SIGNAL SEPARATION MOTIVATED BY HUMAN AUDITORY PERCEPTION: APPLICATIONS TO AUTOMATIC SPEECH RECOGNITION

Richard M. Stern
Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA

ABSTRACT

The human auditory system uses a number of well-identified cues to segregate and separate individual sound sources in a complex acoustical environment. For example, researchers in auditory scene analysis have long identified cues such as common onset, correlated fluctuations in instantaneous amplitude and frequency, harmonicity, and common interaural time and amplitude differences as ways of identifying which components of a complex signal are derived from a common source. It is widely believed that the use of these cues to achieve such "grouping" and signal separation should be very useful in improving the accuracy of automatic speech recognition in very difficult environments such as competing speech, background music, and transient noise, and this has been a goal of several research groups in computational auditory scene analysis. This talk describes and discusses several ways in which signals can be separated using physiologically-motivated cues, along with the potential benefit to be derived from such separation for automatic speech recognition.

1. INTRODUCTION[1]

Signal separation remains one of the most challenging and compelling problems in auditory perception, and a good solution for many core signal separation problems is necessary to improve the accuracy of contemporary automatic speech recognition systems in many practical environments.

As technology for automatic speech recognition is transferred from the laboratory environment into practical applications, the need to ensure robust recognition in a wide variety of acoustical environments becomes increasingly apparent. While algorithms designed to cope with the effects of unknown additive noise and unknown linear filtering are plentiful in number, today's applications also demand good performance in many more difficult environments. Some of the most challenging environments for speech recognition systems today include:

- Speech in high noise, with signal-to-noise ratios (SNRs) at or below 0 dB
- Speech in the presence of background speech
- Speech in the presence of background music
- Speech in highly reverberant environments

Conventional signal processing provides only limited benefit for these problems, even today.

In the spirit of this meeting, the goal of my talk is to suggest ways in which Al Bregman's huge corpus of creative research in auditory streaming and auditory scene analysis [3] can be exploited to improve the accuracy of automatic speech recognition systems. I will begin by briefly summarizing and commenting on some aspects of current state-of-the art speech recognition. I will then discuss ways in which cues that may be useful in separating speech signals can be extracted in ways that are based on some of the principles of auditory scene analysis.

2. ROBUST AUTOMATIC SPEECH RECOGNITION

The general topic of robust speech recognition has received a great deal of attention over the past decade. There are many sources of acoustical distortion that can degrade the accuracy of speech recognition systems. For many speech recognition applications the two most important sources of environmental degradation are unknown additive noise (from sources such as machinery, ambient air flow, and speech babble from background talkers) and unknown linear filtering (from a room, and spectral shaping by microphones or by the vocal tracts of individual speakers). Other sources of degradation include transient interference to the speech (such as doors slamming or telephones ringing), nonlinear distortion (arising from sources such as phase jitter in telephone systems), and "co-channel" interference by individual competing talkers. Similarly, there are many approaches to robust recognition, including the use of statistical estimation of and compensation for the effects of degradation, the use of physiologically-motivated signal processing techniques that mimic processing by the human auditory system, and the use of arrays of microphones. These approaches and others are reviewed in (among other places) [10], [18], [19], [20], and [21]. Most research in robust recognition has been directed toward compensation for the effects of additive noise and linear filtering.

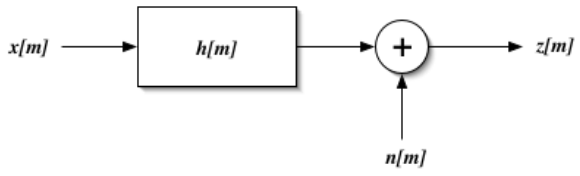

Figure 1. A model of environmental distortion including the effects of additive noise and linear filtering.

1. This paper is a rough transcription of my talk of the same name at the NSF Symposium on Speech Separation in Montreal on November 1, 2003. It is somewhat informal and speculative in nature, and is not presented at a level of scholarship and citation that would be appropriate for an archival publication.

2.1. Statistical approaches to robust recognition

Figure 1 describes the implicit model for environmental degradation introduced in 1990 [1] and now used in many signal processing algorithms developed at CMU and elsewhere. It is assume that the "clean" speech signaL $x[m]$ is first passed through a linear filter with unit sample response $h[m]$ whose output is then corrupted by uncorrelated additive noise $n[m]$ to produce the degraded speech signal $z[m]$. Under these circumstance, the goal of compensation is, in effect, to undo the estimated parameters characterizing the unknown additive noise and the unknown linear filter, and to apply the appropriate inverse operation. The popular approaches of spectral subtraction (*e.g.* [2]) and homomorphic deconvolution [24] are special cases of this model, in which either additive noise or linear filtering effects are considered in isolation. When the compensation parameters are estimated jointly, the problem becomes a nonlinear one, and can be solved using algorithms such as codeword-dependent cepstral normalization (CDCN) [1] and vector-Taylor series compensation (VTS) [14].

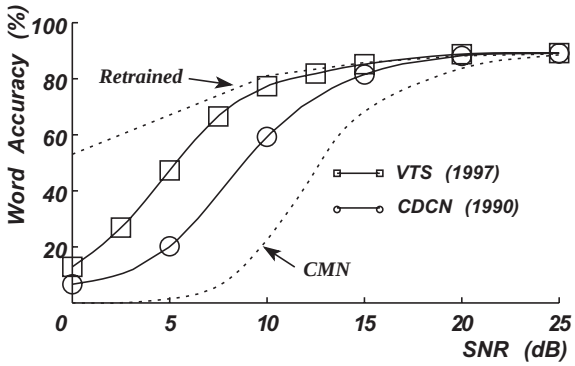

Figure 2. Comparison of recognition accuracies on the DARPA 5000-word Wall Street Journal task using CMN, CDCN, VTS, and complete retraining. From [14].

Figure 2 shows recognition accuracies for a standard dictation task obtained using the CMU SPHINX speech recognition system for speech in broadband noise plotted as a function of signal-to-noise ratio (SNR) [14]. The curve on the right represents the accuracy obtained using features derived from Mel frequency cepstral coefficients (MFCC) using cepstral mean normalization (CMN), which represents baseline performance for this particular system on this task with no particular compensation scheme used. The curve on the far left represents system performance obtained when the system is completely retrained for a particular noisy environment, which represents in a sense the upper bound in performance imposed by the particular noisy environment, given the type of signal processing and speech recognition algorithms used. The intermediate curves represent the recognition accuracy obtained using the CDCN [1] and VTS [14] algorithms, which were introduced in 1990 and 1997, respectively. The use of VTS provides an improvement of approximately 7 dB in SNR compared to the baseline processing. While that may not appear to be very much improvement, it can be the difference between virtually chance recognition performance and best possible performance at intermediate SNRs in the range of 5 to 10 dB, which is in important operating region.

Nevertheless, statistical parameter estimation compensation methods are not without their shortcomings. Figure 3 compares the *improvement* in word error rate (WER) obtained using CDCN for

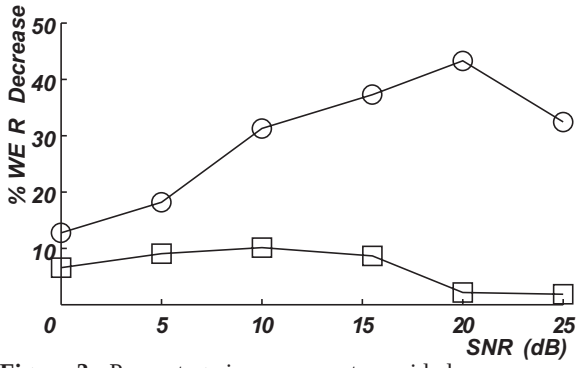

Figure 3. Percentage improvement provided by the CDCN algorithm for speech in the presence of white noise (circles) and background music (squares). From [15].

a similar speech recognition to that of Figure *2*. (Results with VTS would be similar). It is expected that the improvement provided by CDCN and similar algorithms would be small at very high SNRs (because the interfering signal introduces very little degradation at those SNRs) and at very low SNRs (because the noise produces almost complete degradation no matter what form of compensation is attempted. More interesting is the performance of CDCN compensation at intermediate SNRs, where the WER is decreased by almost 50 percent with background noise, but never by more than 10 percent in the presence of background music [15]. I believe that the failure of CDCN and similar compensation algorithms to provide meaningful compensation in the presence of background music to several factors including the nonstationarity of background music as well as its speechlike nature. A wide variety classical noise and channel compensation algorithms will exhibit similar deficiencies.

I believe that viable solutions to the problems of speech recognition at low SNRs and in the presence of transient and other types of time-varying interference must be based on the identification of the speech signal to be recognized, along with its explicit separation from the interfering signal or signals. This can in principle be accomplished by a number of techniques including any of several "missing-feature" approaches to noise compensation, as well as the techniques that are collectively referred to as "computational auditory scene analysis" (CASA). Researchers in CASA attempt to develop computational techniques that mimic the processes that are believed to mediate the identification and separation by humans of the separate components of a complex acoustical sound field. These approaches are discussed in the following two sections.

2.2. Missing-feature approaches to robust recognition

One potentially useful approach to speech recognition in the presence of the type of transient interference that is not handled well by algorithms like CDCN is the use of "missing-feature" techniques. Briefly, in missing-feature approaches, one attempts to determine which cells of a spectrogram-like time-frequency display of speech information are unreliable (or "missing") because of degradation due to noise or some other type of interference. The cells that are determined to be "missing" are either ignored in subsequent processing and statistical analysis, or they are "filled in" by optimal estimation of their putative values. While missing-feature approaches were initially motivated by similar techniques developed in image classification to deal with the problem of par-

tially-occluded objects and have been developed by a number of research groups, it is fair to say that Martin Cooke and his colleagues at the University of Sheffield have produced the most comprehensive and widely-adopted approaches to the problem (*e.g.* [5]).

The upper panel of Figure 4 (in the right hand column of this page) shows a spectrogram of an utterance recorded in quiet. The central panel of that figure shows the same utterance after it is mixed with white noise at an SNR of 15 dB. It can be seen that the major effect of the noise is to fill in the "valleys" of the spectrogram. The lower panel of Figure 4 shows the same spectrogram, but the pixels that have an effective SNR of less than zero dB are indicated by dark blue solid pixels.

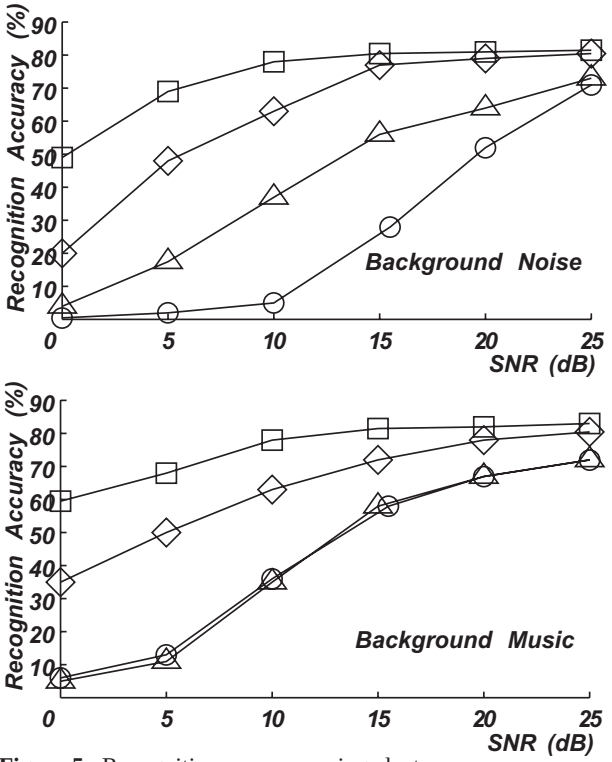

Figure 5. Recognition accuracy using cluster-based missing feature reconstruction (squares), covariance-based missing feature reconstruction (diamonds), simple spectral subtraction (triangles), and cepstral mean normalization only (circles) techniques for speech in the presence of white noise (upper panel) and music (lower panel) when perfect a priori information is available concerning which incoming features are "missing." Data from [16].

Figure 5 compares the speech recognition accuracy that is obtained using two types of missing-feature reconstruction techniques with baseline processing and simple spectral subtraction in the presence of artificially-added white Gaussian noise (upper panel) and background music derived from the DARPA Hub 4 task (lower panel), as a function of SNR [16]. The two missing-feature techniques that are used in these experiments, cluster-based reconstruction and covariance-based reconstruction, reconstruct the incoming feature vectors rather than modify the internal representation used by the classifier, as is more common (*e.g.* [5]). Recognition accuracy using missing-feature techniques can be quite good, even at low SNRs, while compensation using spectral subtraction does not improve performance at all in the presence of

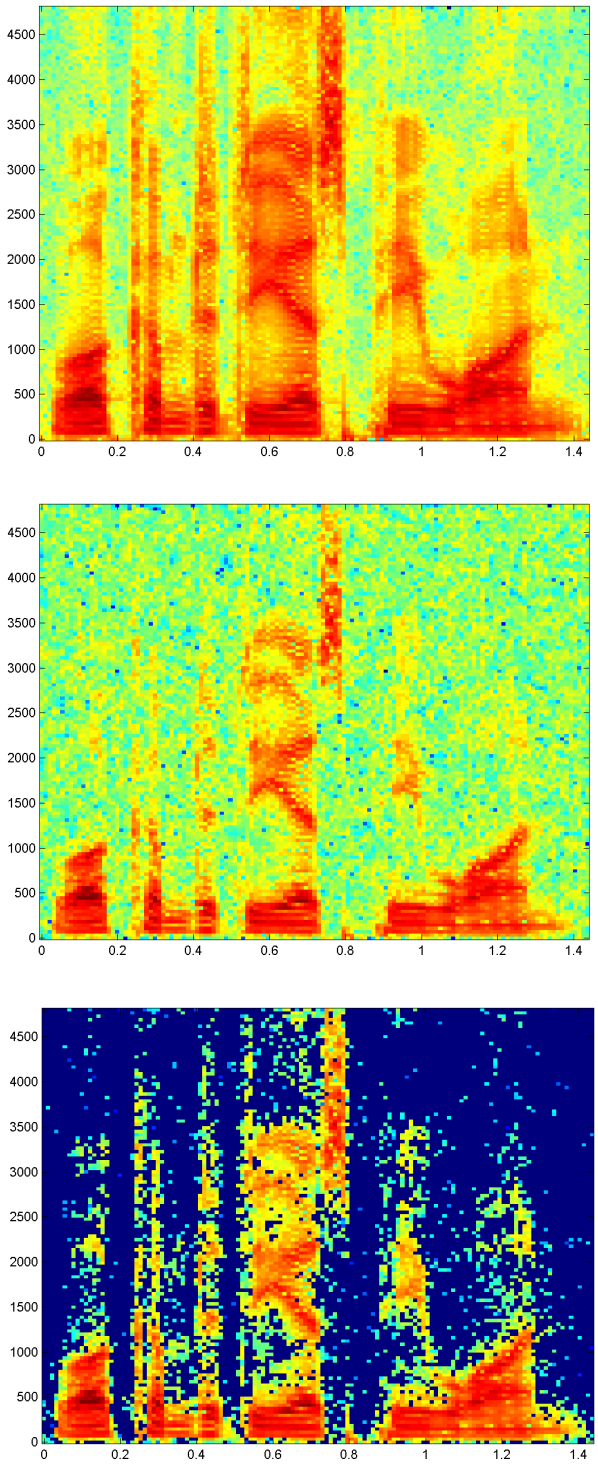

Figure 4. Spectrograms of speech recorded in quiet (upper panel) and subjected to artificially-added white noise with an SNR of 15 dB (central panel). Pixels that exhibit an SNR of less than zero dB are deemed missing and are depicted as solid dark regions.

music. (Algorithms like CDCN and VTS would perform similarly to spectral subtraction for stimuli like these.) Nevertheless, these results were obtained assuming perfect *a priori* knowledge of which pixels in the spectrogram-like representation are "missing" and which pixels are "present" (or more accurately which pixels are damaged and which are undamaged by the effects of noise). This type of information is normally not available in the recognition process and the blind determination of which pixels are or are not missing is in general a very difficult task.

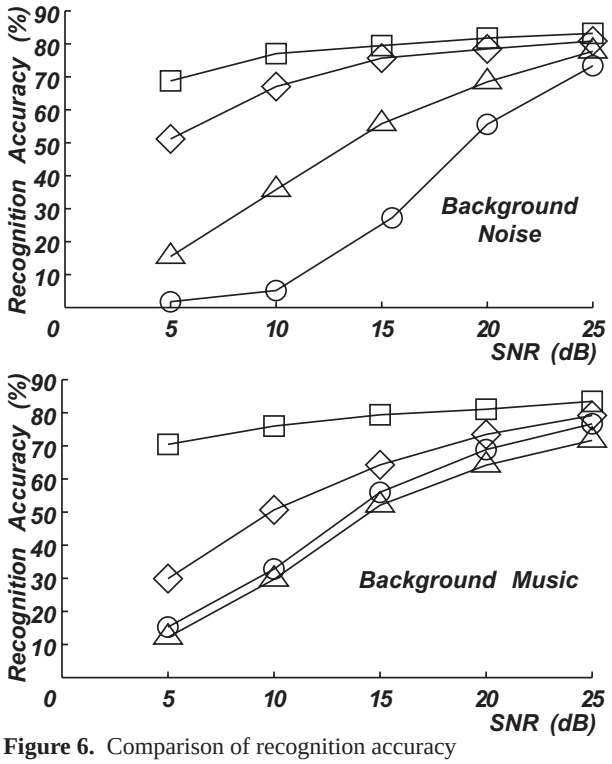

Figure 6. Comparison of recognition accuracy using cluster-based missing feature techniques assuming perfect "oracle" a priori knowledge of which features are missing (squares) and using blind identification of missing features (diamonds), using spectral subtraction as the basis for missing-feature decisions (triangles), and baseline processing (circles) for speech in the presence of while noise (upper panel) and music (lower panel). Data from [16], [17].

Figure 6 presents a more realistic picture of the current state-of-the-art in missing feature recognition in that it compares the recognition accuracy obtained with the missing features identified blindly using Bayesian techniques [17] with oracle missing-feature identification, along with results obtained using spectral subtraction to obtain missing-feature decisions, and baseline processing. The most effective features used by the Bayesian classifier track the fundamental frequency of voiced speech segments and estimate the fraction of total energy in a frame that is observed at frequencies that are harmonic multiples of the fundamental. The performance obtained with blind estimation of missing features approaches that observed with perfect oracle missing-feature identification in the case of background noise. Recognition accuracy obtained using cluster-based missing-feature compensation is not quite as good in the presence of background music, but it is still quite a bit better than the accuracy obtained with statistical estimation techniques like spectral subtraction, which do not provide any meaningful benefit at all.

2.3. Summary

While conventional techniques that compensate for the effects of additive noise and linear filtering of speech sounds can provide substantial improvement in recognition accuracy when the cause of the acoustical degradation is quasi-stationary, little improvement is observed at SNRs below approximately +5 dB. The use of techniques based on missing-feature analysis can provide substantial benefit at lower SNRs, but they are critically dependent on the ability to identify correctly which pixels actually are missing. The recognition of speech at lower SNRs, and especially speech in the presence of transient sources of interference including especially background speech and background music remain essentially unsolved problems at present.

3. APPLICATIONS OF AUDITORY SCENE ANALYSIS TO AUTOMATIC SPEECH RECOGNITION

Over a period of several decades, Al Bregman and his colleagues have compiled a monumental corpus of experimental results and schematic modeling that attempt to identify ways in which the human auditory system segregates and identifies components of a complex sound field (*e.g.* [3], [7]). While this work had originally been called "auditory streaming" by Bregman, it is now commonly known as "auditory scene analysis." The computational simulation and emulation of many of the processes identified by Bregman and his colleagues has become a popular topic of research by computer scientists and engineers in recent years, and these efforts are collectively referred to as "computational auditory scene analysis (CASA)."

Bregman *et al.* have identified many types of cues that can be used for auditory scene analysis of speech signals, including (among several others) fundamental frequency and harmonic relationships, spatial location cues, and correlated frequency and amplitude changes. In this section I will discuss some attributes about these cues and how they may be applied to improve automatic speech recognition accuracy.

3.1. Fundamental frequency and harmonic relationships

It has already been noted that pitch information can be extremely useful in the Bayesian determination of which pixels are missing in missing-feature analysis. In principle, the accurate identification and tracking of the fundamental frequency of voiced segments can be used to isolate the fundamental frequency and its harmonics from the background. In assessing the potential utility of pitch estimates as the basis for improved signal processing to achieve robust speech recognition, some key questions are how well speech can be separated from noise, how well speech signals can be separated from one another, and the extent to which this separation can improve recognition accuracy.

In order to assess some of these issues informally I made use of samples of speech in the Arctic database, which includes a phonetically-balanced corpus of read speech combined with electrolaryngograph (EGG) recordings collected by John Komenik and Alan Black as a resource for speech synthesis. Figure 7 shows an example sentence from the Arctic database, with the speech in the upper panel and the corresponding EGG recording in the lower panel. It is quite easy to extract an accurate pitch track from the EGG recordings.

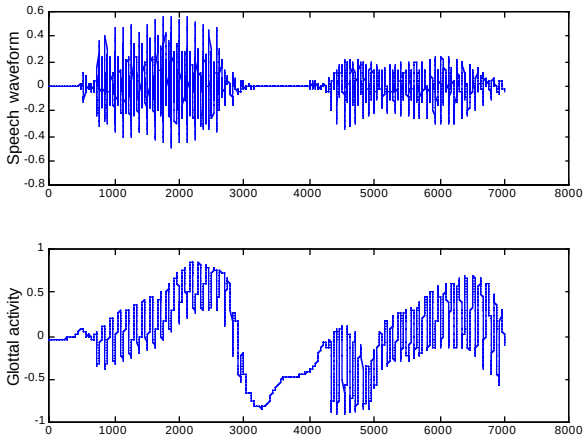

Figure 7. Examples of recordings from the Arctic database. A speech segment is shown in the upper panel, with the corresponding electrolaryngograph (EGG) shown in the lower panel.

In an informal pilot study, I analyzed two utterances from the Arctic database, by one male and one female speaker. The analysis and subsequent resynthesis of the speech were explored using two methods. The first approach, called synchronous heterodyne analysis (SHA), multiplies the incoming signal by a sine wave and cosine wave at the fundamental frequency, squares, and sums over time as shown in Figure 8.

In the second approach, called comb-filter analysis (CFA), the speech signal is passed through a comb filter with the transfer function

$$H(z) = \frac{z^{-P}}{1 - gz^{-P}}$$

This filter has a response with sharp peaks at integer multiples of the reciprocal of the parameter P, which represents the nominal period of the signal. Varying the parameter P in accordance with the estimated fundamental frequency, and using values of 0.8 to 0.9 for the parameter g, it is possible isolate the speech from background interference. Clearly neither SHA nor CFA provide any benefit for unvoiced segments of speech sounds or for whispered speech.

Speech in isolation was analyzed and resynthesized using the SHA and CFA methods. For both male and female speakers informal listening suggests that intelligibility is fair to good using the SHA method and good to excellent using the CFA method. Male and

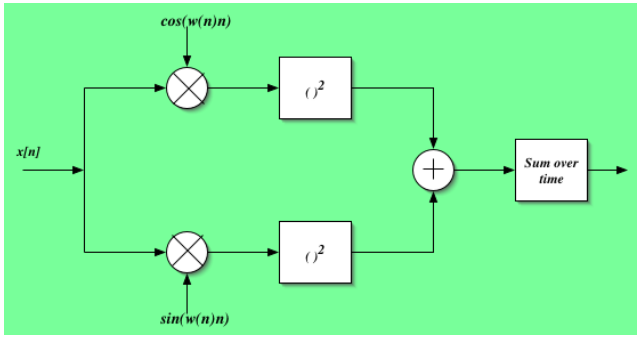

Figure 8. Synchronous heterodyne analysis used for separating speech sounds.

femail speech sounds were also added together at 0 dB SNR and attempted to separate them using SHA and CFA, with the heterodyne frequency in SHA or the fundamental frequency in CFA tuned to the target speaker. Using both SHA and CFA, the separated speech of the male was fair to good in intelligibility, while the separated speech of the female was poor to fair. I believe that this asymmetry in performance is a consequence of the differing spectral regions of male and female speech. Specifically, when the male is the target speaker, the higher fundamental frequency of the interfering female causes her speech components to be spaced farther apart in frequency, imposing a smaller amount of degradation on the upper components of the target male. Conversely, when the female is the target, the upper partials of the speech of the interfering male are relatively dense, and they are more likely to interfere with the perceptually-important lower harmonics of the speech of the target female.

While neither the SHA or CFA technique have yet been used in actual speech recognition experiments, I regard them as promising, both for speech at lower SNRs, and for speech in the presence of interfering speech and music. Again, it must be stressed that the results described above were obtained using perfect "oracle" knowledge of the fundamental frequency of the target speaker. While fundamental frequency extraction continues to be the object of a great deal of attention in recent years (*e.g.* [8], 12]), pitch tracking, and especially tracking the pitch of multiple speech or music sources, remains a very difficult problem. As noted above, these techniques are not useful for unvoiced speech segments.

3.2. Spatial location cues

Sound sources arriving from different azimuths produce interaural time delays (ITDs) and interaural intensity differences (IIDs) as they arrive at the two ears. It is well known that human listeners can use spatial information to improve the intelligibility of speech in the presence of other speech or noise interference [26]. The binaural hearing mechanism can focus attention on a target speaker in a complex acoustical environment, or it can focus on the direction of arrival of the direct sound field of a target speaker in a reverberant environment. The mechanisms underlying these abilities are not completely understood, and as Zurek has noted, some improvement is to be expected simply by attending solely to the ear that is closer to the target speech source [26].

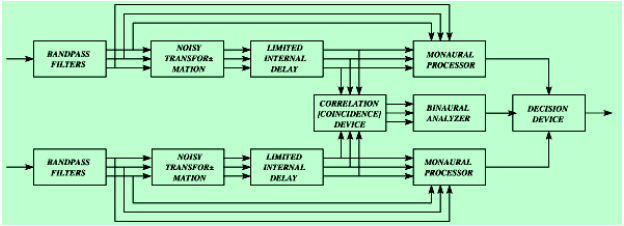

Figure 9. Generic model of binaural processing proposed by Colburn and Durlach [4]. The parallel sets of arrows indicate multiple parallel channels of information in the model.

Most models of binaural perception (*e.g.* [4], [22]) assume that peripheral auditory processing includes bandpass filtering and nonlinear rectification of the incoming sounds, followed by a cross-correlation analysis of the bandpass-filtered and rectified signals, with subsequent analysis (at least for simple stimuli) based on consideration of ITD and IID information as a function of frequency. This processing is summarized by the block diagram of Figure 9. Most models of binaural interaction assume that

ITD information is extracted using a coincidence-analysis mechanism first proposed by Lloyd Jeffress in 1948 [9]. Figure 10 shows the putative representation of interaural timing information in response to bandpass noise presented with an ITD of –1.5 ms with a center frequency of 500 Hz and bandwidths of 50 Hz (upper panel) and 800 Hz (lower panel).

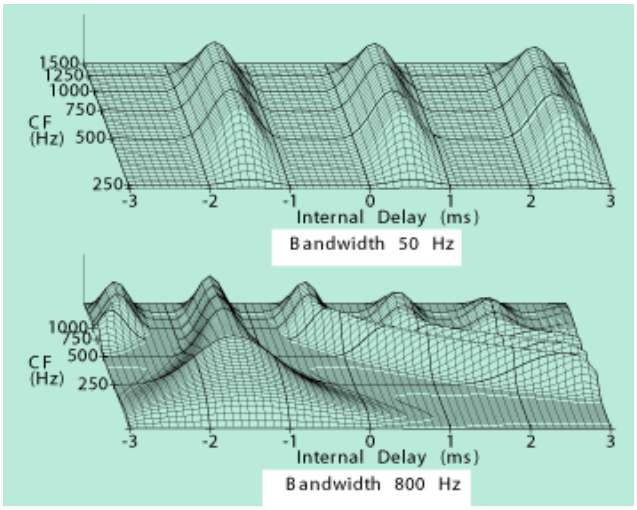

Figure 10. The putative response of an ensemble of Jeffress-Colburn coincidence-counting units to low-frequency bandpass noise with a center frequency of 500 Hz and an ITD of –1.5 ms. Upper panel: response to bandpass noise with a bandwidth of 50 Hz. Lower panel: response to bandpass noise with a bandwidth of 800 Hz. From [22].

There is some disagreement concerning the extent to which ITD information is used by humans as the basis of signal separation. While the results of one influential study by Culling and Summerfield imply that simultaneously-presented unmodulated whispered are not separated by their ITDs [6], more recent studies have shown that these ITD can be a useful cue in fostering identification of simultaneously-presented speech sounds when they are presented with natural amplitude and frequency modulations [23]. In any case, even if the auditory system does not make efficient use of interaural timing information, these physical cues are still available for computational auditory processors. In the 1980s and 1990s, many types of signal processing such as spectral subtraction that did not improve human intelligibility still proved to be useful for improving speech recognition accuracy. My work in this area is motivated by the belief that we should be inspired by but not limited by our knowledge of the auditory system.

The block diagram of one system that used ITD information to improve speech recognition accuracy is shown in Fig. 11 [25]. The input signals are first delayed in order to compensate for differences in the acoustical path length of the desired speech signal to each microphone. (This is the same processing performed by conventional delay-and-sum beamforming.) The signals from each microphone are passed through a bank of bandpass filters with different center frequencies, passed through nonlinear rectifiers, and the outputs of the rectifiers at each frequency are correlated. (The correlator outputs correspond to outputs of the coincidence counters at the internal delays of the "ridges" in Fig. 10 at –1.5 ms.) The result of this operation is a form of *N*-dimensional cross-correlation, which reduces to conventional cross-correlation operation for two inputs. The outputs of the multi-dimensional cross-correlation operation are considered as if they were energy esti-

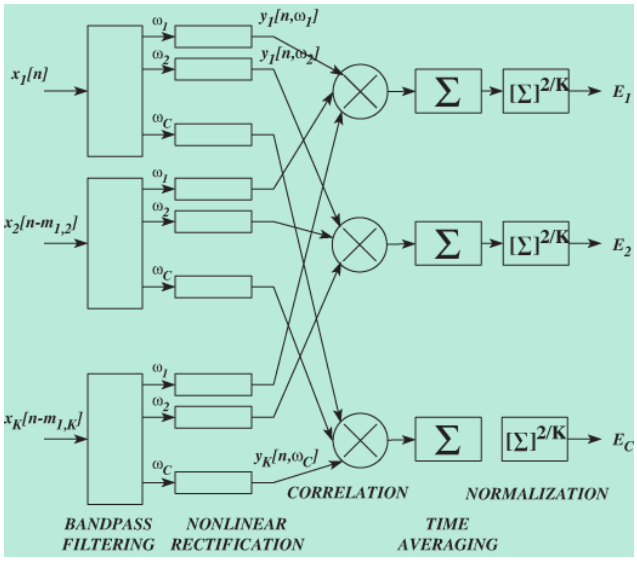

Figure 11. Block diagram of multi-microphone cross-correlation-based processing system. From [25].

mates short-time energy estimates in each of the frequency channels, and they are subsequently converted into 12 cepstral coefficients using the cosine transform. These cepstral coefficients along with an additional coefficient representing the power of the signal are used as features for speech recognition in the conventional fashion.

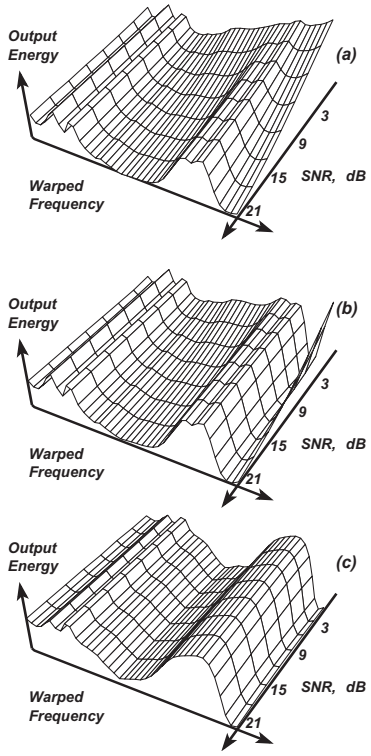

Figure 12. Estimates of frequency-warped spectra for the vowel segment /a/ for various SNRs using (a) 2 input channels and zero delay, (b) 2 input channels and 125-μs delay to successive channels, and (c) 8 input channels and 125-μs delay. From [25].

Figure 12 demonstrates the validity of the such processing in the context of an analysis of a sample of the digitized vowel segment /a/ corrupted by artificially-added white Gaussian noise at global SNRs of 0 to +21 dB. The speech segment was presented to all microphone channels identically (to simulate a desired signal arriving on axis) and the noise was presented with linearly increasing delays to the channels (to simulate an off-axis corrupting signal impinging on a linear microphone array). The processing of such a system was simulated using 2 and 8 microphone channels, and time delays for the masking noise of 0 and 125 μs to successive channels.

The curves of Figure 12 describe the effect of SNR, the number of processing channels, and the delay of the noise on the spectral profiles of a sample of the vowel segment /a/. (The frequency representation for the vowel is warped in frequency according to the nonlinear spacing of the auditory filters.) The upper panel summarizes the results that are obtained using 2 channels with the noise presented with zero delay from channel to channel (which would be the case if the speech and noise signals arrive from the same direction). Note that the shape of the vowel, which is clearly defined at high SNRs, becomes almost indistinct at the lower SNRs. The center and lower panels show the results of processing with 2 and 8 microphones, respectively, when the noise is presented with a delay of 125 μs from channel to channel (which corresponds to a off-axis source location for typical microphone spacing). As the number of channels increases from 2 to 8, the shape of the vowel segment in Figure 12 becomes much more invariant to the amount of noise present. In general, it was found in these experiments that the benefit to be expected from increases sharply as the number of microphone channels is increased. It was also observed (unsurprisingly) that the degree of improvement increases as the simulated directional disparity between the desired speech signal and the masker increases. It was concluded from these pilot experiments that the cross-correlation method described can provide very good robustness to off-axis additive noise, and in practice, this approach did provide a moderate benefit over the recognition accuracy obtained using conventional delay-and-sum processing [25].

These studies, which were conducted in the early 1990s, were not continued because of the lack of available computational resources at that time. I believe that further improvements are likely to be obtained once greater attention is paid to the nature of the bandpass filtering, within-channel nonlinearities, and correlation operations. Correlation-based approaches such as these can be applied to unvoiced as well as to voiced segments of speech.

A final item of note with regard to spatial processing is that high levels of reverberation are extremely detrimental to recognition accuracy. Frame-based compensation strategies such as those discussed in Sec. 2.1 fail because the effects of reverberation are generally spread over multiple analysis frames. In addition, traditional adaptive filtering methods, which have also been considered for this purpose, depend on the statistical independence of target and masker. Since in reverberant environments, the "noise" consists of reflected and attenuated copies of the target speech signals, noise-canceling adaptive filter strategies (such as those that use the LMS algorithm) are not effective. Sub-band processing in a similar fashion has been somewhat effective in reverberation, but it has not yet been applied to a wide range of problems. I believe that techniques based on auditory perception and physiology, missing-feature recognition, and CASA techniques should be more effective in characterizing and ultimately ameliorating the effects of reverberation.

3.3. Correlated frequency and amplitude changes

Although cues based on fundamental frequency and sound source location are potentially extremely valuable in separating signals for automatic speech recognition, pitch cues are ineffective for unvoiced segments and location multichannel recordings with location information are not always available. Even with just a single channel, unvoiced speech segments, and/or imperfect pitch estimates, we expect to be able to separate multiple sound sources using by extracting and clustering sounds according to small ("micro") modulations in frequency and amplitude. The use of such physical cues for sound separation and auditory scene analysis has been supported by many psychoacoustical studies in recent years (*e.g.* [*3*], [*7*]).

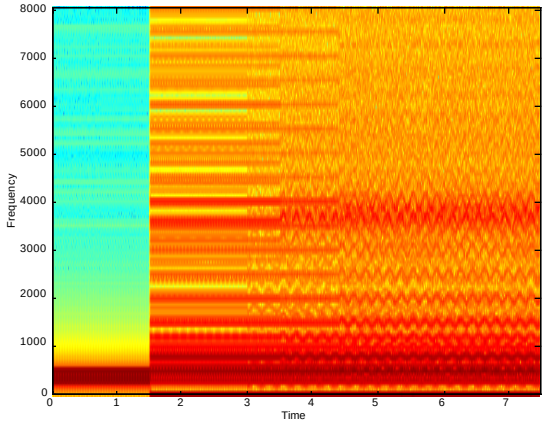

Figure 13. Spectrogram of the signal used by Chowning to demonstrate the importance of correlated frequency modulation in auditory object formation.

Figure 13 is a spectrogram of a signal that John Chowning developed in the early 1980s to demonstrate the perceptual salience of correlated frequency modulation in fusing and separating components of a complex signal. These signals are described in detail in Bregman's treatise [3]. The signal from 0 to 1.5 ms consists of three sine waves, at 300, 400, and 500 Hz. Taken as individual sine waves, these three frequencies form a major triad in the second inversion, but the components are more likely to be heard as a fused complex tone with fundamental frequency 100 Hz. From 1.5 to 3 ms the three sine waves are replaced by three sets of 10 harmonics at integer multiples of 300, 400, and 500 Hz, with spectral envelopes that are derived from three different vowel sounds. Again this signal is perceived as a complex tone with fundamental frequency 100 Hz, but with a sharper timbre because of the presence of 27 additional upper-frequency partials. Beginning at 3 ms, the signals consist of the same three overtone series, but each is separately modulated at 4.5, 5, and 5.5 Hz, respectively. Once the frequency modulation is applied, the harmonics associated with each fundamental frequency segregate from one another and become easily perceived as three separate complex tones with fundamental frequencies of 300, 400, and 500 Hz.

My research group is currently developing ways to develop through computational means a form of signal separation based on the identification of common locations in a time-frequency display like a spectrogram that exhibit covarying amplitude and frequency modulation. This is a difficult task because of the need to achieve sharp frequency resolution while allowing for temporal fluctuations, but the ability of the human auditory system to

accomplish these tasks remains a powerful existence proof and motivation to move this work forward.

4. SUMMARY

I believe that computational auditory approaches are potentially extremely useful in ameliorating some of the most difficult speech recognition problems, specifically the recognition of speech presented at low SNRs, speech masked by other speech, speech masked by music, and speech in highly reverberant environments. The solution to these problems using CASA techniques is likely to depend on the ability to develop several key elements of signal processing, including the reliable detection of fundamental frequency for isolated speech and for multiple simultaneously-presented speech sounds, the reliable detection of modulations of amplitude and frequency in very narrowband channels, and the development of across-frequency correlation approaches that can identify frequency bands with coherent micro-activity as they evolve over time. I am extremely optimistic that effective solutions for these problems are within reach in the near future.

REFERENCES[2]

[1] A. Acero, and R. M. Stern, Environmental Robustness in Automatic Speech Recognition, *Proc. ICASSP*, Albuquerque, New Mexico, 1990.

[2] [6] S. F. Boll, Suppression of Acoustic Noise in Speech Using Spectral Subtraction, *IEEE Trans. Acoustics, Speech and Signal Processing,* **27:** 113-120, 1979.

[3] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound,* Cambridge: MIT Press 1990.

[4] H. S. Colburn and N. I. Durlach, N. I. Models of binaural interaction, in Handbook of Perception, E. C. Carterette and M. P. Friedman, ed., New York: Academic Press, pp. 467-518, 1978.

[5] M. Cooke, P. Green, L. Josifovski, and A, Vizinho, Robust Automatic Speech Recognition with Missing Features and Unreliable Acoustic Data, *Speech Communication,* **34:** 267-285, 2001.

[6] J. F. Culling and Q. Summerfield, Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay, *J. Acoust. Soc. Am*. **98**: 785-797, 1995.

[7] C. J. Darwin and R. P. Carlyon, Auditory Grouping, in *Handbook of Perception and Cognition, Vol. 6: Hearing*, pp. 347-386, B. C. J. Moore., Ed. New York: Academic Press, 1995.

[8] A. de Cheveign and A. Baskind, F0 Estimation of One or Several Voices, *Proc. Eurospeech,* pp. 833-836, 2003.

[9] L. A. Jeffress,. A place theory of sound localization, *J. Comparative and Physiological Psychology* **41**: 35-39, 1948.

[10]B.-H. Juang, Speech Recognition in Adverse Environments, *Computer Speech and Language*, **5**: 275-294, 1991.

[11]W. Lindemann, Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals, *J. Acoust. Soc. Am*. **80**, 1608-1622, 1986.

[12]H. Kawahara, I. Matsuda-Katsuse, and A. de Cheveign , Restructuring Speech Representations using a Pitch Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds, *Speech Communication,* **27**: 175-185, 1999.

[13]J. Komenek and A. Black *CMU_ARCTIC Databases,* 2003. Available at `http://www.festvox.org/cmu_arctic`.

[14]P. J. Moreno, B. Raj, and R. M. Stern, A Vector Taylor Series Approach For Environment-Independent Speech Recognition, *Proc. ICASSP*, Atlanta, GA, May 1996.

[15]B. Raj, V. N. Parikh, and R. M. Stern, The Effects Of Background Music On Speech Recognition Accuracy, *Proc. ICASSP*, Munich, Germany, April 1997.

[16]B. Raj, M. L. Seltzer, and R. M. Stern, Reconstruction of Missing Features for Robust Speech Recognition, *Speech Communication Journal*, accepted for publication, 2004.

[17]M. L. Seltzer, B. Raj, and R. M. Stern, A Bayesian Framework for Spectrographic Mask Estimation for Missing Feature Speech Recognition, *Speech Communication Journal,* accepted for publication, 2004.

[18]R. Singh, R. M. Stern, and B. Raj, Signal and Feature Compensation Methods for Robust Speech Recognition, Chapter in *CRC Handbook on Noise Reduction in Speech Applications*, Gillian Davis, Ed. CRC Press, 2002.

[19]R. Singh, B. Raj, and R. M. Stern, Model Compensation and Matched Condition Methods for Robust Speech Recognition, Chapter in *CRC Handbook on Noise Reduction in Speech Applications*, Gillian Davis, Ed. CRC Press, 2002.

[20]R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, Signal Processing for Robust Speech Recognition, Chapter in *Speech Recognition*, pp. 351-378, C.-H. Lee and F. Soong, Eds., Boston: Kluwer Academic Publishers, 1996.

[21]R. M. Stern, B. Raj, and P. J. Moreno, (1997). Compensation for Environmental Degradation in Automatic Speech Recognition, *Proc. ETRW on Robust Speech Recognition for Unknown Communication Channels*, April, 1997, Pont-au-Mousson, France, pp. 33-42.

[22]R. M. Stern and C. Trahiotis, Models of Binaural Interaction, in *Handbook of Perception and Cognition, Volume 6: Hearing*, pp. 347-386, B. C. J. Moore., Ed. New York: Academic Press, 1995.

[23]R. M. Stern, C. Trahiotis, A. M. Ripepi, Some Conditions Under Which Interaural Delays Foster Identification, in *Dynamics of Speech Production and Perception*, G. Meyer and P. Divenyi, Eds., Amsterdam: IOP Press, 2004.

[24]T. G. Stockham, T. M. Cannon and R. B. Ingebretsen, Blind Deconvolution Through Digital Signal Processing, *Proc. IEEE*, **63**: 678-692, 1975.

[25]T. M. Sullivan and R. M. Stern, Multi-Microphone Correlation-Based Processing for Robust Speech Recognition, *Proc. ICASSP*, Minneapolis, Minnesota, 1993.

[26]P.M. Zurek, Binaural Advantages and Directional Effects in Speech Intelligibility, in *Acoustical Factors Affecting Hearing Performance II,* G. A. Studebaker and I. Hochberg, Eds. Boston: Allyn and Bacon, 1993.

2. Copies of most papers for which I am a co-author can be found at `http://www.cs.cmu.edu/~robust/papers.html`.