Multichannel Speech Recognition using Distributed Microphone Signal Fusion Strategies

Marek B. Trawicki, Michael T. Johnson, and
An Ji
Marquette University
Department of Electrical and Computer Engineering
Speech and Signal Processing Laboratory
P.O. Box 1881
Milwaukee, WI 53201-1881
{marek.trawicki, mike.johnson, an.ji}@marquette.edu

Tomasz S. Osiejuk
Adam Mickiewicz University
Department of Behavioural Ecology
Umultowska 89, 61-614
Poznán, Poland
http://www.behaecol.amu.edu.pl
osiejuk@amu.edu.pl

Abstract

Multichannel fusion strategies are presented for the distributed microphone recognition environment, for the task of song-type recognition in a multichannel songbird dataset. The signals are first fused together based on various heuristics, including their amplitudes, variances, physical distance, or squared distance, before passing the enhanced single-channel signal into the speech recognition system. The intensity-weighted fusion strategy achieved the highest overall recognition accuracy of 94.4%. By combining the noisy distributed microphone signals in an intelligent way that is proportional to the information contained in the signals, speech recognition systems can achieve higher recognition accuracies.

1. Introduction

Over the past several decades, there has been extensive research in the speech and signal processing community on development of signal enhancement and robust recognition algorithms using both single-channel microphones and microphone arrays. While the current state-of-the-art single-channel methods for speech enhancement [1] and speech recognition [2] work reasonably well for many applications such as hands-free and mobile communication, the performance of the algorithms still deteriorates under highly noisy conditions. On the other hand, the current state-of-the-art microphone array methods for speech enhancement [3] and recognition [4] have shown improvements over single-channel microphones for many applications such as hearing aids since the additional microphones allow the array to better suppress noise from different directions and only focus on the signal of interest. In both single-channel microphones and microphone arrays, the microphones are arranged in a structured microphone environment. Whereas single-channel microphones require the subjects to be situated relatively close to the microphone, microphone arrays [5] need close-spacing of the microphones to satisfy the spatial aliasing criterion and a priori knowledge of the array geometry. Clearly, single-channel microphones and microphone arrays are a restricted domain of possible microphone configurations. Speech enhancement and recognition systems that employ single-channel microphones or microphones arrays are unable to fully exploit all the available acoustic and spatial information from the environment.

Recently, research in signal enhancement and speech recognition has begun to focus more on the larger domain of distributed microphones [5]. Figure 1 illustrates an example of a typical distributed microphone scenario for the general case of omnidirectional sources in a diffuse noise field. Although there is not nearly as much research in this area, distributed microphones are becoming more common in practice for applications such as speaker spotting and tracking systems and generalize both the structured microphone environments of single-channel microphones and microphone arrays to an unstructured microphone environment. Specifically, distributed microphones involve an arbitrary placement of the microphones at potentially far distances from each other and the source with longer, unknown time-delays and larger, unknown signal attenuation. Researchers have found that microphones distributed over a wide area of interest have the potential to better reduce noise in both acoustic signals and feature vectors by exploiting significant acoustic and spatial information of the speech and noise sources [6]. Through the utilization of distributed microphones, speech enhancement and recognition systems can improve over simply selecting the closest microphone in the presence of background noise.

Despite the recent interest in distributed microphone environments, there are currently not any standard state-of-the-art methods for distributed speech enhancement or speech recognition. McCowan and Sridharan [7] performed sub-band processing of microphone array signals for speech recognition by integrating dynamically-weighted models trained on each sub-array frequency bands based on sub-band speech energy. By also using microphone arrays, Seltzer, Raj, and Stern formulated full-band [4] and sub-band [8] beamforming methods for optimally combining microphone array signals to generate the sequence of features that maximize the likelihood of producing the correct hypothesis. In contrast, Shimizu, Kajita, Takeda, and Itakura [9] developed methods using a fixed sound source that perform speech recognition for each microphone and selects the highest likelihood or equally weights and combines the feature vectors from the microphones. The approaches would not work for the generalized case of distributed microphones because of the large spacing between the microphones and varying location of the sound source. As an alternative method, it would be better to combine the distributed microphone signals in an intelligent way that was proportional to the information they contained to achieve higher recognition accuracies.

In this paper, the purpose is to perform speech recognition on vocalizations collected in a distributed microphone environment using a variety of heuristic fusion strategies for channel weighting, including signal amplitude, signal variance, source distance, and distance squared strategies.

2. Distributed microphone corpus

The 8-channel distributed microphone Ortolan Bunting (*Emberiza Hortulana*) unidirectional vocalization corpus was collected from the Passeriformes (song birds) in their natural habitat. The 44.1 kHz, 16-bit wave data files were captured using 6 different microphone systems that each contained 8 channels from the last two breeding seasons through multiple sites at Glesmyra Peatbog in Hedmark County, Norway. For each of the microphone systems, the individual microphones were not distributed in a regular pattern but instead were designed to fit the natural territorial boundaries of the birds. Overall, the distances between the focal bird and distributed microphones varied between 5–50 m with an average distance of 30 m. A reduced repertoire consisting of four different songtypes, *ab*, *cd*, *eb*, and *jufb*, was used for this experiment.

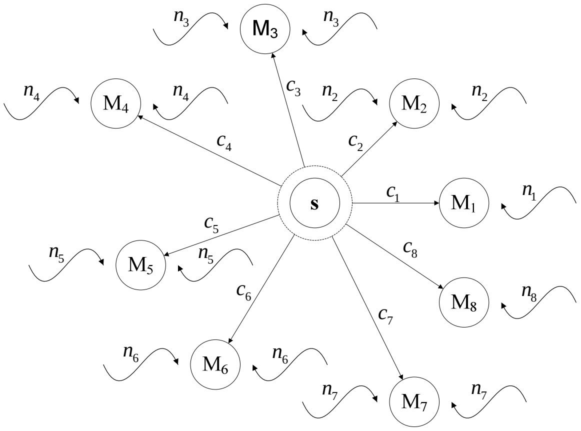

Figure 1. Distributed Microphone Environment

3. Fusion strategies

The general distributed microphone scenario of omni-directional sources is similar to traditional antenna theory [10]. The clean signal $s(t)$ captured at each particular microphone i is correlated across all microphones M , time-delayed by τ_i , and corrupted by additive and uncorrelated noises $n_i(t)$ based on the model

$$y_i(t) = a_i s(t - \tau_i) + n_i(t), \tag{1}$$

where $0 \le t \le T$ and $0 \le i \le M$ denote the time and microphone channels, $y_i(t)$ describes the individual noisy observation signals, and a_i corresponds to the attenuation factors from the source $s(t)$ to the individual microphone signals i . In (1), the attenuation factors a_i measure the decrease in energy of the clean signal $s(t)$ as it propagates through the noisy air medium towards each microphone i . This decrease in energy is proportional to increasing distance from the source because of absorption, scattering, and reflection effects in three dimensions [11]. By assuming *a priori* knowledge of the microphone and source locations, the signals can be time aligned and (1) can be expressed as

$$y_i(t) = a_i s(t) + n_i(t), \qquad (2)$$

where the delay τ_i has been removed, which is necessary for fusing the distributed microphone signals.

Based on the assumption that the clean signal $\hat{s}(t)$ can be estimated as a linear combination of the individual observation signals $y_i(t)$, we can write the estimation formula

$$\hat{s}(t) = \sum_i w_i y_i(t), \qquad (3)$$

where $w_i(t)$ is a microphone-dependent weighting factors. There are several possible heuristic approaches for the weighting factors w_i in the distributed microphone fusion model [12]. As a simple fusion strategy, the individual noisy observation signals $y_i(t)$ can be assigned unity weighting factors

$$w_i = 1, \qquad (4)$$

which assumes that each of the noisy observation signals $y_i(t)$ have equal importance for producing the estimate of the clean signal $s(t)$. The unity weighting factors in (4) treat the cleaner and noisier signals equally, whereas combining the noisy observation signals $y_i(t)$ with unequal weights, where the cleaner signals would have more weight and the noisier signals would have less weight, would be expected to ultimately produce an enhanced clean signal $\hat{s}(t)$ that best reduces the effects of the noises $n_i(t)$ at each of the microphones.

It should be noted that the enhanced clean signal $\hat{s}(t)$ from the microphone fusion model in (3) has the same form as a beamformed output source signal [5]. While the spatial aliasing criterion is not satisfied for multichannel observation signals $y_i(t)$ collected by distributed microphones, the aperiodic and random placement of the microphones eliminates the grating lobes in the radiation pattern and produces a distinct main lobe for the clean signal $s(t)$ of interest analogous to microphone arrays [13]. Since beamforming with microphone arrays provides a platform for the extension to distributed microphones, the weights of the delay-and-sum beamformer serve a comparable role as the weighting factors w_i of the distributed microphones given as

$$w_i = 1/M, \qquad (5)$$

where M is the total number of microphones in the distributed microphone environment. Since scaling has no impact on recognition results, the weighting factors in (5) are equivalent to the unity weighting factors from (4) [5].

Possibilities for weighting schemes could involve many different factors, including distance from the subject to the microphone i, the number of microphones M, and estimated amplitudes and estimated variances of the clean signal $s(t)$ and noisy observation signals $y_i(t)$ [12]. Based on the propagation of the clean signal $s(t)$ to the individual microphones i, the weighting factors $w_i(t)$ can be estimated based on sound pressure (pressure deviation from ambient pressure caused by sound waves), which for omnidirectional sound sources is inversely proportional to physical distance

$$w_i = 1/d_i \qquad (6)$$

or through sound intensity (sound power per unit area), which for omnidirectional sound sources is inversely proportional to the square of the distance [11]

$$w_i = 1/d_i^2 . \qquad (7)$$

As a final set of fusion strategies, the weighting factors w_i for the distributed microphones can be computed through amplitudes and variances of the noisy observation signals $y_i(t)$. For non-omnidirectional sound sources, the measurements of actual sound amplitude and power may be more indicative of sound pressure and sound intensity than physical distances. The approach can be utilized to combine the noisy observation signals $y_i(t)$ through noisy amplitudes

$$w_i = \sqrt{\sigma_{y_i}^2} \qquad (8)$$

or noisy variances

$$w_i = \sigma_{y_i}^2, \qquad (9)$$

which can be estimated directly from the individual channel signals.

4. Experimental results

The weighting factors represented by equations (5), (6), (7), (8), and (9) were independently implemented. These are denoted “equal weighting,” “inverse distance weighting,” “inverse distance squared weighting,” “signal amplitude weighting,” and “signal power weighting.”

The 8-channel distributed microphone noisy observation signals were combined together as described above as a front-end to the speech recognizer for performing song-type classification. Based on results of previous work with song-type classification on single-channel Ortolan Bunting vocalizations [14], the analysis conditions for the distributed microphone corpus were frames of 5 ms with 50% overlap with 12 Generalized Cepstral Coefficient (GFCC) [15] features computed from the 26-channel filterbanks [16] and appended with the delta and delta-delta coefficients. The left-to-right song-type Hidden Markov Models (HMMs) [17] consisted of 18-states with a single diagonal-covariance Gaussian Mixture Model (GMM) underlying each state with approximately an equal split of the four song-types across each of the 8-channel microphones under matched training and testing conditions. HMM implementation was done using the Hidden Markov Model Toolkit (HTK) software toolkit [18].

Recognition accuracy results are shown in Table 1 and Table 2. Accuracies are shown for each syllable (C = Correct and T = Total) and overall. All methods outperform the baseline. The inverse distance squared strategy achieved the highest overall song-type recognition accuracy at 94.4% over simply utilizing the baseline closest channel at 90.7%.. Conversely, the signal power weighting and signal power weighting from (9) and (8) had recognition accuracies of 91.2% and 92.0%, which were the lowest of the fusion strategies. The equal weighting of the noisy observation signals provided surprisingly good recognition results at 93.3%, which was only 1.1% lower than the best results with the inverse distance weighting fusion strategy. Ultimately, the inverse distance squared fusion strategy using the acoustic signals over combining the extracted speech feature vectors obtained the best enhancement and recognition results for the distributed microphone corpus.

5. Conclusion

Distributed microphones generalize single-channel microphones and microphone arrays to unstructured microphone environments and potentially better reduce background noise for speech recognition systems. By comparing various fusion strategies for the distributed microphone experiments, the best way to combine the multichannel signals is through an inverse distance squared strategy with an accuracy of 94.4% for song-type classification. In contrast, the baseline closest single-channel produced an accuracy of 90.7%. Overall, the intelligent combination of distributed microphone signals through heuristic approaches produced higher recognition accuracies than simply selecting the closest channel.

6. Acknowledgements

The authors would like to thank NSF (Grant No. IIS-0326395) and DOE (GAANN Grant P200A010104) for supporting this work.

7. References

[1] R. Martin, "Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 13, pp. 845-856, 2005.

[2] L. Deng, J. Droppo, and A. Acero, "Estimating Cepstrum of Speech Under the Presence of Noise Using a Joint Prior of Static and Dynamic Features," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 218-233, 2004.

[3] B. D. V. Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," in *IEEE ASSAP Magazine*, 1988.

[4] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-Maximizing Beamforming for Robust Hands-Free Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 489-498, 2004.

[5] M. Brandstein and D. Ward, *Microphone Arrays*. New York, NY: Springer-Verlag, 2001.

[6] J. Polastre, R. Szewczyk, and A. Mainwaring, "Chapter 18: Analysis of Wireless Sensor Networks for Habitat Monitoring," in *Wireless Sensor Networks*. Norwell, MA: Kluwer Academic Publishers, 2004.

[7] I. A. McCowan and S. Sridharan, "Microphone Array Sub-Band Speech Recognition," presented at International Conference on Acoustics, Speech, and Signal Processing, 2001.

[8] M. L. Seltzer and R. M. Stern, "Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments," *IEEE Transactions on Audio, Speech, and Language*

Table 1. Fusion Recognition

[redacted]	*AB*		*CD*		*EB*		*JUFB*		**Accuracy**
	C	**T**	**C**	**T**	**C**	**T**	**C**	**T**	
Closest (Baseline)	337	359	267	272	33	69	92	104	90.7%
Signal Amplitude	349	359	249	272	55	69	80	104	91.2%
Signal Power	355	359	249	272	52	69	84	104	92.0%
Inverse Distance	341	359	268	272	41	69	95	104	92.7%
Equal Weighting	350	359	257	272	55	69	88	104	93.3%
Inverse Distance^2	338	359	267	272	60	69	94	104	94.4%

Table 2. Channel Recognition

[redacted]	*AB*		*CD*		*EB*		*JUFB*		**ACCURACY**
	C	**T**	**C**	**T**	**C**	**T**	**C**	**T**	
Closest (Baseline)	337	359	267	272	33	69	92	104	90.7%
Channel #1	207	330	257	272	32	63	35	88	70.5%
Channel #2	183	345	262	272	9	68	78	94	68.3%
Channel #3	191	348	207	266	36	69	71	104	64.2%
Channel #4	191	355	202	260	50	69	83	104	66.8%
Channel #5	157	326	257	272	14	68	57	75	65.5%
Channel #6	194	314	222	272	8	55	42	74	65.2%
Channel #7	130	333	207	272	19	67	70	91	55.8%
Channel #8	46	356	257	272	31	69	30	103	45.5%

Processing, vol. 14, pp. 2109-2121, 2006.

[9] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech Recognition Based on Space Diversity Using Distributed Multi-Microphone," presented at International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, 2000.

[10] B. D. Steinberg, *Microwave Imaging with Large Antenna Arrays*. New York, NY: John Wiley & Sons, 1983.

[11] W. Benenson, J. W. Harris, H. Stocker, and H. Lutz, *Handbook of Physics*. New York City: Springer, 2002.

[12] H. F. Silverman, W. R. P. III, and J. M. Sachar, "An Experiment that Validates Theory with Measurements for a Large-Aperture Microphone Array," presented at International Conference on Acoustics, Speech, and Signal Processing, 2001.

[13] B. D. Steinberg, *Principles of Aperture and Array System Design*. New York, NY: John Wiley & Sons, 1976.

[14] M. B. Trawicki, M. T. Johnson, and T. S. Osiejuk, "Automatic Song-Type Classification and Speaker Identification of Norwegian Ortolan Bunting (Emberiza Hortulana)," presented at IEEE International Conference on Machine Learning in Signal Processing (MLSP), 2005.

[15] P. J. Clemins, M. B. Trawicki, K. Adi, J. Tao, and M. T. Johnson, "Generalized Perceptual Features for Vocalization Analysis Across Multiple Species," presented at International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France, 2006.

[16] P. J. Clemins, "Automatic Classification of Animal Vocalizations," in *Electrical and Computer Engineering*. Milwaukee, WI: Marquette University, 2005.

[17] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, vol. 3, pp. 4-16, 1986.

[18] Cambridge University Engineering Department, *Hidden Markov Model Toolkit (HTK) Version 3.2.1 User's Guide*. Cambridge, MA, 2002.