BRIDGING THE GAP: TOWARDS A UNIFIED FRAMEWORK FOR HANDS-FREE SPEECH RECOGNITION USING MICROPHONE ARRAYS

Michael L. Seltzer

Microsoft Research
Redmond, WA, USA
mseltzer@microsoft.com

ABSTRACT

In this paper we describe two families of algorithms for hands-free speech recognition using microphone arrays. Enhancement-based approaches use a cascade of independent processing blocks to perform speech enhancement followed by speech recognition. We discuss the reasons why this approach may be sub-optimal and motivate the need for a solution that tightly integrates all processing blocks into a common unified framework. This leads to a second family of algorithms called unified approaches which considers all processing stages to be components of a single system that operates with the common goal of improved recognition accuracy. We describe several examples of such algorithms that have been shown to outperform more traditional signal-processing-based approaches. In doing so, we hope to convey the benefits of performing hands-free speech recognition in this manner and motivate further research in this area.

Index Terms— microphone array processing, speech recognition, beamforming

1. INTRODUCTION

Current automatic speech recognition (ASR) technology has progressed to the point where applications are being deployed for use in real-world environments. There is a class of these applications for which the use of a close-talking microphone by the user is undesirable, for reasons of either safety or convenience. In such scenarios, because the microphone resides some distance from the user, the captured speech signal is subject to distortions caused by environmental noise and reverberation. In these situations, the use of a microphone array has been proposed as a means of sound capture that can enhance the captured signal, and thereby reduce the detrimental effect that these distortions have on the performance of the downstream speech recognition system.

One of the most challenging aspects of distant-talking speech recognition is variety of deployment environments encountered. For example, a speech recognition system in an automobile, e.g. [1], must be robust to significant amounts of noise but only low reverberation. On the other hand, a meeting room environment typically has a much higher SNR but has moderate to high amounts of reverberation and the additional challenge of overlapping talkers [2]. Information kiosks [3] or mobile devices [4] can be used in highly variable environments which create especially challenging scenarios.

In this paper, we first review the most important and commonly used traditional microphone array processing algorithms. These were all originally developed for signal enhancement applications, and have been applied by many researchers to recognition tasks. Algorithms applied in this way can be considered *enhancement-based approaches* to improving recognition performance. That is, by generating cleaner speech signals, improved speech recognition accuracy can be obtained.

Performing microphone-array-based speech recognition in this manner has been shown to give improvements in recognition accuracy over single channel techniques. However, we argue that because speech enhancement and speech recognition are two fundamentally different tasks, performing hands-free speech recognition using an enhancement-based approach is inherently sub-optimal. This mismatch creates a gap in performance between current performance and what may be possible with a more optimal approach.

The remainder of the paper is spent discussing research that attempts to bridge this gap. We describe recent work from a variety of researchers that employs much tighter integration between the front-end and back-end processing and as a result, is often able to obtain significant improvements over more traditional approaches. We describe these algorithms as *unified approaches* to hands-free speech recognition. We show that these unified approaches can operate at a variety of points along the processing chain, including the beamforming, post-filtering, feature extraction, or decoding stages. Finally, we discuss some computational considerations and some general principles for training and evaluating a speech recognition system for hands-free environments.

2. ENHANCEMENT-BASED APPROACHES TO HANDS-FREE ASR

The most common method of performing speech recognition using a microphone array is to apply either a fixed or adaptive beamforming algorithm to the multi-channel captured audio, followed by a post-filtering operation on the resulting output signal. The output signal from the post-filter is then passed to the recognizer for feature extraction and decoding.

2.1. Fixed beamformers

A fixed beamformer is one whose weights are precomputed and held fixed during deployment. The weights are independent of the observed target and/or interference signals, and depend only on the assumed source and/or interference location. The most common fixed beamformers used for speech recognition applications are the delay-and-sum beamformer (DS) and the superdirective (SD) beamformer [5].

Both of these beamformers can be derived as special cases of the minimum variance distortion response (MVDR) beamformer. In the MVDR beamformer, the weights are chosen so as to minimize the output power of the array subject to a zero-distortion constraint in the look direction. The closed-form MVDR solution can be obtained based solely on knowledge of the look direction and an assumed distribution of the ambient noise. The noise is typically assumed to be zero mean, so its distribution can be completely defined by its covariance matrix.

When a spherical or cylindrical coherence function is substituted for the noise covariance matrix, the MVDR beamformer is equivalent to the superdirective beamformer (SD). When the noise is assumed to be equal power but uncorrelated across microphones, the noise covariance matrix is a scaled version of the identity matrix. Using such a model in the MVDR solution results in the delay-and-sum beamformer.

There are numerous examples of both the delay-and-sum and superdirective beamformers used for speech recognition applications in the literature, e.g. [6].

2.2. Adaptive beamformers

Typically, the location of discrete noise sources, such as a radio or an interfering talker is unknown *a priori*. In these scenarios, it may be advantageous to use adaptive beamformers. The most common adaptive beamformer is the Generalized Sidelobe Canceller (GSC) [7]. The GSC is similar to the fixed MVDR beamformer in that it tries to minimize output power subject to the same distortionless constraint in the direction of the target signal. However, the minimization is done adaptively based on observed samples. This gives the array the ability to steer spatial nulls in the direction of discrete interference sources.

While adaptive beamformers have a theoretical advantage over fixed beamformers, the improvements in recognition accuracy over fixed beamformers is typically not that significant [8]. This may be because of reverberation which breaks the assumption of single source and interference directions, source localization errors which result in signal cancellation, and other reasons [9]. As a result, adaptive beamformers have been applied to speech recognition tasks [10, 11] but are less widespread than their fixed beamformer counterparts.

2.3. Post-filtering

While these beamforming algorithms can reduce the noise in the received signal significantly, they are incapable of removing the noise entirely in any realistic environment. As a result, the output signal generated by a beamforming stage is typically processed with a single channel post-filter. These filters can be based on conventional single channel speech enhancement algorithms, e.g. Wiener filtering or spectral subtraction, or can exploit information from all array channels [6, 12]. In [6], significant gains in recognition accuracy were obtained using a post-filter after a fixed superdirective beamformer.

3. LIMITATIONS OF ENHANCEMENT-BASED APPROACHES

In general, traditional beamformers have been applied successfully to farfield speech recognition tasks, but the improvements in recognition accuracy have not been as great as improvements in signal to noise ratio (SNR) or perceptual quality might indicate. In addition, they have also not performed as well as simple feature enhancement algorithms. In this section, we propose two potential explanations for this sub-optimal performance.

3.1. Beamformer constraints are too strict

Almost all traditional beamformers applied to speech processing and recognition do not utilize any prior information about the source signal. As a result, to ensure there is no distortion caused in the received signal, they impose a requirement that the beamformer must produce unity gain and zero phase distortion in the assumed direction of arrival. While a perfectly reasonable assumption, there are two reasons why it is perhaps too strong for speech recognition applications. First, the features used by most state-of-the-art recognizers, e.g. MFCC or PLP features, are derived from the magnitude of the spectrum only. The phase information is thrown away, and therefore, a requirement of zero phase distortion is actually unnecessary.

In addition, most recognition systems perform some version of Cepstral Mean Normalization (CMN). This technique removes any spectral tilt, i.e. linear channel distortion, from an input sequence of feature vectors. Therefore, the requirement for unity gain on the output is also perhaps unnecessary.

Thus, it is possible that requiring the beamformer to enforce a constraint that has little bearing on recognition accuracy may result in sub-optimal usage of the beamformer parameters. Of course, the proper way to design a beamforming algorithm that suitably removes or weakens these constraints remains an open and interesting area of research.

3.2. Objective functions are mismatched

Most traditional microphone array processing methods were developed as extensions of narrowband signal enhancement algorithms. The goal of these algorithms was to enhance the received signal that was corrupted by noise or multi-path effects during transmission and as such, objective criteria based on SNR, or squared error of the waveform or spectrum was appropriate. On the other hand, speech recognition is a pattern recognition task that uses a maximum likelihood criterion to hypothesize a word sequence based on a set of statistical models (HMMs) and a given sequence of input features derived from the waveform. Such a mismatch in objective functions means that enhancing the output signal does not necessarily improve speech recognition performance. The array algorithm can only be expected to do so if its objective function matches that of the recognizer, i.e. its output generates a sequence of feature vectors that maximizes, or at least increases, the likelihood of the correct transcription, relative to other hypotheses.

4. UNIFIED APPROACHES TO HANDS-FREE ASR

In the previous section, we described potential reasons why traditional enhancement-based approaches may be sub-optimal for speech recognition tasks. We now turn our attention to several recent algorithms that are based on the notion that better performance can be obtained if the different processing blocks involved in a microphone-array-based speech recognition system operate in a more unified manner.

4.1. Array processing methods

As described in Section 3, the optimal set of array parameters are those that result in feature vectors that give the correct word sequence the highest likelihood. Finding this set of parameters is the goal of the LIMABEAM family of algorithms [13, 14]. In these algorithms, the array parameters are chosen so as to maximize the likelihood of the correct word sequence, represented by the most likely state sequence that corresponds to the correct transcription. LIMABEAM uses an unconstrained beamformer optimized using gradient-based techniques.

In reality, the correct state sequence is unknown, so two solutions are proposed. The first uses LIMABEAM as a calibration operation under the assumption that the user speaks an enrollment utterance with a known transcription. In the second approach, the state sequence is treated as a hidden variable and optimized using a generalized EM algorithm. In practice, the E-step is simplified by extracting the single best state sequence based on the current set of array parameters. The M-step then updates the array parameters based on this state sequence.

A speech recognition model is also used for array parameter optimization in the phase-based masking algorithm described in [15]. The Phase Error Filter (PEF) is constructed based on the deviation of the observed phase difference from the expected phase difference for a given direction of arrival. As with other masking algorithms, there is a tradeoff between noise reduction and musical noise, and in the PEF, this tradeoff is controlled by a single parameter. The authors show that different values are optimal for different SNRs and propose a recognizer-based method for selecting the optimal value automatically.

The authors use a Gaussian mixture model (GMM) in the cepstral domain in order to optimize this parameter for a given utterance. The GMM can be considered a simplified speech recognition system, with all Gaussians merged into a single state. As in Unsupervised LIMABEAM, the target state sequence is unknown *a priori* and a generalized EM algorithm is derived. Interestingly, the authors show that convergence can be obtained after only a few iterations, something that was not observed in the LIMABEAM algorithms. This approach is shown to outperform both a generic DS beamformer and the DS beamformer followed by a Wiener post-filter.

4.2. Post-filtering methods

As described in Section 2.3, a post-filter is often applied to the array output signal for further noise suppression. This post-filter would be then followed by feature extraction in speech recognition applications.

However, significant gains can be obtained by replacing the post-filter *before* feature extraction by a feature compensation algorithm *after* feature extraction. Feature compensation algorithms typically compute the MMSE estimate of the clean speech feature vector given the observed noisy feature vector and a prior model of clean speech. Several researchers have shown significant gains in recognition accuracy in array contexts using feature compensation algorithms such as CDCN and VTS [8, 11].

4.3. Acoustic modeling & decoding methods

Finally, a third category of unified approaches to hands-free speech recognition perform the integration in the acoustic models of the recognition engine itself. In this section we describe two different approaches, one that primarily addresses additive noise and one that addresses reverberation.

In [16] an algorithm is described that closely couples a GSC in the acoustic front-end with HMM compensation in the recognizer. This is performed by running a robust GSC in order to generate an enhanced target signal. However, the system also takes the output of the adaptive filters that follow the blocking matrix and uses this as a running estimate of noise in the received signal. Feature extraction is performed on both the estimated target and noise signals, and both sets of features are passed to the the recognizer, where VTS model compensation is performed. This approach is intuitively appealing as it works in the model domain of the recognizer which has been shown to outperform front-end based compensation schemes. It also takes advantage of the fact that the "lower branch" of the GSC generates an accurate estimate of the noise with the source signal attenuated. This approach is shown to provide good performance in a task with overlapping talkers, demonstrating its ability to handle non-stationary interference sources.

All of the methods discussed thus far have obtained improvements in recognition accuracy by more closely aligning the objective functions of multiple processing blocks. In contrast, the algorithm proposed in [17] proposes a change in how a single block operates, namely the HMM decoding. In this work, the convolution of the clean speech and room impulse response that is assumed to occur in reverberant speech is modeled in the mel spectral (melspec) domain. That is, the reverberant melspec features are assumed to have been generated from the convolution of the melspec features of the clean speech and the melspec representation of the room impulse response. The decoding algorithm computes the optimal state sequence over clean speech models given a sequence of melspec feature vectors computed from an observed reverberant signal and the melspec representation of the room impulse response, trained offline prior to decoding.

There are several interesting aspects of this approach. First, like the GSC+VTS algorithm, it does not rely on an estimate of the clean speech features prior to recognition. In addition, small perturbations in the room impulse response which make dereverberation a difficult task in the signal domain are smoothed out by working in the feature domain. The algorithm was evaluated on various rooms with different reverberation times and obtained performance that exceeds conventional decoding using either clean speech models or matched condition models trained on reverberant speech.

5. OTHER CONSIDERATIONS

5.1. Complexity of Unified Approaches

In general, the unified approaches described in Section 4 are significantly more complex than the traditional algorithms in Section 2 both from a computational and implementation standpoint. One reason for the additional computational complexity is that the simple linear relationship between the clean speech, additive noise, and reverberation that exists in the signal domain becomes a highly nonlinear relationship in the feature domain. This typically results in algorithms that require iterative nonlinear optimization methods that use gradients that relate parameters through multiple layers of intermediate variables. The complexity in implementation comes from the simple fact that state-of-the-art speech recognizers are complicated software engineering systems. Implementing an algorithm that is tightly integrated to such a system can be a significant engineering challenge. While both of these are potential limitations of unified approaches, they are also avenues for improvement and further research.

5.2. Training and testing a speech recognizer for distant-talking ASR

Speech recognition systems perform best when the speech observed in deployment is closely matched to that seen in the training data. Because the exact conditions that will occur in deployment are unknown, *multi-style* training is typically performed, where the recognizer is trained using data from a mixture of likely environments. For all recognition tasks, this gives performance that is far superior to that obtained by training the recognizer using clean speech data, e.g. [18]. Collecting actual speech data in the field is of course optimal but this is expensive and time-consuming. Using recorded impulse responses and recorded samples of ambient noise, it is possible to easily create a synthetic training data that will generate good quality acoustic models [19].

In addition, further improvements can be obtained by processing the training data through the same pipeline that will be performed in deployment. For hands-free speech recognition, this may include segmentation (voice activity detection), sound source localization, beamforming, postfiltering, and other stages. This Noise Adaptive Training (NAT) enables the recognizer to learn how to appropriately model the joint effect of all front-end processing on the speech signal [20].

When evaluating an algorithm for hands-free speech recognition, it is important to know how it performs in the context of any other recognizer-based compensation schemes that may be applied. For example, in evaluation systems built for large vocabulary meeting recognition tasks, the processing during decoding typically involves many compensation steps, such as speaker adaptation and vocal tract length normalization [2]. Therefore, it is valuable to compare the performance of new algorithms for hands-free speech recognition to both existing enhancement-based approaches described in Section 2 but also to these well-known recognizer-based compensation schemes. Hands-free algorithms that generate complementary improvements to these existing recognizer-based techniques are especially valuable.

6. CONCLUSION

In this paper we described two families of algorithms for hands-free speech recognition using microphone arrays. *Enhancement-based approaches* use a cascade of independent processing blocks to perform recognition. In contrast, *unified approaches* consider all processing stages to be components of a single system that operates with the common goal of improved recognition accuracy. By operating in this unified manner, these algorithms have bridged the gap between the previously disparate processing stages involved in hands-free speech recognition. In doing so, they have also begun to bridge the gap in performance between hands-free speech recognition systems and close-talking systems. We believe that further research in this direction will lead to continued improvement in hands-free speech recognition performance.

7. REFERENCES

[1] Y. Grenier, “A microphone array for car environments,” in *Proc. ICASSP*, San Francisco, CA, Mar. 1992.

[2] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan, “The AMI system for the transcription of speech in meetings,” in *Proc. ICASSP*, Honolulu, Hawaii, Apr. 2007.

[3] J. L. Gauvain, J. J. Gangolf, and L. Lamel, “Speech recognition for an information kiosk,” in *Proc. ICSLP*, Philadelphia, PA, Oct. 1996, pp. 849–852.

[4] A. Acero, N. Bernstein, R. Chambers, Y.-C. Ju, X. Li, J. Odell, P. Nguyen, O. Scholz, and G. Zweig, “Live search for mobile: web services by voice on the cellphone,” in *Proc. ICASSP*, Las Vegas, NV, Apr. 2008.

[5] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*, Springer-Verlag, New York, 2001.

[6] I. A. McCowan and H. Boulard, “Microphone array postfilter based on noise field coherence,” *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[7] L. J. Griffiths and C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas Propagat.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[8] M. L. Seltzer, *Microphone Array Processing for Robust Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, July 2003.

[9] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, “Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement,” in *Proc. ICASSP*, Phoenix, AZ, May 1999, vol. 5, pp. 2965–2968.

[10] I. A. McCowan, D. C. Moore, and S. Sridharan, “Near-field adaptive beamformer for robust speech recognition,” *Digital Signal Processing*, vol. 12, no. 1, pp. 87–106, Jan. 2002.

[11] W. Herbordt, T. Horiuchi, M. Fujimoto, T. Jitsuhiro, and S. Nakamura, “Noise-robust hands-free speech recognition and communication on PDAs using microphone array technology,” in *Proc. ASRU*, San Juan, Puerto Rico, Nov. 2005.

[12] C. Marro, Y. Mahieux, and K. U. Simmer, “Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering,” *IEEE Trans. Speech Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[13] M. L. Seltzer, B. Raj, and R. M. Stern, “Likelihood maximizing beamforming for robust hands-free speech recognition,” *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 489–498, Sept. 2004.

[14] M. L. Seltzer and R. M. Stern, “Subband likelihood maximizing beamforming for speech recognition in reverberant environments,” *IEEE Trans. Audio Speech Lang. Processing*, vol. 14, no. 6, pp. 2109–2121, Nov. 2006.

[15] G. Shi, P. Aarabi, and H. Jiang, “Phase-based dual microphone speech enhancement using a prior speech model,” vol. 15, no. 1, pp. 109–118, Jan. 2007.

[16] X. Zhao and Z. Ou, “Closely-coupled array processing and model-based compensation for microphone array speech recognition,” vol. 15, no. 3, pp. 1114–1122, Mar. 2007.

[17] A. Sehr, M. Zeller, and W. Kellermann, “Distant-talking continuous speech recognition based on a novel reverberation model in the feature domain,” in *Proc. Interspeech*, Pittsburgh, PA, Sept. 2006.

[18] H. G. Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluations of speech recognition systems under noisy condidions,” in *ISCA ITRW ASR*, Paris, France, Sept. 2000.

[19] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, “Training of HMM with filtered speech material for hands-free recognition,” in *Proc. ICASSP*, Phoenix, AZ, Mar. 1999.

[20] L. Deng, A. Acero, M. Plumpe, and X. Huang, “Large vocabulary speech recognition under adverse acoustic environments,” in *Proc. ICSLP*, Beijing, China, Oct. 2000.