ON THE USE OF EXPLICIT SPEECH MODELING IN MICROPHONE ARRAY APPLICATIONS

Michael S. Brandstein

Division of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
msb@hrl.harvard.edu

ABSTRACT

This paper addresses the limitations of current approaches to distant-talker speech acquisition and advocates the development of techniques which explicitly incorporate the nature of the speech signal (e.g. statistical non-stationarity, method of production, pitch, voicing, formant structure, and source radiator model) into a multi-channel context. The goal is to combine the advantages of spatial filtering achieved through beamforming with knowledge of the desired time-series attributes. The potential utility of such an approach is demonstrated through the application of a multi-channel version of the Dual Excitation speech model.

1. INTRODUCTION

For close-talker environments, single channel speech enhancement systems have been well-studied and proven effective at improving the perceived quality of speech degraded by low to moderate levels of background noise. Examples of these approaches include Spectral Subtraction and its many variations, Wiener Filtering, Adaptive Noise Cancellation, and Comb Filtering. Summaries of these techniques may be found in any of a number of references, for instance [1, 2]. In recent years, more sophisticated speech models have been applied to the enhancement problem. In addition to utilizing the periodic features of the speech, as in the case of comb filtering, these systems exploit the signal's mixture of harmonic and stochastic components. By separating speech into voiced and unvoiced portions and performing a spectral modification to the unvoiced part alone, the Dual Excitation (DE) Speech Model [3] (a cousin to the Multiband Excitation Model popular in the low-rate speech coding field [4, 5]) is able to achieve results free of the tonal artifacts associated with traditional spectral modification methods and to increase speech quality and intelligibility [6]. A similar approach to speech decomposition is adopted by the Harmonic Plus Noise Model [7]. Here the speech is represented in the time-domain as the sum of harmonically linked sinusoids (similar to the Sinusoidal Speech Model [8]) and a noise signal. The method has been applied in speech enhancement and modification scenarios. By associating harmonics with distinct speakers, sinusoidal modeling has also shown some effectiveness for co-channel separation [9].

Microphone arrays have seen increasing application for speech enhancement in challenging acquisition environments, particularly in those situations where the talker is physically separated from the input device. By employing spatial filtering in addition to temporal processing, multi-channel algorithms offer a distinct performance advantage over single-channel techniques in the presence of additive noise, interfering sources, and distortions due to multipath channel effects. The simplest microphone array method is the Delay and Sum (DS) Beamformer which derives its output via averaging of the time-synchronized microphone data. A variety of more sophisticated algorithms exist for adaptively 'steering' the array in the direction of the desired source and simultaneously adjusting the microphone 'weightings' to minimize the contributions of noise sources [10]. These techniques usually assume the desired source is stationary and at a known location. While dynamic localization schemes [11] and weighting constraints may be incorporated into the adaptation procedure, these methods are very sensitive to steering errors which limit their noise source attenuation performance and frequently distort or cancel the desired signal. Furthermore, these algorithms are oriented solely toward noise reduction and have limited effectiveness at enhancing a desired signal corrupted by reverberations. A variant approach is based upon attempting to undo the effects of multipath propagation. By multiple beamforming on the direct path and the major images it is possible to use the multipath reflections constructively to increase signal to noise ratios well beyond those achieved with a single beamformer. The result is a matched filtering process [12] which is effective at enhancing the quality of reverberated speech as well as attenuating noise sources. Unfortunately, this technique has a number of practical shortcomings. The matched filter is derived from the source location-dependent room response and as such is difficult to estimate dynamically. In [12] the room responses are calculated from a model of the enclosure geometry or measured *a priori* in actual rooms. The channel responses obtained in this manner do not address the issue of non-stationary (or unknown) source locations or changing acoustic environments. These problems are addressed in [13] by attempting to adaptively estimate the channel responses and incorporate the results into an adaptive beamforming process.

2. LIMITATIONS OF THE CURRENT FIELD

Single channel enhancement schemes have achieved only moderate success. While capable of improving perceived quality in restrictive environments (additive noise, no mul-

tipath, high to moderate SNR, single source), non-model based approaches (Spectral Modification, Wiener Filtering, Adaptive Noise Cancellation) provide no mechanism for addressing reverberant distortions, competing sources, and severe noise conditions. The model-based techniques achieve a performance improvement over their earlier single channel counterparts, both in speech quality and intelligibility. Additionally, these methods, by virtue of their specific parameterization of the speech signal, offer some applicability to the more general acquisition problem. Currently, however, these model-based estimation schemes are limited by their single channel application.

While single-channel techniques exploit various features of the speech signal itself, multi-channel methods have focused primarily on improving the quality of the spatial filtering process. Beamforming research has dealt with algorithms to attenuate undesired sources and noise, track moving sources, and deconvolve channel effects. These approaches, while effective to some degree, are fundamentally limited by the nature of the distant talker environment. Array design methods are overly sensitive to variations in their assumptions regarding source locations and radiation patterns and inflexible to the complex and time-varying nature of the enclosure's acoustic field. A talker turning his head or motion as little as a few centimeters is frequently sufficient to compromise the optimality of these schemes in practical scenarios. Similarly, matched filtering processing, while shown to be capable of tracking source motion to a limited degree, is not adaptable at rates sufficient to capture effectively the motions of a realistic talker. This point is illustrated in plots A) and B) of Figure 1 where the spectra of a voiced speech segment are plotted for two closely-spaced (10cm separation) source locations in the center of a simulated noiseless 4mx4mx3m rectangular room with plane reflective surfaces and uniform, frequency-independent reflection coefficients equivalent to a 400ms reverberation time. Room impulse responses were generated for 8 microphones with 25cm spacing positioned along one wall of the enclosure using the image model technique [14] with intra-sample interpolation and up to sixth order reflections. Both the microphones and sources were assumed to have cardioid patterns and the sources were oriented toward the center of the array. The bold lines correspond to the spectrum of the original speech while the dotted lines plot the spectra of the data received at each of 8 microphones placed on one wall of the enclosure. The reverberation effects are multiplicative in the frequency domain and vary considerably from channel to channel. Note that even for this very simple simulation there are significant variations in the channel responses when the source is moved just a few inches. The implication is that any system which attempts to estimate the reverberation effects and apply some means of inverse filtering would have to be adaptable on almost a frame-by-frame basis to be effective.

The current approach in microphone array research, to identify and compensate for environmental enclosure effects, is an extremely difficult (if indeed solvable) problem. While avenues are and will continue to be made, it does not seem likely that such schemes alone will achieve the stated goal of this research, namely to acquire a high-quality speech signal from an unconstrained talker in a hands-free environment surrounded by interfering sources (the "cocktail party" problem). Any effective solution will require some application of knowledge regarding the desired signal content. For single channel enhancement methods this has proven, out of necessity perhaps, to be a path to higher performance. These observations suggest that the incorporation of speech modeling with the benefits of spatial filtering offered through array technology is a logical step towards advancing the field of distant-talker speech acquisition.

3. A MULTI-CHANNEL SPEECH MODEL

As a specific example of this proposed speech modeling/spatial filtering fusion, consider a potential application of the Dual Excitation Speech Model [3]. In the single-channel DE model a windowed segment of speech, $s[n]$, is represented as the sum of two components: a voiced signal, $v[n]$, and an unvoiced signal, $u[n]$. In the frequency-domain, the relationship may be expressed as:

$$S(\omega) = V(\omega) + U(\omega) \tag{1}$$

where $S(\omega)$, $V(\omega)$, and $U(\omega)$ correspond to the Fourier Transforms of $s[n]$, $v[n]$, and $u[n]$, respectively. The voiced portion is assumed to be periodic over the time window and may be represented as the sum of the harmonics of a fundamental frequency, ω_0:

$$V(\omega) = \sum_{m=-M}^{M} A_m W(\omega - m\omega_0) \tag{2}$$

where $W(\omega)$ is the Fourier Transform of the window, A_m is the complex spectral amplitude of the m^{th} harmonic, and M is the total number of harmonics ($M = \lfloor \pi/\omega_0 \rfloor$). Following [4], the fundamental frequency and harmonic amplitudes are estimated through minimization of the mean-squared error criterion:

$$\mathcal{E} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S(\omega) - \sum_{m=-M}^{M} A_m W(\omega - m\omega_0)|^2 d\omega \tag{3}$$

This non-linear optimization problem may be decoupled efficiently by noting that for a given fundamental frequency, the harmonic amplitudes which minimize the error are found through the solution of a set of uncoupled linear equations. The optimal parameter set may then be calculated through global minimization of the error function in (3) versus all fundamental frequencies of interest.

The estimated unvoiced signal plus noise spectrum, $\hat{U}(\omega)$, is then found from the difference spectrum:

$$\hat{U}(\omega) = S(\omega) - \hat{V}(\omega) \tag{4}$$

where $\hat{V}(\omega)$ is the estimated voiced spectrum derived from (2) using the estimated values of ω_0 and A_m.

The utility of the DE model for improving speech degraded by background noise lies in its independent enhancement of the voiced and unvoiced components of the speech. Assuming that the degrading noise is independent of the harmonic structure, the voiced spectrum is subjected to only a minor thresholding operation relative to

the background noise power. The bulk of the enhancement is achieved by nulling out the unvoiced portions of strongly voiced harmonics and applying modified Wiener filter to the remaining unvoiced spectral regions.

A number of methods are available for extending the DE model within a multi-channel context to improve its effectiveness for the additive noise case and to address the more general distant-talker scenario involving multipath channels and multiple sources. Consider first the extension of the DE error criterion in (3) to include data from N channels:

$$\mathcal{E}_N = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{1}{N}\sum_{i=1}^{N}G_i(\omega)S_i(\omega) - \sum_{m=-M}^{M}A_m W(\omega - m\omega_0)\right|^2 d\omega \qquad (5)$$

where $G_i(\omega)$ is a filter associated with the i^{th} channel and $S_i(\omega)$ is the short-term spectrum of the data received at the i^{th} microphone. Alternatively, for environments where the dominant degradation effect is reverberant, it may be advantageous to recast the above error criterion as the L_2 norm in the log spectrum domain.

The voiced signal estimate, $\hat{V}_N(\omega)$, derived from the parameters minimizing (5) would then be used to produce the unvoiced signal plus noise spectrum from:

$$\hat{U}_N(\omega) = \frac{1}{N}\sum_{i=1}^{N}H_i(\omega)(G_i(\omega)S_i(\omega) - \hat{V}_N(\omega)) \qquad (6)$$

The channel filters, $G_i(\omega)$, in a fashion similar to multi-channel approaches summarized earlier, could be designed to provide appropriate spatial filtering, addressing issues of noise-reduction, attenuation of interfering sources, and dereverberation. Additionally, the channel-dependent weighting filters, $H_i(\omega)$, could be incorporated as a multi-channel post-processor to exploit known signal characteristics.

In the simplest case of an additive noise condition, the extension of the Dual Excitation model to a plurality of channels would stand to improve its enhancement performance by virtue of the data averaging alone. With the inclusion of the spatial filtering afforded through Equations (5) - (6) it is possible to give the DE model a robustness to channel effects and interfering sources. With regard to multiple sources, the error criterion in (5) could be extended explicitly to include L sources and N channels by:

$$\mathcal{E}_{LN} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{1}{L}\sum_{j=1}^{L}\left(\frac{1}{N}\sum_{i=1}^{N}G_{ij}(\omega)S_i(\omega) - \sum_{m=-M}^{M}A_{mj}W(\omega - m\omega_{0j})\right)\right|^2 d\omega \qquad (7)$$

where $G_{ij}(\omega)$ is the spatial filter associated with the i^{th} channel and j^{th} signal source, ω_{0j} is the fundamental frequency of the j^{th} source, and A_{mj} is the amplitude of the m^{th} harmonic associated with the j^{th} source. Using this approach it would be possible to track individual sources through a combination of location and pitch data. Such a multi-channel DE model would have the ability to isolate and enhance a desired source signal by employing both spatial and signal-content information.

4. SIMULATION

To illustrate the effectiveness of such an approach, again consider the example of the voiced speech segment in Figure 1. Plots C) and D) show the relationship between the Delay and Sum Beamformer and the voiced signal estimate, $\hat{V}_N(\omega)$, derived from the proposed multichannel scheme for the two closely spaced source positions. Each set of results was generated using delays appropriate for the source 1 location. This would correspond to a 10cm mis-aim in the source 2 case. As the plots suggest, by exploiting the periodic nature of the desired signal, the proposed scheme achieves a better fit to the original signal spectrum. Unlike the Delay and Sum method, the approach is relatively insensitive to imperfect knowledge of the source location suggesting a robustness to the small, but nominal, variations encountered in a practical operating environment. This result is confirmed by more quantitative methods, such as SNR and log spectral distortion scores.

5. DISCUSSION

The multi-channel speech enhancement problem has the potential to benefit by being cast as a multi-dimensional estimation scheme based upon a specific parameterization model, rather than purely in a spatial filtering context as it is today. The Dual Excitation Model is just one example of the application of a sophisticated speech model to the problem. Similarly, the multi-channel method presented represents only one of many possible enhancement schemes. Future work along these lines will investigate strategies to fuse proven methods in the single and multiple channel enhancement fields as well as seek to develop novel and effective algorithms.

REFERENCES

[1] J. Lim, editor. *Speech Enhancement*. Prentice Hall, 1983.

[2] J. Deller, J. Proakis, and J. Hansen. *Discrete-Time Processing of Speech Signals*. Prentice Hall, first edition, 1987.

[3] J. Hardwick. *The Dual Excitation Speech Model*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, June 1992.

[4] D. Griffin and J. Lim. Multiband excitation vocoder. *IEEE Trans. Acoust., Speech, Signal Processing*, vol.36(8):1223–1235, August 1988.

[5] M. Brandstein, J. Hardwick, and J. Lim. The multiband excitation speech coder. In B.S. Atal, V. Cuperman, and A. Gersho, editors, *Advances in Speech Coding*, pages 215–224. Kluwer Academic Publishers, 1990.

[6] J. Hardwick, C. Yoo, and J. Lim. Speech enhancement using the dual excitation speech model. In *Proceedings of ICASSP93*, pages II–367–II–370. IEEE, 1993.

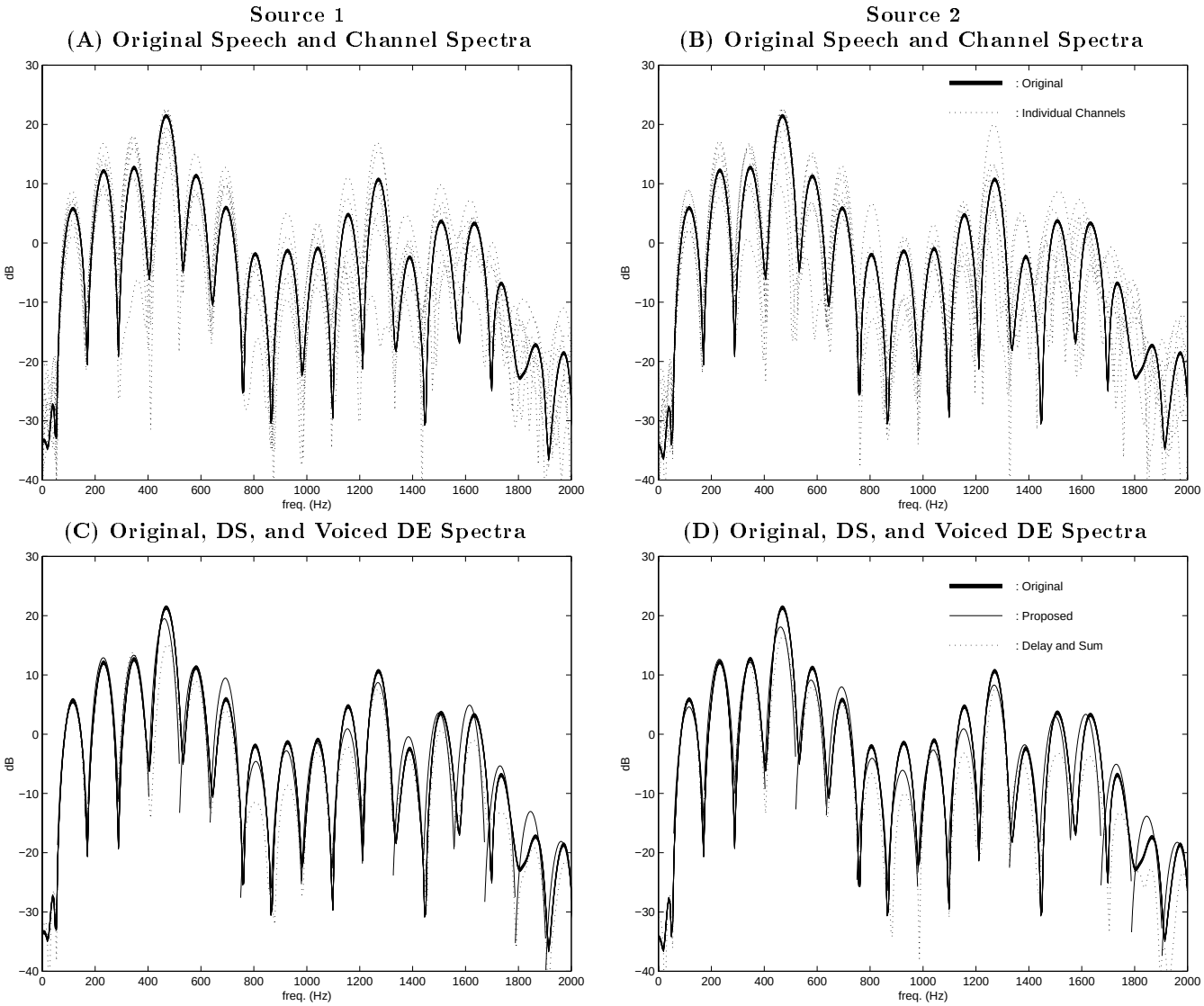

Figure 1. Simulation Results: Spectra of a a voiced speech segment simulated at two closely-spaced source locations.

[7] J. Laroche, Y. Stylianou, and E. Moulines. Hns: Speech modification based on a harmonic + noise model. In *Proceedings of ICASSP93*, pages II–550–II–553. IEEE, 1993.

[8] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744–754, August 1986.

[9] R. Danisewicz and T. Quatieri. An approach to co-channel talker interference suppression using a sinusoidal model for speech. Technical Report 794, Lincoln Laboratory, MIT, Bedford, MA, February 1988.

[10] D. Johnson and D. Dudgeon. *Array Signal Processing-Concepts and Techniques*. Prentice Hall, first edition, 1993.

[11] M. Brandstein and H. Silverman. A practical methodology for speech source localization with microphone arrays. *Computer, Speech, and Language*, 11(2):91–126, April 1997.

[12] J. Flanagan, A. Surendran, and E. Jan. Spatially selective sound capture for speech and audio processing. *Speech Communication*, 13(1-2):207–222, 1993.

[13] S. Affes and Y. Grenier. A signal subspace tracking algorithm for microphone array processing of speech. *IEEE Trans. Speech Audio Proc.*, 5(5):425–437, September 1997.

[14] J. B. Allen and D. A. Berkley. Image method for efficiently simulating small room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, April 1979.