts in the recorded audio.

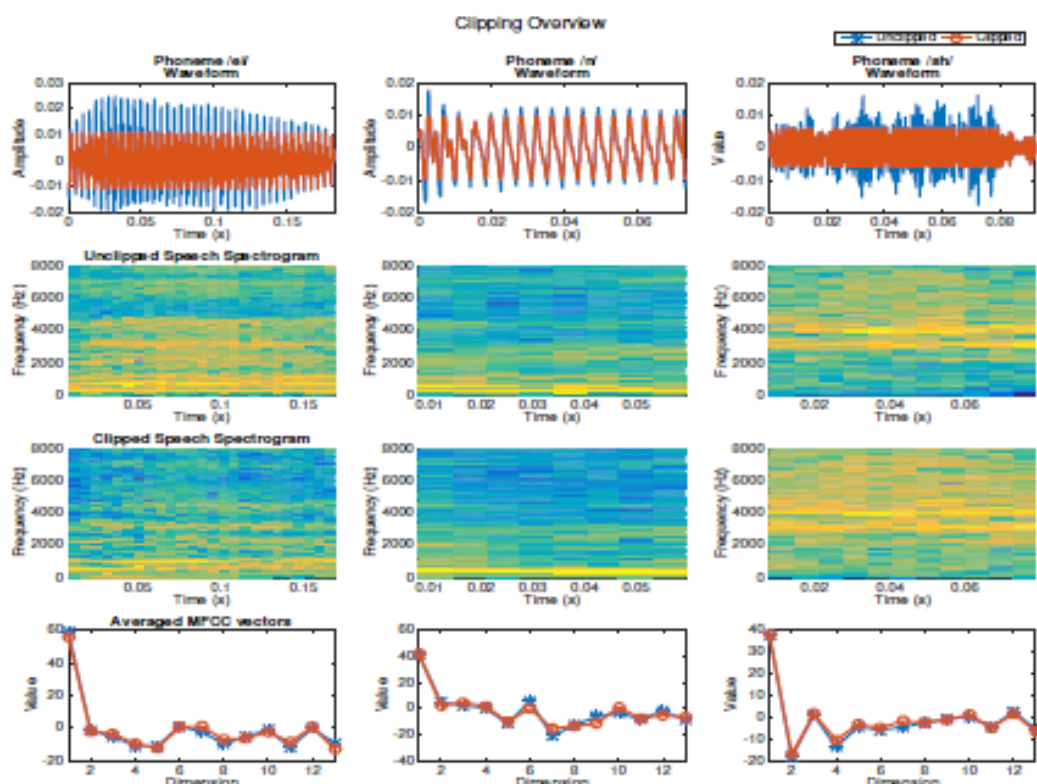


Figure 13 Comparison of original clean and clipped speech that includes waveform, spectrogram, and averaged MFCC feature vectors for three phonemes (/ei/, /n/, /sh/)

We illustrate an example of the effects of clipping for speech signals in Figure 13. The first row presents time domain waveforms of the /ei/, /n/, and /sh/ phonemes from left to right respectively, with blue representing the original unclipped signal, and red the corresponding clipped version. We create these visuals by isolating the three phonemes and increasing the gain of each until 10% of the audio time domain samples are clipped.

We show spectrograms of the original unclipped waveforms in the second row, versus clipped waveform spectrograms in the third row. Comparing the corresponding spectrogram plots of each phoneme, we observe that the largest impact of clipping is distortion harmonics appearing in the higher frequencies of the spectrogram. This manifests as more energy within the upper regions of the spectrogram. We note that the lower frequencies, and the overall shape of the speech energy information still remains, and the formant structure appears to be present and intact. This would imply that the audio is still intelligible, and that, for this level of clipping, the content is not degraded to the point that it is very difficult to understand. Our subsequent listener test also

supports this observation, with the speakers information content is very clear, but with noticeable high frequency artifacts present such as pops and hisses. The fourth row of Figure 13 displays the averaged mel-frequency cepstral coefficients (MFCC) plots for the 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 show how different clipped versus unclipped files would be interpreted by automatic speech systems. The overall shape of the clipped-audio MFCC vectors remains relatively similar to their unclipped counterparts. However, there is sufficient variation in individual coefficients which could impact the performance of speech systems.

We propose a Clipping Detection and Tagging (ClipDaT) algorithm to detect signal clippings for naturalistic audio streams. Signal clipping may occur at the maximum level of the output digital representation. It may also occur when the speech signal is re-transmitted or converted to other audio file formats, where a renormalization of the overall gain is introduced after clipping.

The ClipDaT algorithm searches for strings of consecutive audio samples whose amplitudes reach a specific +/- maximum sample found in the file, or whose amplitudes come very near to this +/- maximum. We describe the clipping detection procedure in the following pseudo-code:

Algorithm 1 Sequential Clipping Detection

Let $x(n), n = 0, 1, \dots, N - 1$ be the initial discrete time signal

for $i = 0 \rightarrow N - 1$ samples **do**

 Find the maximum absolute value of the signal x_{max}

 Find all M peaks within a very small range $x_{max} \pm \epsilon$ of the speech amplitude distribution, denote peaks as $\{M_0, M_1, \dots, M_{M-1}\}$

 Estimated clipping threshold θ is given by

$$\theta = \frac{1}{M-1} \sum_{i=0}^{M-1} M_i$$

 If $x(n) \geq \theta$ or $x(n) \leq -\theta$

 counts $\leftarrow$ counts + 1

 Clipping rate $\gamma = \frac{counts}{N}$

end for

The first step is to perform an initial pass through the audio file to determine where the +/- maximum sample values are located. During a second pass, the algorithm searches for these +/- maximum values. Once one of these extreme samples is encountered, the immediately succeeding samples are analyzed in turn to decide if the extreme sample is the start of a string of clipped samples, or an isolated peak itself. If two consecutive extreme samples are found, we tag it as a clipping event. For the following samples to be deemed part of this clipping event, they must remain within a prescribed range of the extreme value. We set the clipping threshold by computing the average of speech signal peak values. We adopt this approach after analyzing a number of clipped speech files and discover that a high amplitude input that causes clipping may not result

in a steady and flat waveform. We observe some slight variation in amplitude values once the pre-amplifier has been saturated.

We use four speech quality measures to evaluate the clipped speech data. The first, Perceptual Evaluation of Speech Quality (PESQ) [62] is a full reference measure derived from the ITU-T P.862 recommendation regarding speech quality for telephony applications. The PESQ speech quality compares the perceptual difference between clean and degraded signals. The output from the PESQ algorithm is a mean opinion score (MOS) from -0.5 to 4.5, with 4.5 being the highest score, representing excellent quality.

The speech-to-noise-ratio (STNR) algorithm developed by NIST (NIST-STNR) estimates the distance between the speech and noise content of a bi-modal histogram distribution of an audio file. It is a non-reference quality measure and higher values represent a better quality measurement.

Waveform amplitude distribution analysis signal-to-noise ratio (WADA-STNR) is the third quality measure evaluated. Kim and Stern [63] proposed WADA-STNR as a non-reference quality measure. This approach tries to model the amplitude distribution of a speech waveform with the Gamma distribution function, while assuming that the background noise could fit a Gaussian distribution.

Finally, we measure the Sources to Artifacts Ratio (SAR) of each clipped file compared to its corresponding clean file using the Blind Source Separation Evaluation toolbox [64].

Table 25 Table 2: Comparison of four speech quality measures for original clean speech, and different levels (0.5%-10.0%) of waveform clipping on TIMIT data.

Clipping Rate(%)	Quality Measure			
	NIST STNR	WADA STNR	PESQ	SAR
0	49.6	80.7	4.1	—
0.5	48.8	83.8	3.9	17.9
1	48.2	83.1	3.7	15.7
5	44.6	77.7	3.1	10.4
10	40.8	71.5	2.7	8.3

Table 25 presents results obtained when evaluating the data from all 630 speakers of the artificially clipped TIMIT dataset with the aforementioned speech quality measures. All four speech quality measures are clearly sensitive to the presence of clipping, with scores diminishing as the level of introduced clip- ping increases for each algorithm. We can see that the WADA SNR reported for the unclipped audio recordings is lower than both 0.5% and 1.0% clipped recordings, suggesting that the original TIMIT dataset may have some slight distortion. In addition, we can observe a clear degrading trend compared to the original unclipped audio. There is a high and consistent correlation between clipping rates and these quality measures, where 1% to 5% clipping having a much larger impact on speech quality scores compared with the distortion change from 0.5% to 1% clipping.

Additionally we further investigate the effect of clipping using an i-Vector Speaker Identification approach [65] [66] [31] with PLDA-based scoring method [30] [67] on the NIST SRE-08 data. NIST SRE-08 data includes multiple microphone and recording environments such as conversational telephone speech and speech data recorded with a microphone in an interview scenario. Some speakers in the telephone conversational data are bilingual and their evaluation data may include non-English speech as well. Here, we focus on the core short2-short3 test condition of the SRE08 data. This condition includes two- channel telephone conversational excerpts of about five minutes total duration. We perform gender dependent speaker recognition on all male data. All training and test data involve only English language telephone speech spoken by Native American English speakers. The speech data are sampled at 8 kHz and coded in PCM format. We use all previous SRE data to train the background model and a total of 39,433 test trials are used in this evaluation.

We first extract 60 dimensional MFCCs, including delta and delta-delta acceleration vectors. A 2048 mixture Gaussian Background Model is trained on previous SRE data. We extract 400 dimensional i-Vectors ω for each enrolled speaker and test utterance. Finally, we use the Probabilistic Linear Discriminant Analysis (PLDA) scoring method to perform the log-likelihood ratio test.

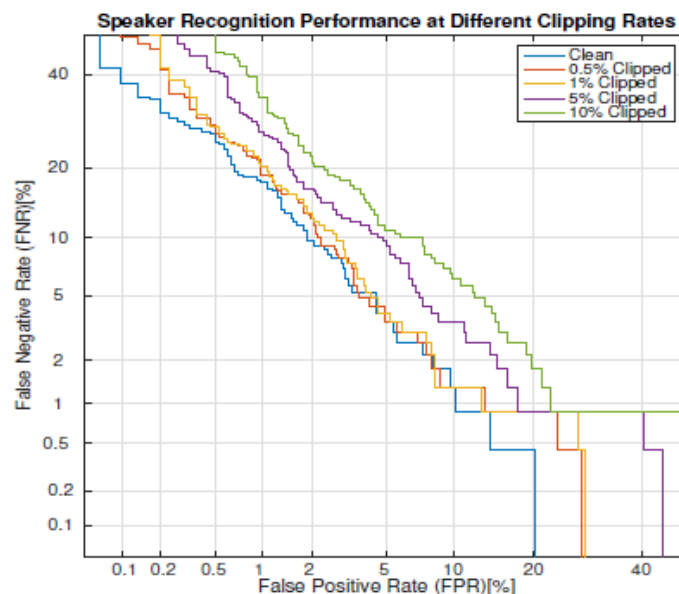


Figure 14 DET curve results using i-Vector with PLDA approach on NIST SRE-08 corpus: performance for 0%, 0.5%, 1%, 5%, and 10% clipping distortion.

Figure 14 shows the detection error trade-off (DET) curve of our speaker recognition system at 5 different clipping rates. From the figure and Table 26, we observe that with an increase in clipping rate, speaker verification performance drops correspondingly. When clipping rate increases from 5% to 10%, we see a significant performance loss with respect to Equal Error Rate (EER) and minimum Detection Cost Function (DCF). Due to the non-linear characteristic of speech signal clipping, the speaker dependent information contained in voiced segments are more likely to be clipped, if the input gain is set too high for a given recording. If only a very small amount of

samples (e.g. less than 1%) are clipped, it may not have a notably negative affect for speaker recognition. This is because at 1% clipping rate, sufficient high energy voiced speech content which possesses significant speaker discriminant information remains intact.

Table 26 EER and Minimum DCF results on NIST SRE-08 corpus: performance for clean trained speaker models and clean, 0.5%, 1%, 5%, 10% clipped test data.

Test Data	EER(%)	DCF
Clean	4.8	0.0257
0.5% clipped	4.4	0.0287
1% clipped	4.8	0.0289
5% clipped	6.6	0.0339
10% clipped	8.3	0.0408

Here we have established clipping has a degrading effect on the accuracy of speech systems including automatic speaker recognition. The effects range from a trivial drop in accuracy when only a small percent of data is affected, to a significant performance loss when clipping contamination is widespread. We find that if only a small number of files within a dataset suffer from clipping contamination, the negative performance impact will likewise be minimal. As the clipping rate increases to 5%, the negative impact that may reduce system performance is much higher. We should be aware of this phenomenon before processing speech data, and be responsible for inspecting whether the severity of clipping within a given data set is too high.

4.5.2. Automatic Sentiment Detection in Naturalistic Audio

Text based sentiment detection is an established field in natural language processing (NLP). Sentiment analysis / opinion mining, analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [68].

Speech based sentiment extraction is an emerging and challenging field. In this study, robust methods are presented to extract sentiment/opinion from natural audio sources. A hybrid system is developed which utilizes a robust Automatic Speech Recognition (ASR) system in tandem with NLP based sentiment analysis techniques to detect sentiment of audio streams. Unlike text based sources, audio sources have a high degree of variability both in terms of expressing opinion as well as the mode of expression of the opinion. There are a range of challenges for sentiment extraction in highly natural speech sources including:

- 1) *Domain and Vocabulary*: The speaker can express opinions about any topic, (e.g., products, movies, politics, social issues, games, etc.) Hence the ASR system should be efficient to handle a wide range of domains and vocabulary. The language model should be comprehensive.

- 2) *Speaker Variability and Speaker Accents*: ASR system should be robust to speaker variability which includes a wide range of English accents from all over the world.
- 3) *Noisy Audio and Channels*: Inconsistent recording equipment and different mode/distance of recording, inconsistent acoustic and background environment conditions make the sentiment detection problem challenging. Also, background music/talk, intentional music mixing, reverberation issues make the problem harder.
- 4) *Natural and Spontaneous*: Detecting audio sentiment in natural and spontaneous speaker settings and various speaker interactive scenarios (i.e., 1-way, 2-way, public speech etc.) is challenging.

A simple way to perform sentiment detection for audio data is to process ASR output transcripts through a text-based sentiment detection system (such as the ME method described in previous sections). In this study, this system constitutes the baseline for audio based sentiment detection experiments.

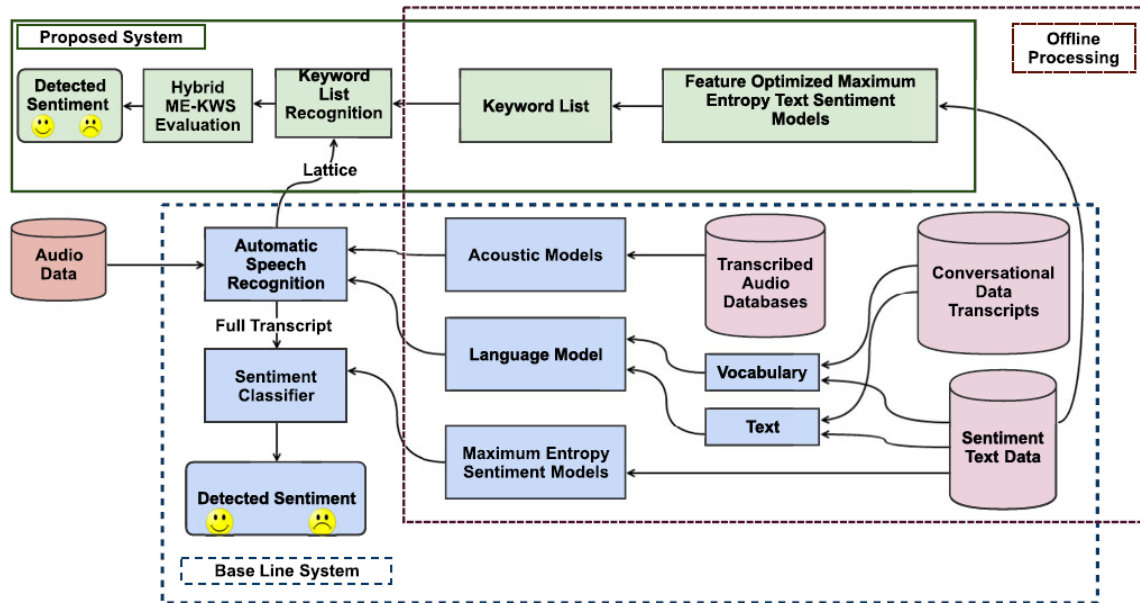


Figure 15 Baseline audio sentiment detection system uses Automatic Speech Recognition (ASR) to generate transcripts which are then processed by the sentiment classifier. Proposed sentiment detection system uses Keyword Spotting (KWS) to detect sentiment-bearing terms in the audio data. An iterative feature pruning technique is proposed to generate the sentiment-bearing terms for KWS. Offline development of customized text based sentiment models.

Figure 15 gives an overview of how ASR based sentiment extraction works. The automatic speech recognition (ASR) system used in this study is the KALDI speech recognition system [69].

The acoustic models are trained using a mix-style approach, where acoustic data from multiple corpora are used. We have used Conversational Telephone Corpora consisting of Switch- board and Fisher corpora to train our acoustic models. The above corpora constitutes around 2000 hours of speech data. Part of the training data is corrupted with additive noise at various SNRs (signal to noise ratio) to create a mixture of clean and noisy conditions (to mimic various kinds of noises like office, fan, cafeteria, car road which we expect to see in the evaluation corpora used in this study). The acoustic models use Mel frequency cepstral coefficients (MFCCs) along with deep neural networks (DNNs). The ASR vocabulary size is about 120 K words. The vocabulary contained in the 64 K sentiment feature set (described in previous sections) is part of the ASR vocabulary. Additionally, the language model for the ASR is trained on data from two sources, namely, CTS (conversational telephone speech) corpora and reviews dataset.

Triphone language models are trained using various text sources like Switchboard [70], Fisher [71] and UW191 [72] (191 M) words collected from the web by the University of Washington) and entire six sentiment corpora comprised of:

- 1) Amazon Product Reviews [73], [74], [75],
- 2) Pros & Cons database [76],
- 3) Comparative Sentence Set Database [77],
- 4) Opinion Lexicon Database [73],
- 5) Scale database [78], and
- 6) R-T polarity database [73].

Altogether, the combined text dataset for language model training contained more than 1 billion words. In this manner, the language context of the sentiment features is captured in the language model.

The ASR uses a 2-pass decoding strategy, where an fMLLR transform is estimated in the first pass and used to normalize the MFCCs. In the second pass, the normalized MFCCs are used with DNN models to obtain the output recognition lattice. Subsequently, one-best transcripts are obtained from the lattice. Finally, the one-best transcripts are analyzed by the Maximum Entropy sentiment detection model to obtain the sentiment decision for the audio file.

Sentiment bearing terms are relatively sparse in both speech and text. Therefore, a large majority of the ASR transcript, (as well as the errors in the transcript), have limited impact on sentiment detection accuracy. When compared to ASR, Keyword Spotting (KWS) provides the ability to focus on relevant terms and phrases, and ignore irrelevant (from an application perspective) background text. Therefore, it is suggested that KWS would be better suited for sentiment detection as opposed to employing full ASR transcripts directly.

In order to further explain the above mentioned argument, we present a simple experiment that demonstrates the impact of ASR WER (word error rate) on sentiment detection. In this experiment, we add controlled amounts of substitution errors (to simulate a required word error

rate) in our text based evaluation dataset. While simulating substitution errors, words were randomly chosen and replaced by other words from the ASR vocabulary. This process generates various noisy versions of the evaluation dataset. Using our best sentiment detection model, we then compute sentiment detection accuracy for these noisy variants, and compare this accuracy to the original clean version.

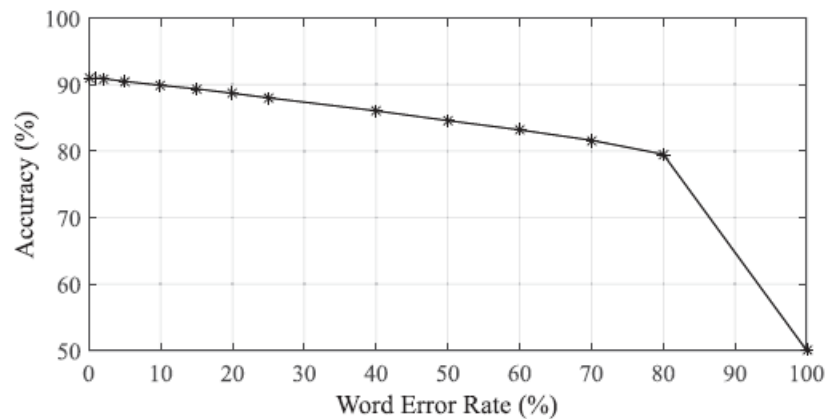


Figure 16 A plot of the variation in sentiment detection accuracy vs. simulated word error rate (WER). The trend suggests that only a small fraction of words can effectively determine sentiment, and most words are irrelevant from a detection perspective. This implies that KWS may be a feasible and superior solution for sentiment detection over ASR.

Figure 16 shows sentiment detection accuracy obtained on noisy variants of the evaluation dataset vs. the amount of controlled WER introduced in the transcript. This shows the tolerance of sentiment detection to ASR word errors. It can be seen that as simulated WER increases from 0% to 80%, there is a gradual, almost linear drop in average sentiment detection accuracy from 91.7% to 80%. When simulated WER increases beyond 80%, we see a sharp decrease in accuracy (and the system reaches random performance, (i.e., 50%) at 100% simulated WER). This simple experiment shows that automatic sentiment detection in audio can tolerate high amounts of WER. In other words, successful sentiment detection depends on finding the few effecting sentiment relaying terms/phrases. This finding is further strengthened by empirical observations that sentiment bearing terms are relatively rare in text and audio (for example, approximately 2-to-5% relative frequency of occurrence in re- views). Taken collectively, this implies that sentiment detection should be possible with KWS using a relatively small keyword list, and it is likely to outperform full transcript ASR. Though an ASR system might generate word errors differently than a random substitution word error simulation; however the sentiment distance between the correct ASR word errors vs randomly selected word errors should be equally large/divergent. Therefore we feel the simulation accurately represents the system's ability to accurately access sentiment of the content. It is difficult to force a predetermined WER in ASR systems for Keyword Spotting.

The KWS system developed is an extension of our KALDI based large vocabulary ASR system. Using the word lattice generated by the ASR, a finite state transducer (FST) based method is used to search the word lattices for keywords [72]. In parallel, the word lattices are converted in to

phone lattices, and the PCN-KWS (phone confusion network keyword spotting) method employed to search for keywords as well [70], [71]. Subsequently, the search results from the two methods are combined (by simple likelihood combination) to yield the final keyword result list. The KWS system assigns log-likelihood values to keyword hypotheses in the audio. The higher the value of the log-likelihood, the more confident the hypothesis. Keywords with a likelihood below a certain pruning threshold are rejected, and not used for further computation. In our experiments, we have rejected keywords with log-likelihood below 3 (this threshold value is empirically determined and yields a good equal error rate across many KWS experiments).

Maximum entropy is used to compute the sentiment detection likelihood:

$$p(y_j|x_k) = \left\{ \frac{1}{Z_\lambda(x)} \sum_{i=0}^N \exp(\lambda i j f_i(x_k, y_j)) \right\} \quad (13)$$

The function $f_i(x_k, y_j)$ in this equation assumes binary values of 0 and 1 either the absence or presence of evidence (i.e., sentiment bearing feature), respectively. Unlike text, using KWS for detecting sentiment bearing terms in audio produces the likelihood of the sentiment bearing feature (a probability number that is between 0 and 1). In order to incorporate keyword detection confidence into the sentiment detection likelihood computation, we propose to set $f_i(x_k, y_j)$ equal to the exponent of the log-likelihood. In this manner, $f_i(x_k, y_j)$ can now assume values between 0 and 1, and it represents the confidence of detecting the sentiment bearing feature. The benefit of computing $f_i(x_k, y_j)$ in this manner is that features that are detected with higher confidence have a greater say in determining the overall sentiment detection outcome. In other words, let c_k be the confidence score of k th detected sentiment bearing feature. Now, $f_i(x_k, y_j)$ is computed as:

$$f_i(x_k, y_j) = c_k, \text{ where } 0 \leq c_k \leq 1 \quad (14)$$

Using the above equation, we compute a new value of the maximum entropy (ME) sentiment detection likelihood which now incorporates keyword n confidence into scoring.

The star ratings for all reviews in the six sentiment corpora (noted above) were normalized by first transforming all ratings to a 5-point scale. Subsequently, ratings of 3 are discarded, ratings below and equal to 2 are labeled negative, while ratings above and equal to 4 are rated positive. The total dataset contains approximately 53 million reviews, 43 million positive, and 10 million negative. This dataset is used to create balanced training and evaluation datasets for our experiments. In particular, 8 million reviews (4 million positive and 4 million negative reviews) are used for training, 1.1 million reviews for evaluation (550 K positive and 550 K negative), and 2 million reviews for development (1 million positive and 1 million negative). All datasets are chosen randomly and have no overlap.

Unlike text sentiment detection which has numerous data sources, audio corpora resources dedicated specifically for sentiment detection task are very limited [79]. In the proposed work we have developed databases in a very controlled manner keeping audio sentiment detection as the main focus. In this context, we have developed two new audio datasets to support our sentiment detection experiments.

YouTube videos [80] are an ideal source to collect speech data where speakers follow a natural and spontaneous speaking style while sharing their opinion on a wide variety of topics. We have collected a set of 85 videos (44 positive and 41 negative) that contain people speaking spontaneously. These videos cover a wide range of topics including product reviews, movies, social issues and political opinions. The dataset contains a balanced speaker distribution. It has 55 male speakers, 30 female speakers, and 5 videos with multiple speakers. The average duration of the videos is about 5 minutes with individual video durations ranging from 2 to 15 minutes. The total duration of the evaluation dataset is about 7.5 hours. The audio quality, recording equipment, channel characteristics, and accents/dialects vary across videos.

In order to establish the sentiment ground truth on the YouTube videos, three listeners evaluated the videos and rated them for positive and negative sentiment. The listeners were asked to judge the videos for positive or negative sentiment.

The evaluators were asked to watch the video and listen to the speaker while wearing headsets. They were asked to set volume to a comfortable level and the experiment was conducted in an office environment. Each evaluator judged the sentiment of the speaker in each video on a 5-point scale, where, 1, 2, 3, 4, and 5 meant extremely negative, negative, neutral, positive, and extremely positive, respectively. In order to determine the final ground truth, the ratings of the three evaluators were averaged to obtain one rating per video. Finally, average ratings of 1 and 2 were considered negative, 4 and 5 positive, and 3 neutral. These decisions were used as ground truth for the experiments presented in this study.

The YouTube dataset mentioned in the previous section has some drawbacks. It is hard to control factors such as age, gender and accent (and these are relevant variants towards measuring performance of a speech based system). Additionally, the sentiment expressed in YouTube videos is less likely to be subtle (i.e., speakers tended to take extreme positions). Finally, it is hard to get a diverse set of opinions on the same topic (i.e., we find it hard to control for topic as a variable in the study).

In order to overcome the above mentioned issues, we have developed the UT-Dallas Sentiment (UT-Opinion) Audio Archive. In this corpus, each participant was asked 10 questions to illicit opinion. The questions are:

- 1) Are you aware of our college Mascot TEMOC? Do you like it?
- 2) Which is the last movie that you saw? Did you like it?
- 3) Do you remember your last vacation? How was your experience?
- 4) Are you satisfied with the library and other campus facilities?
- 5) Are you satisfied with the university's bus/transport facilities?
- 6) Do you like living in Dallas?
- 7) What is your opinion about Google Glass?
- 8) Do you remember the last time you went for a fancy restaurant dinner? What was your experience?
- 9) Recall the last course you completed. Did you like the class?
- 10) Would you like a dog as your pet?

The questions were asked in an interview style, with the interviewer and participant facing each other. The responses were recorded using a Samsung hand-held tablet. The distance between the participant and the tablet microphone was between 1-to-2.5 ft. After that participant answered the question subjectively, they were also asked to rate their opinion on a scale of 1-to-5 (where, 1, 2, 3, 4, and 5: meant extremely negative, negative, neutral, positive, and extremely positive, respectively). This data was collected in various acoustic settings inside the campus, namely, classrooms, hallways, offices, library, gym and street. Altogether, 63 people participated in the study, 43 male and 20 female. Students, staff and faculty participated in the data collection. The ages varied from 19 to 62 years old. The corpus consists of a diverse speaker set including native speakers of American English as well as Arabic, Spanish, Farsi, Mandarin, Korean, Hindi, Tamil, Kannada, Marathi, Bengali, Telugu, Italian and French.

Evaluators listened to the audio data and generated the sentiment ground truth. From the tabulated results, the Inter-Rater agreement for YouTube database is 94.5% and 80.1% for UT-Opinion dataset. It can be observed that when compared to YouTube dataset UT-Opinion has a lower agreement rate. This shows that the sentiment detection for UT-Opinion database is relatively not as easy when compared to YouTube database for human annotators. This is an indication that the sentiment detection for UT-Opinion dataset could also be a challenging task for machines.

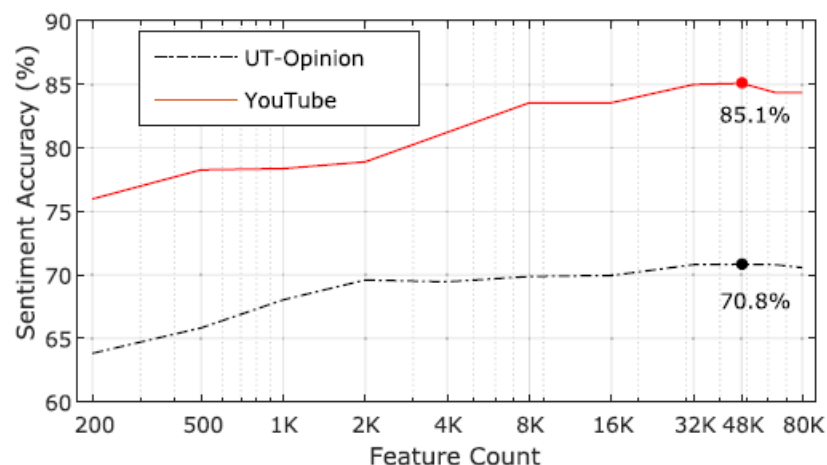


Figure 17 Accuracy distribution curve for proposed hybrid ME-KWS audio sentiment system. This gives information regarding the most optimal keyword list to be used.

In the first audio experiment, we try to determine the best keyword list for audio based sentiment detection. By using all keyword lists extracted we evaluate sentiment detection accuracy for both YouTube and UT-Opinion datasets against each keyword list. Figure 17 shows sentiment accuracy vs. keyword list size. It shows the impact of choosing larger sized keyword sets in order to improve the overall system accuracy. The intention is to show the saturating behavior of accuracy as we choose more and more keywords. It is observed that accuracy for both YouTube and UT-Opinion first increases as we increase the keyword list size from 200 to 32 K. After 32 K, the accuracy seems to level off before eventually decreasing. Since the smallest keyword lists have the most effective sentiment bearing features, the accuracy initially rises quickly.

As the keyword size grows, the new keywords are relatively less effective, and therefore accuracy does not rise as rapidly. In KWS systems, it is commonly understood that increasing the size of the keyword list can potentially increase the number of false alarms (i.e., the system falsely detects a keyword that is not actually present in the audio). In other words, bigger lists can be less accurate leading to high false alarms. From the trend in Figure 17, it appears that incorrectly detected features (due to errors in KWS) in the biggest keyword lists overwhelm the system resulting in a loss of accuracy. It is possible that a high number of false alarms contribute towards this outcome. Since the 48K keyword set gives the best performance, we choose this keyword set for all remaining experiments.

Table 27 shows the EER of the baseline ASR transcript based sentiment system and the proposed Hybrid ME-KWS based sentiment detection systems. It can be observed that the proposed method outperforms the baseline system on both YouTube and UT-Opinion corpora. The proposed system yields a +4% absolute improvement on the YouTube dataset (19% relative reduction in error) and

Table 27 results for the proposed hybrid ME-KWS and ASR

TRANSCRIPT+ME system		
Dataset	Method	
	ASR Transcripts with Max Entropy	Proposed Hybrid ME+KWS (Total Error Reduction)
YouTube	18.41%	14.9% (19%)
UT-Opinion	31.8%	29.2% (8.2%)

a +2.6% absolute improvement on UT-Opinion dataset (8.2% relative reduction in error). Additionally, the performance of both ASR and KWS systems are better on the YouTube dataset when compared to UT-Opinion. The results show that the proposed KWS method for sentiment detection outperforms the baseline ASR method for both corpora. The results also show that sentiment detection task on UT-Opinion is more difficult, perhaps owing to wider variations in accents, noise, and subtly in sentiment expression. Another main reason is that UT-Opinion audio files are shorter in duration and captured in field interview scenarios. Since YouTube files are longer, there is more evidence for consistent sentiment detection, making the task relatively easier. In YouTube, in general, people express stronger opinion than the UT-opinion corpus (we note that the topics in UT-opinion are generally less controversial than many examples in YouTube), which also makes the problem relatively easier.

Table 28 Performance across various parameters in ut-opinion database

Variable	Type	Person Count	Sentiment Detection EER	
			ASR Transcript	Proposed Hybrid ME-KWS
Accent	Native	40	25.8%	23.5%
	Non Native	23	38.1%	34.9%
Environment	Office	41	20.6%	19.1%
	Hallway + Library + Gym	12	43.0%	39.3%

Table 28 gives a detailed distribution of performance across various parameters of the developed UT-Opinion database. The performance is discussed in terms of accent and environment. It can be observed that the performance is improved in both the cases by using the proposed method. There is a very good improvement in detecting non-native speaker's sentiment using ME-KWS system. Hallway and Gym are very noisy places. The performance in such environments have also increased by the using the proposed method. Human inter-rater agreement being relatively low for UT-Opinion is further proof that the sentiment detection for this data is a challenge.

But, when compared to YouTube audio database and text based systems the performance is almost less by 15%. Performance on text corpora are high because the data is rich in opinion. In text based reviews, a case where a reviewer doesn't have a strong opinion are seldom less. Reviews with strong sentiment based words are easier to classify. Similarly, in YouTube database the opinions of the expressed by particularly strong and detailed. Hence the performance of the YouTube is relatively comparable to text based systems given some issues introduced by either speech recognition or speech signal related issues. But in UT-opinion the question are more natural and things that people come across in their everyday life.

5 CONCLUSION

Under the Advancements in Robust Speech Processing effort the University of Texas at Dallas', Center for Robust Speech Systems has researched, developed and delivered novel speech processing systems to address a wide variety of real world speech processing issues.

In developing robust solutions to speech processing systems a wide variety of approaches have been employed, applying current state of the art approaches to address issues with data mismatch for speaker and language identification, distortion, longitudinal speaker identification, diarization, robust speech recognition, distortion assessment and mitigation and sentiment classification. All algorithms have been rigorously evaluated and documented, with all corresponding documentation and publications delivered to AFRL.

All relevant algorithms and software have been delivered throughout the course of this effort via the MS-ACID software tool box and standalone software deliveries as required. Software deliveries have been implemented and documented to enable the fast and efficient transition of the tools to AFRL software baselines.

6 References

- [1] U. Yapanel and J. Hansen, "A New Perceptually Motivated MVDR-Based Acoustic Front-End (PMVDR) for Robust Automatic Speech Recognition," *Speech Communications*, vol. 50, pp. 142-152, 2008.
- [2] G. Zhou, J. Hansen and 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.
- [3] A. Ziaei, A. Sangwan and J. Hansen, "A Speech System for Estimating Daily Word Counts," in *Interspeech*, Singapore, 2014.
- [4] T. Hasan, O. Sadjadi, G. Liu, N. Shokouhi, H. Boril and J. Hansen, "CRSS Systems for 2012 NIST Speaker Recognition Evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
- [5] G. Liu and J. Hansen, "An Investigation into Back-End Advancements for Speaker Recognition in Multi-Session and Noisy Enrollment Scenarios," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1978-1992, 2014.
- [6] T. Hasan, R. Saeidi, J. Hansen and D. Van Leeuwen, "Duration Mismatch Compensation for i-Vector based Speaker Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
- [7] F. Kelly, A. Drygailo and N. Harte, "Speaker verification in score-ageing-quality classification space," *Computer Speech and Language*, vol. 27, pp. 1068-1084, 2013.
- [8] K. Kryszczuk and A. Drygailo, "Improving biometric verification with class-independent quality information," *Signal Processing IET*, vol. 3, pp. 310-312, 2009.
- [9] D. Leeuwen and N. Brummer, "An INtroduction to Application-Independent Evaluation of Speech Recognition Systems," *Speaker Classification*, vol. I, pp. 310-321, 2007.
- [10] M. I. Mandasari, R. Saeidi, M. McLaren and D. A. Leeuwen, "Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 11, pp. 2425-2438, 2013.
- [11] H. Boril and J. Hansen, "Reduced Complexity Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments," in *ISCA INTERSPEECH-2009*, Brighton, 2009.
- [12] J. Hansen and V. Varadarajan, "Analysis and Normalization of Lombard Speech under different types and levels of noise with application to In-Set/Out-of-Set Speaker

- Recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 2, pp. 366-378, 2009.
- [13] V. Varadarajan, J. Hansen and A. Ikeno, "UT-Scope: A Corpus for Speech under Cognitive/Physical task stress and Emotional," in *ELRA- LREC-2006: Lang. Resources & Eval. Conf.*, Genoa, 2006.
 - [14] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, pp. 12-40, 2010.
 - [15] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *ISCA ODYSSEY*, 2012.
 - [16] S. Ranjan, A. Misra and J. Hansen, "Curriculum learning based probabilistic linear discriminant analysis for noise robust speaker recognition," in *Proc. ISCA INTERSPEECH*, 2017.
 - [17] L. Ferrer, "Promoting robustness for speaker modeling in the community: The PRISM evaluation set," in *Proc. NIST Speaker Recognit.*, 2011.
 - [18] NIST, "SRE 2010," 2010. [Online]. Available: <https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2010>. [Accessed 17 April 2017].
 - [19] Y. Lei, L. Burget, M. Ferrer, M. Graciarena and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2012.
 - [20] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-vectors: Robust dnn embedding for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.
 - [21] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng and H. Li, "Unsupervised domain adaption via domain adversarial training for speaker recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2018.
 - [22] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096-2030, 2016.
 - [23] F. Bahmaninezhad and J. Hansen, "Generalized discriminant analysis (GDA) for improved i-vector based speaker recognition," in *INTERSPEECH*, 2016.
 - [24] S. Gu, Y. Tan and X. He, "Discriminant Analysis via support vectors," *Neurocomputing*, vol. 73, no. 10, pp. 1669-1675, 2010.

- [25] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788-798, 2011.
- [26] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka and N. Brummer, "Discriminatively trained probabilistic linear discriminant," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011.
- [27] S. Sadjadi and J. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197-200, 2013.
- [28] D. Leeuwen, "A note on performance metrics for speaker recognition using multiple conditions in an evaluation," *Tech. Rep.*, 9 June 2008.
- [29] F. Kelly, N. Harte and D. Leeuwen, "Effect of longterm ageing on i-vector speaker verification," in *Interspeech 2014*, Singapore, 2014.
- [30] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey*, Brno, 2010.
- [31] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, Florence, 2013.
- [32] S. Pigeon, P. Druyts and P. Verlinde, "Applying Logistic regression on fusion of the NIST '99 1-speaker submission," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 237-248, 2000.
- [33] D. A. van Leeuwen and N. Brummer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification I ser. Lecture notes in Comp. Sci.*, New York, Springer, 2007, pp. 330-353.
- [34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, pp. 2672-2680, 2014.
- [35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, "Improved techniques for training gans," *Advances in Neural Information Processing Systems*, pp. 2234-2242, 2016.
- [36] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep Neural networks with multitask learning," in *ICML*, 2008.
- [37] A. Ali, S. Vogel and S. Renals, "Speech Recognition challenge in the wild: Arabic MGB-3," in *IEEE ASRU*.

- [38] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskeyer and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.
- [39] C. Yu, C. Zhang, S. Ranjan, Q. Zhang, A. Misra, F. Kelly and J. H. Hansen, "UTD-CRSS system for the NIST 2015 language recognition i-vector machine learning challenge," in *IEEE ICASSP*, 2016.
- [40] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," *IEEE SLT*, 2018.
- [41] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett and N. Dahlgren, *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus*, 1993.
- [42] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Thirteen International Conference on Artificial Intelligence and Statistics*, 2010.
- [43] H. Dubey, L. Kaushik, A. Sangwan and J. Hansen , "A speaker diarization systems for studying peer-led team learning groups," in *ISCA INTERSPEECH*, 2016.
- [44] J. Hansen, H. Dubey, A. Sangwan, L. Kaushik and V. Kothapally, "UTDallas-PLTL: Advancing multistream speech processing for interaction assessment in peer-led team learning," *The Journal of the Acoustical Society of America*, vol. 143, no. 3, pp. 1869-1869, 2018.
- [45] J. Hansen, H. Dubey and A. Sangwan, "CRSSLDNN:Long-duration naturalistic noise corpus containing multi-layer noise recordings for robust speech processing," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1797-1797, 2018.
- [46] A. Graves, N. Jaitly and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.
- [47] H. Sak, A. Senior and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," arXiv, 2014.
- [48] A. Graves, G. Hinton and M. Abdelrahman, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
- [49] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Cantanzaro, Q. Cheng and G. Chen, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016.

- [50] M. Elfeky, M. Bastani, X. Velez, P. Moreno and A. Waters, "Towards acoustic model unification across dialects," in *IEEE Spoken Language Technology Workshop*, 2016.
- [51] S. Mirsamadi and J. Hansen, "On multi-domain training and adaptation of end-to-end rnn acoustic models for distant speech recognition," in *Interspeech*, 2017.
- [52] T. N. Sainath, R. Prabhavalkar, S. Kumar, S. Lee, A. Kannan, D. Rybach, V. Schogol, P. Nguyen, B. Li, Y. Wu and et. al., "No need for a lexicon? Evaluating the value of the pronunciation lexica in end-to-end models," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2017.
- [53] R. Giri, M. L. Seltzer, J. Droppo and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *International Conference on Acoustics Speech and Signal Processing*, 2015.
- [54] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui and B. Ramabhadran, "Efficient knowledge distillation from an ensemble of teachers," in *Interspeech*, 2017.
- [55] G. Hinton, O. Vinyals and J. Dean, "Distilling the knowledge in a neural network," *stat*, vol. 1050, p. 9, 2015.
- [56] S. Watanabe, T. Hori, J. Le Roux and J. R. Hershey, "Student-teacher network learning with enhanced features," in *International Conference on Acoustics Speech and Signal Processing*, 2017.
- [57] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom and A. Rosenberg, "Knowledge distillation across ensembles of multilingual models for low-resource languages," in *International Conference on Acoustics Speech and Signal Processing*, 2017.
- [58] R. Hsiao, T. Ng, L. Zhang, S. Ranjan, S. Tsakalidis, L. Nguyen and R. Schwartz, "Improving semi-supervised deep neural network for keyword search in low resource languages," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [59] T. Asami, R. Masamura, Y. Yamaguchi, H. Masataki and Y. Aono, "Domain adaptation of dnn acoustic models using knowledge distillation," in *International Conference on Acoustics Speech and Signal Processing*, 2017.
- [60] R. Huang, J. Hansen and P. Angkititraku, "Dialect/accent classification using unrestricted audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 453-464, 2007.
- [61] H. Sak, A. Senior, K. Rao and F. Beaufays, *Fast and accurate recurrent neural network acoustic models for speech recognition*, arXiv preprint, 2015.

- [62] A. Rix, J. Beerends, M. Hollier and A. Hekstra, "Perceptual evaluation of speech quality (pesq) - a new method for speech quality assessment of telephone networks and codecs," in *International Conference on Acoustics Speech and Signal Processing*, 2001.
- [63] R. Stern and Kim, "Robust signal-to-noise ratio estimation distribution analysis," in *Interspeech*, 2008.
- [64] E. Vincent, R. Gribonval and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on audio Speech and Language Processing*, vol. 14, no. 4, pp. 1462-1469, 2006.
- [65] P. Kenny, G. Boulianne, P. Ouellet and p. Dumouchel , "Joint factor analysis versus eigenchannels in speaker recognition".
- [66] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget and J. Cernocky, "Full-covariance ubm and heavy-tailed plda in i-vector speaker verification," in *International Conference on Acoustics Speech and Signal Processing*, 2011.
- [67] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *International Conference on Computer Vision*, 2007.
- [68] B. Liu, *Sentiment Analysis and Opinion Mining*, Williston, VT: Morgan & Claypool publishers, 2012.
- [69] D. Povey, A. Ghoshal, G. Boulianne, L. Burget , O. Glembek , N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Big Island, Hawaii, 2011.
- [70] J. Godfrey, E. Holliman and J. McDaniel, "SWITCHBOARD: Tele- phone speech corpus for research and development," in *Int. Conf. Acoust., Speech, Signal Process.*, 1992.
- [71] C. David, D. Miller and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text," in *Proc. 4th Int. Conf. Lang. Resources Eval.*, 2004.
- [72] "Web Data Collection," 2011. [Online]. Available: http://ssli.ee.washington.edu/ssli/projects/ears/WebData/web_data_collection.html.
- [73] "Sentiment Analysis," [Online]. Available: www.cs.uic.edu/liub/FBS/sentiment-analysis.html.
- [74] J. McAuley, C. Targett, Q. Shi and A. Hengel, "Image-based recommendations on styles and substitutes," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2015.
- [75] J. McAuley, R. Pandey and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proc. 21th ACM SIGKDD Int.*, 2015.

- [76] M. Ganapathibhotla and B. Liu, "Mining opinions in comparative sentences," in *Proc. 22nd Int. Conf. Comput. Linguistics*, 2008.
- [77] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2006.
- [78] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics*, 2005.
- [79] G. McKeown, M. Valstar, R. Cowie, M. Pantic and M. Schroder, "The SEMAINE database: Annotated multi-modal records of emotionally coloured conversations between a person and a limited agent," *IEEE Transactions Affect. Comput.*, vol. 3, no. 1, pp. 5-17, 2011.
- [80] "Youtube," [Online]. Available: www.youtube.com.
- [81] X. Fan and J. Hansen, "Speaker Identification within Whispered Speech Audio Streams," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1408-1421, 2011.
- [82] X. Fan and J. 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, 2013.
- [83] J. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans, Audio, Speech and Language Proc.*, vol. 17, no. 2, pp. 336-378, 2009.
- [84] A. Graves, A. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks".
- [85] T. Asmai, R. Masumura, Y. Yamagushi, H. Masataki and Y. Aono, "domain adaptation of dnn acousti models using knowledge distillation," in *IEEE International Conference on Acoustics Speech and Signal Processing*, 2017.

7 ACRONYMS

A/D	Analog to Digital
ACC	Accuracy
ACL	Anti-curriculum Learning
ADDA	Adversarial Discriminative Domain Adaptation
ADI	Arabic Dialect Identification
AFRL	Air Force Research Laboratory
AGC	Automatic Gain Control
AMI	Augmented Multi-Party Interaction
ASD	Absolute Session Difference
ASR	Automatic Speech Recognition
BNF	Bottleneck Feature
CE	Cross-Entropy
CER	Character Error Rate
CG	Cross Gender
CL	Curriculum Learning
C_{ltr}	Log Likelihood Ratio Calibration Cost
ClipDaT	Clipping Detection and Tagging
CNN	Convolutional Neural Network
CSO	Character Spikes Overlap
CTC	Connectionist Temporal Classification
CV	Cross Validation

DANN	Domain Adversarial Neural Network
DARPA	Defense Advanced Research Projects Agency
DAT	Domain Adversarial Training
dB	Decibel
DER	Diarization Error Rate
DIARNAT	Audio Diarization in Naturalistic Environments
DID	Dialect ID
DNN	Deep Neural Network
EER	Equal Error Rate
EGY	Egyptian
EM	Expectation Maximization
FA	Factor Analysis
FA	False Accept
FST	Finite State Transducer
GAN	Generative Adversarial Network
GLF	Gulf
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
IE	i-Vector Extractor
KL	Kullback-Leibler
KWS	Keyword Spotting
LAV	Levantine
LE	Lombard Effect

LID	Language ID
LR	Likelihood Ratios
LRE	Language Recognition Evaluation
LSTM	Long Short-Term Memory
MA	Multi-Accent
MARP	Multi-Session Audio Research Project
ME	Maximum Entropy
MFCC	Mel-Filtered Cepstral Coefficients
MGB-3	Multi-Genre Broadcast – 3
minDCF	Minimum Decision Cost Function
MLLR	Maximum Likelihood Linear Regression
MSA	Modern Standard Arabic
N+S-MOD-KA	Noise & Speaker Modeling For Knowledge Assessment
NIST	National Institute of Standards
NLP	Natural Language Processing
NOR	Maghrebi
PCA	Principal Component Analysis
PCM	Pulse Code Modulation
PCN	Phone Confusion Network
PESQ	Perceptual Evaluation Of Speech Quality
PLDA	Probabilistic Linear Discriminant Analysis
PLL	Prof. Life Log
PLP	Perceptual Linear Prediction

PLTL	Peer Led Team Learning
PMVDR	Perceptually Motivated Minimum Variance Distortionless Response
PRC	Precision
QMF	Quality Measure Functions
RATS	Robust Automated Transcription Of Speech
RCL	Recall
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SAR	Source-to-Artifacts Ratio
SDC	Shifted Delta Cepstrum
SID	Speaker Identification
SNR	Signal-to-noise Ratio
SRE	Speaker Recognition Evaluation
STNR	Speech-To-Noise Ratio
SV	Speaker Variability
SVDA	Support Vector Discriminant Analysis
SVM	Support Vector Machine
TDNN	Time-Delay Deep Neural Network
TEO	Teager Energy Operator
TV	Total Variability
UBM	Universal Background Model
UBNF	Unsupervised Bottleneck Features
USAF	United States Air Force

UTD-CRSS	University of Texas at Dallas Center For Robust Speech Systems
VAD	Voice Activity Detection
VTL	Vanilla Transfer Learning
WADA	Waveform Amplitude Distribution Analysis
WER	Word Error Rate