AFRL-SR-AR-TR-04-

0639

REPORT DOCUMENTATION PAGE

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	15 Dec 2004	FINAL - 1 SEP 2003 - 30 AUG 2004

4. TITLE AND SUBTITLE
Separation of Speech from Background

5. FUNDING NUMBERS

6. AUTHOR(S)
Dr. John Josephson
James Russell

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Aetion Technologies, LLC
1275 Kinnear Road
Columbus OH 43212

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NL
4015 Wilson Blvd., Rm 713
Arlington VA 22203-1954

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
Ff49620-03-C-0054

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release: Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Phase I has resulted in advances in computational methods, assessments of current strengths and weakness, and assessments of the range of potential applications and their performance requirements. The progress and results from Phase I show that the technology is very promising. However, it has also become clear that the problem is quite challenging, and significant technical weaknesses remain. Phase II would aim to overcome the most significant weaknesses preventing practical levels of performance. These are considered to be: inadequate performance on unvoiced speech materials, weak processing algorithms that mandate large amounts of computation and impose latencies that are too long for near-real-time applications.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Voice, Speech recognition, Binaural processing	12
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASS	UNCLASS	UNCLASS	

"Separation of Speech from Background"

December 15, 2004

SBIR/STTR Program

Issued by Air Force Research Laboratory Under

Contract No. F4962-03-C-0054

Aetion Technologies, LLC
Dr. John Josephson, James Russell
1275 Kinnear Rd
Columbus OH 43212
614 340 1835

Effective Date of Contract: September 01, 2003
Short Title of Work: Final Technical Report
Reporting Period: September 01, 2003 – August 30, 2004

DISCLAIMER

"The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either express or implied, of the Air Force Research Laboratory or the U.S. Government.

1. Summary

Human ability to attend to a single voice in the presence of background interference is remarkable. If this could be imitated in practical technology it would be of great benefit for automatic speech recognition and other applications. Researchers at the Ohio State University have demonstrated computational methods for separating speech from interfering sounds that imitate human auditory processing, and that have achieved levels of performance clearly evident to the untrained ear. Aetion Technologies has partnered with Ohio State to carry this research to commercial application.
Although humans make strong use of binaural processing in the "cocktail party effect" – separating a single voice from interfering background voices - human abilities in monaural separation are remarkable in themselves. While the Ohio State researchers have made progress on imitating human binaural speech separation, their work on monaural separation was more mature. Moreover the immediate market for monaural separation is much larger than that for binaural or n-aural separation, since single-microphone sound sources are ubiquitous (e.g., telephones), while multiple-microphone sources are much less so.

Phase I has resulted in advances in computational methods, assessments of current strengths and weakness, and assessments of the range of potential applications and their performance requirements. The progress and results from Phase I show that the technology is very promising. However, it has also become clear that the problem is quite challenging, and significant technical weaknesses remain. Phase II would aim to overcome the most significant weaknesses preventing practical levels of performance. These are considered to be: inadequate performance on unvoiced speech materials, weak performance on some voiced materials (especially with background voices), and processing algorithms that mandate large amounts of computation and impose latencies that are too long for near-real-time applications.

2. Background: Abductive Inference

Abductive inference - inference to the best explanation - is a distinctive and recognizable pattern of evidential reasoning (Josephson & Josephson, 1994, 96). It is ubiquitous at or near the surface of typical arguments offered in science, intelligence analysis, and ordinary life, and may be considered part of commonsense logic. Abductive arguments are fallible, but there are only a small number of ways in which they can go wrong.

An abductive inference has a pattern that can be described as follows:

> *D* is a collection of data (facts, observations, givens).
> Hypothesis *H* explains *D* (would, if true, explain *D*).
> No other hypothesis explains *D* as well as *H* does.
> Therefore, *H* is probably correct.

Note that the conclusion is justified, not simply as a possible explanation, but as the best explanation in contrast with alternatives. The strength of the conclusion *H*, the force of the probably in the conclusion statement, reasonably depends on the following considerations:

- how decisively *H* surpasses the alternatives,
- how good *H* is by itself, independently of considering the alternatives,
- how thorough was the search for alternative explanations.

Besides the judgment of likelihood, willingness to accept the conclusion also reasonably depends on pragmatic considerations, including:

- how strong is the need is to come to a conclusion at all, especially considering the possibility of gathering further evidence before deciding,
- the costs of being wrong and the benefits of being right.

A hypothesis that leads to false or highly inaccurate predictions is poor by itself, and should not be accepted, even if it appears to be the best explanation when considering all the available data. Failure of predictions counts as evidence against a hypothesis and so tend to improve the chances of other hypotheses emerging as best. Failure of predictions may improve the margin of decisiveness by which the best explanation surpasses the failing alternatives. Thus, abductive inferences are capable of turning negative evidence against some hypotheses, into positive evidence for alternative hypotheses.

John R. Josephson and Susan G. Josephson, ed., (1994, 1996) *Abductive Inference: Computation, Philosophy, Technology*, Cambridge University Press.

3. Project Narrative and Results

Prof. DeLiang Wang and his students at Ohio State have assisted Aetion in reproducing key experiments that have been conducted by Prof. Wang with his colleagues and students in the area of separation of speech from interference, resulting in a usable software suitable for performance testing over a range of sampled materials. Aetion has studied this software to determine improvements needed for practical applications. Ideal processing would occur at real-time speeds, and with latency well under a second; these are presumably the operational demands of the broadest market, which includes improved intelligibility of speech in real-time communications, e.g., for mobile phones. The operational demands are nearly as stringent for interactive speech recognition applications, which also incur the delay of recognition. Aetion has also analyzed the research software from the perspective of *abductive inference* and assessed whether it would be advantageous to re-engineer the software based on the concepts or software for information fusion currently under development by Aetion under other sponsorship.

Phase I sought to determine the feasibility of practical speech separation based on the work of Ohio State researchers.

Phase I work began with the method of Hu and Wang [HuWang04], which is effective for separating speech from background for voiced speech, but which did not address the problem of separation for unvoiced segments such as unvoiced stops and fricatives. During Phase I, OSU (Hu and Wang) succeeded in developing methods that leverage the previous ability to segregate voiced intervals, using the results to constrain the possible intervals of unvoiced speech. The new methods recognize unvoiced speech in the remaining intervals by examining time-frequency regions created by detecting acoustic onsets, by pairing acoustic onsets with offsets to locate "acoustic events", and making use of spectral profile and segment duration to categorize acoustic events as speech (stop, unvoiced fricative, or unvoiced affricate), or as interference. This extends the range of speech materials addressed by the available methods to the full range of natural speech (excluding whispering, and other special cases). This is a significant step toward practicality, but the effectiveness of the methods remains inadequate for most applications. However, it is plausible that these methods are already substantially good enough for preprocessing of speech recorded in noisy environments for off-line automatic speech recognition, which is the least demanding of the envisioned practical applications.

To enable initial experiments with mixed voiced/unvoiced speech, Aetion modified the program codes upon which the earlier results of Hu and Wang were based. A decision was made to experiment with a relatively naive extension of the Hu-Wang algorithm to determine the minimum requirements for producing acceptable results, and to establish a base line of performance. It was hypothesized that a system which simply applied the Hu-Wang algorithm to each of the continuous segments of voiced speech in a sample would produce reasonable results. This conjecture was based on the observation that not only is the majority of continuous speech voiced, but much of the information about certain voiceless consonants, in particular unvoiced stops, is contained in the voiced regions rather than in the unvoiced regions [RabinerJuang93]. Thus, the bulk of the information lost by considering only voiced portions of speech is presumably in voiceless fricatives and affricates. So a system was constructed in which the original, noisy, signal was gated through to the resulting signal during unvoiced segments. This method, while producing results far from optimal, appears to be quite acceptable for consumption by human listeners, who can often readily reconstruct masked information with fairly high fidelity.

Machine listeners, however, are generally not capable of this kind of reconstruction of masked information, and will likely require better treatment of unvoiced segments. It was apparent that the baseline system performs very well at finding the voiced segments in continuous speech, even in the face of very high levels of broadband noise. However, as noise levels increase, the spectral distortion of the segregated speech signal also increases, either from the inclusion of energy from overlapping noise, or the exclusion of portions of the spectrum of the original speech signal because noise dominates in those regions. These behaviors, both positive and negative, of the baseline system are essentially carried over unchanged from the predecessor system. The Hu-Wang system is

just as good at finding a voiced segment amidst noise, and suffers just as much due to broadband noise.

To create an extended version of the Hu-Wang system which can process mixed voiced/unvoiced speech, it was necessary to develop an iterative version of the algorithm which could identify a series of non-overlapping voiced segments. Each of these segments would then receive further processing as detailed in [HuWang04]. In particular, we modified the Initial Grouping algorithm (Section IV-B) and the Pitch Tracking algorithm (Section V-A), while the rest of the method remained completely unchanged. The necessary changes to the Hu-Wang method closely follow the original algorithms and were significantly more simple than we had originally estimated, and suggest how general and flexible the original method is.

Our Initial Grouping algorithm first finds the longest segment, or time-frequency region, in the sample, subject to certain constraints, and classifies it as foreground (speech-dominant) or background (noise-dominant) according to the original algorithm. If the longest segment is classified as speech then all segments which temporally overlap the longest segment in the majority of their frames are classified as speech or background depending on whether they agree with the longest segment for the majority of their overlapping frames. One classified, these segments are removed from further consideration by later iterations of the algorithm. If the longest segment is classified as background, then overlapping segments are not classified at this point. In any case, the algorithm then repeats, using the next longest unclassified segment.

The classification of the longest segment, which normally contains F_0, is subject to certain simple-minded plausibility constraints which heuristically improve the classification of segments. While phonetically motivated, these constraints are tentative and subject to improvement. The first constraint is that some portion of the segment must be less than about 500Hz, which is the upper limit of the range for F_0 plausibility. The second constraint is a minimum duration for the longest segment, since voiced segments of speech are rarely very short. The duration used for testing the system was 100ms, which is certainly too long (50ms is more reasonable), but worked better than the shorter duration which included too many spurious segments. Additional heuristics are probably necessary.

The Pitch Tracking algorithm was also modified to cope with the multiple voiced speech portions output by the grouping algorithm. While the Initial Grouping algorithm, described above, has been modified to operate iteratively over the input, the modified Pitch Tracking algorithm operates mostly unchanged from the original algorithm, which makes several passes over the input, successively refining the F_0 estimate and interpolating over uncertain areas. The new version has been simply modified so that, during each pass, the pitch re-estimation and interpolation operates only on the identified voiced portions of the signal.

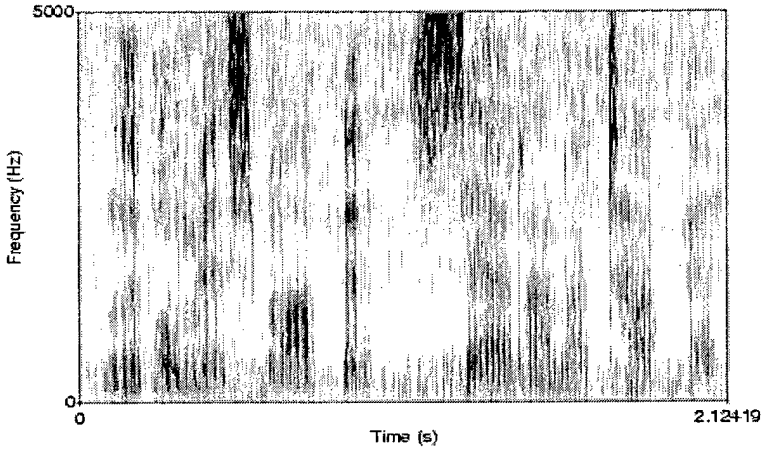

Figure X.a

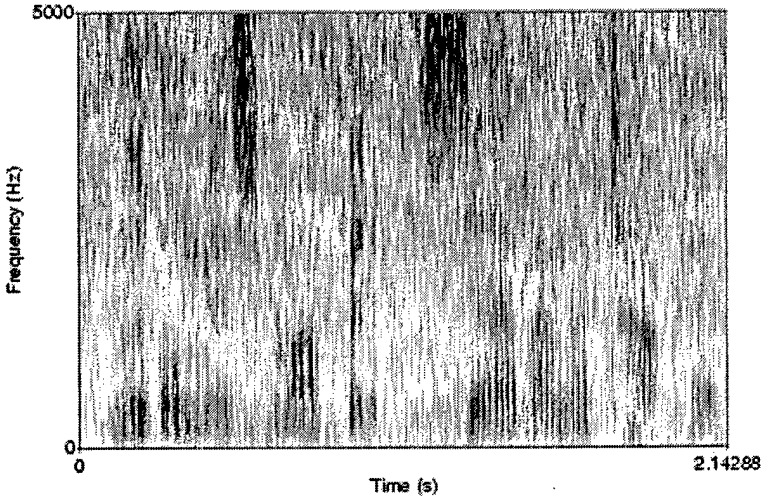

Figure X.b

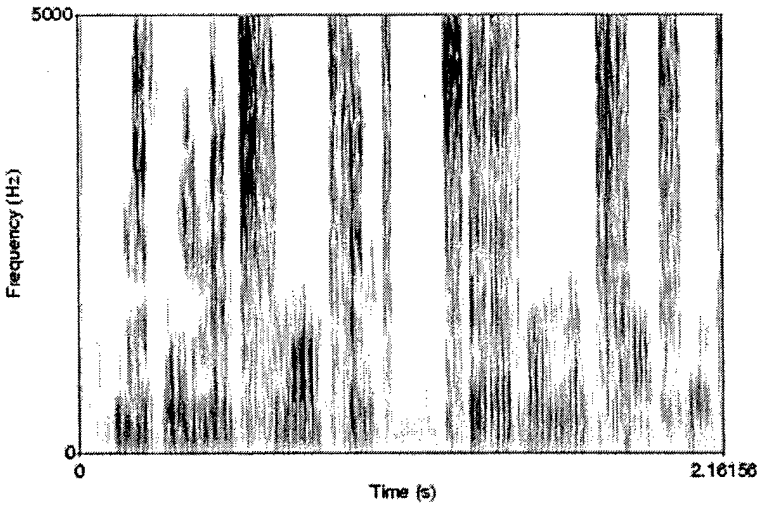

Figure X.c

The results of running the system on a sample utterance can be seen in Fig. X. Figure X.a shows a spectrogram of a recording of one of the authors saying "Little Miss Muffet sat on a tuffet." Figure X.b shows a spectrogram of the same recording mixed with broadband noise, at a SNR of –2.5dB. Finally, Figure X.c shows the result of running the speech segregation algorithm on the noisy mixture. Broadband noise was selected because it represents a near worst-case situation for the segregation algorithm. In general the result of the method becomes less satisfactory as the spectral overlap between signal and inclusion increases.

It is apparent that the method performs very well at finding the voiced segments in continuous speech, even in the face of very high levels of broadband noise. However, as noise levels increase, the spectral distortion of the segregated signal increases commensurately, either from the inclusion of energy from overlapping noise, or the exclusion of portions of the spectrum of the original signal because noise dominates in those regions. These behaviors, both positive and negative, of the extended system are essentially carried over unchanged from the underlying system. The Hu-Wang system is just as good at finding a voiced segment amidst noise, and suffers just as much due to broadband noise.

Aetion informally evaluated the intelligibility of the results produced by the baseline system in various situations, using the sound mixtures discussed in [HuWang04] as well as mixtures of their own devising. The findings were in basic agreement with other informal evaluations of similar systems, which have found that intelligibility is not significantly improved, despite improvements in S/N ratio (E.g., Ellis, 1996). In situations where the baseline system exhibits near-optimal behavior, such as intermittent

noise, low-amplitude intrusions, narrow-band intrusions, or high-frequency intrusions, the intelligibility of the pre-segregation mixture was generally not significantly impaired, leaving little scope for improvement. In hard cases, such as broadband noise or competing speech of comparable amplitude, the distortion of the resulting signal roughly compensates for the energetic or informational masking caused by the intrusion. The intelligibility gain from eliminating the latter is cancelled out by the intelligibility reduction caused by the former. On the other hand, it became relatively clear during the experimentation that in many signal/intrusion combinations the attentional demands on the listener are somewhat reduced by segregation, resulting in a corresponding reduction in listening effort. However, listening effort is much harder to measure than intelligibility.

Aetion also analyzed the computational demands of the Hu-Wang algorithms to determine their suitability for practical applications. As they were designed, the algorithms are suitable only for off-line processing. Wang and Brown had a number of intuitions about how real-time performance could be achieved [WangBrown], and our proposals for improvement are closely related to these. There are two components which constrain the Hu-Wang algorithms to off-line processing. The first is that the segmentation and grouping mechanisms operate on global properties of the utterance, or, as Aetion extended it, on properties global to each voiced segment of the utterance. Decisions about grouping are made, and in some cases, re-made, based on these global properties. The second component constraining the Hu-Wang algorithms to off-line processing is extraction of the dominant pitch, which provides an initial estimate for the fundamental frequency (F0). To determine F0, the system finds the peak of a summary correlogram, which pools a correlogram for each of the 128 channels into which the original signal is filtered. This is extremely computationally expensive, and not only imposes an inherent latency, but pushes the limits of practical near-real-time computation, especially for low-power applications. This is probably the most significant hurdle to the suggestion [WangBrown] that the basic method could be turned into a real-time system with parts implemented on an analog VLSI chip.

The previous work in monaural speech segregation, upon which Aetion's ideas build, [HuWang][WangBrown], has shown excellent results in identifying the location of speech in signals consisting of a mixture of voiced speech and a wide variety of intrusive noises, including telephone rings, white noise, music, and background speech. It also is successful in isolating and preserving the majority of that part of the signal representing voiced speech and rejecting the majority of that part of the signal which is noise. However, the current approach suffers from several problems which limit the applicability of the method in real-world situations. Aetion will suggest several means to address those shortcomings with an eye toward turning the existing methods into ones that could be deployed in the field.

The immediate motivation for these suggestions stems from the results of current experiments which suggest a wider domain of applicability than previously considered. The existing work focuses on single, short, fully voiced utterances mixed with intrusions. The assumption is made that the voiced utterance extends through the majority of the

sound sample being processed. Moreover, the method depends on global properties of the utterance, which means processing must take place off-line. This obviously limits the potential uses of the method as-is. Our experiments suggest that the method can be successfully expanded to continuous utterances consisting of alternating voiced and unvoiced segments, which brings to mind on-line, real-time processing. The segregation of unvoiced segments is the subject of ongoing research.

The specific suggestions are divided into two major categories. The first set of suggestions deal with enhancements to the existing framework which will allow speech segregation to proceed in real time with minimum latency. The second set of suggestions deal with enhancements to increase the fidelity and intelligibility of the resulting sound streams, making it a more practical method for computational auditory scene analysis (CASA).

In previous work [WangBrown] it was suggested that the basic method could be turned into a real-time system, with parts implemented on an analog VLSI chip. Despite this suggestion, no progress has been made on real-time implementations of their method. The most significant hurdle is the dependence of the algorithm on a large number of autocorrelation calculations, which are quite computationally expensive. The number of calculations currently required exceeds the limits of practical near-real-time computation. Moreover, the use of autocorrelation imposes an inherent latency, the duration of which may be unacceptable in certain applications. The most critical use of these autocorrelations comes in the calculation of the fundamental frequency (F_0).

We propose to solve this problem by designing a system of coupled oscillators which will rapidly converge to the fundamental frequency of the speech portion of the signal, or, in the absence of a strongly harmonic signal, exhibit uncorrelated behavior. We expect that this system of oscillators, like those that modulate the behavior of the cochlea, will be able to determine F_0, even if the fundamental frequency is missing and only its harmonics remain. This system of oscillators is essentially a computationally inexpensive method for determining, among all plausible fundamental frequencies, the one whose harmonic spectrum accounts for the majority of the observed frequency peaks in the input system. (That is, it implements a form of "abductive inference" or best-explanation reasoning.)

We furthermore anticipate that this system of coupled oscillators will be able to act as a pitch tracker and, in effect, maintain and continuously update a hypothesis about both the fundamental frequency and whether or not there is currently a speech signal at all. We expect the response of the system to be fast enough to quickly transition between non-speech dominant and speech dominant segments of the input.

The second hurdle to real-world use of the existing method is its performance in the face of noise which significantly overlaps spectrally with the speech signal. This is a known

shortcoming of the system [HuWang]. At frequencies below 1kHz, the current system performs extremely well, except perhaps in the case where the intrusive sound is competing speech. At frequencies above 1kHz, where much of the information in speech is found, however, the system performs much less adequately, even though the latest version represents a dramatic improvement over all predecessors. When the speech is mixed with broad-spectrum noise, more than half of the speech energy can be lost, and almost half the noise retained. Moreover, the resulting signal is highly distorted and its intelligibility is unlikely to be any better than the original, noisy signal. This phenomenon is observed with even relatively small amounts of broad-spectrum noise of the kind that would likely be observed in the field, such as instrument noise, interference, wind noise, and so on.

The current system (without the recent improvements for unvoiced speech segments from the work of Guoning Hu) achieves the level of performance that it does by looking in the high frequency bands for correlations between F_0 and the *envelope* of the signal rather than in the signal itself. It does this because the high frequency bands are wide enough to contain multiple, unresolved harmonics, and because at high frequencies, small errors in the estimate of F_0 are exaggerated, which would cause less resilient behavior. It exploits the phenomenon of beats, whereby two adjacent harmonics of a fundamental exhibit an envelope which fluctuates at the fundamental frequency.

We hypothesize that the difficulty with the current system results from the binary nature of its decision making. The system decides for every frequency band at every 10ms time slice whether or not that bin contains a signal which is speech dominant or non-speech dominant. Once that decision is made, the entire signal in that bin is either kept or rejected, resulting in either the possible retention of significant amounts of noise, or rejection of significant amounts of speech. We suggest that this coarseness is the cause of much of the distortion in the resultant signal, and furthermore we suggest that significant improvements can be made to the results by making much finer-grained decisions.

We propose a modification and extension of the current method, wherein we are not satisfied by a mere correlation of the envelope with F_0, but instead impose an expectation of the shape of that envelope, keeping the portion of the signal which matches the expected shape of the envelope and rejecting that portion which violates our expectation. We are thus making a decision which is not just a 0-1 decision but falls on the [0,1] continuum, and is not made at a granularity of 10ms, but evolves continuously in time.

We have taken this approach because it appears that the beat phenomenon is not the most productive way to view the cause of high-frequency band envelope fluctuation at F_0. The oral cavity acts as a damped resonator for the glottal pulse, and thus all harmonics of F_0 show a characteristic envelope beginning with an attack which occurs at some delay after the onset of the glottal pulse, followed by a decay with some particular time constant.

We believe that the shape of these envelopes can be roughly characterized for different frequency bands. The speech signal predominates only over predictable short intervals synchronized with the glottal pulses, and this finer temporal view of the phenomena can be exploited for better grouping of the speech signal, and thus better separation from background sounds.

Real-time constraints will limit the complexity of the calculations that can be done to approximate the envelopes, but we believe the result should be more satisfactory than the current system in any case. More sophisticated approaches can be taken in off-line processing, and we believe that the outlined approach holds even greater promise there.

A side-effect of making a continuous, rather than binary, decision about what part of the signal to keep and which to reject is that the background signal can be better reconstructed in its own right. This has obvious applicability in the case of overlapping speech signals. The system should have the ability to mask off the louder, foreground voice, and allow reconstruction of the previously incomprehensible background voice.

An additional problem for the current system, which prevents more widespread application, is how to achieve sequential grouping; that is, how to associate streams which are discontinuous in time, yet come from the same source. This is a fundamental problem in ASA which is particularly important for applications in which the sources are highly confusable, such as multiple-speaker co-channel segregation. Many approaches to this problem have been suggested, including associations by F_0 or timbre, formant tracking, speaker identification, and various statistical and model-based approaches. Unfortunately, none of the systems work particularly well in isolation, and it not obvious how to combine the systems. An abductive solution should be helpful here. Abductive inference would be used to integrate hypotheses generated from cues which are derived from the input signal by existing methods and make a decision, based on possibly ambiguous data, about which hypothesis best accounts for the observed signal.

Furthermore, we believe that the current system can be extended with a technique which would jointly model both foreground and background signals, which, unlike the current system, would generate enough information about the background remainder during the extraction of the foreground signal that a useful background signal could be extracted as well. For instance, we determined that when the current system is presented with a stimulus with competing speakers, not enough of the background speech remains after the foreground speech has been segregated to be able to understand the background speech. The human auditory system is capable of this under many conditions, so it is clearly possible to do. If the system were so equipped then multiple layers, corresponding to multiple sound sources, could be peeled off the signal like an onion.

References

Ellis, D.P.W. (1996). *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, Dept. of Elec. Eng & Comp. Sci., M.I.T., June 1996.

[CampbellEtAl04] Campbell, S., D. L. Wang, and C. Jayaprakash, "Synchronization Rates in Classes of Relaxation Oscillators", *IEEE Transactions on Neural Networks*, Vol. 15, 2004, pp.684–697, 1027-1038.

[HuWang03] Hu, G. and D. L. Wang, "Separation of stop consonants", *ICASSP 2003*, Vol. 2, pp. 749-752.

[HuWang04] Hu, G. and D. L. Wang, "Monaural Speech Segregation Based on Pitch Tracking and Amplitude", *IEEE Transactions on Neural Networks*, Vol. 15, 2004, pp. 1135–1150.

[RabinerJuang93] Rabiner, L. and B. H. Juang, *Fundamentals of Speech Recognition*, PTR Prentice-Hall, 1993, p. 33.

[WangBrown99] Wang, D. L. and G. J. Brown, "Separation of Speech from Interfering Sounds Based on Oscillatory Correlation", *IEEE Transactions on Neural Networks*, Vol. 10, 1999, pp. 684–697.