IDIAP RESEARCH REPORT

A Neural Network based Regression Approach for Recognizing Simultaneous Speech

Weifeng Li [a] Kenichi.Kumatani
John Dines [a] Mathew Magimai.-Doss [a]
Herve Bourlard [a]

IDIAP–RR 08-10

April 2008

submitted for publication

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Av. des Prés–Beudin 20 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 08-10

A Neural Network based Regression Approach for Recognizing Simultaneous Speech

Weifeng Li Kenichi.Kumatani John Dines Mathew Magimai.-Doss
Herve Bourlard

April 2008

submitted for publication

Abstract. This paper presents our approach for automatic speech recognition (ASR) of overlapping speech. Our system consists of two principal components: a speech separation component and a feature estmation component. In the speech separation phase, we first estimated the speaker's position, and then the speaker location information is used in a GSC-configured beamformer with a minimum mutual information (MMI) criterion, followed by a Zelinski and binary-masking post-filter, to separate the speech of different speakers. In the feature estimation phase, the neural networks are trained to learn the mapping from the features extracted from the pre-separated speech to those extracted from the close-talking microphone speech signal. The outputs of the neural networks are then used to generate acoustic features, which are subsequently used in acoustic model adaptation and system evaluation. The proposed approach is evaluated through ASR experiments on the *PASCAL Speech Separation Challenge II* (SSC2) corpus. We demonstrate that our system provides large improvements in recognition accuracy compared with a single distant microphone case and the performance of ASR system can be significantly improved both through the use of MMI beamforming and feature mapping approaches.

1 Introduction

A recent thrust of ASR research has focused on techniques to efficiently integrate inputs from multiple distant microphones (multi-channel) for multiparty meetings (where more than one speakers can be active at the same time). There are two common approaches to the separation of overlapping speech: blind source separation (BSS) [1] and beamforming techniques. BSS exploits the assumption of statistical independence or de-correlated components between the overlapped signals in order to separate them, while beamforming provides an enhanced version of the input speech based on the location of the speakers. The most fundamental and important multi-channel method is the microphone array beamforming method, which consists of enhancing signals coming from a particular location by filtering and combining the individual microphone signals. The simplest technique is *delay-sum* (DS) beamforming, which performs a summation of delayed microphone inputs, where the delays are calculated to compensate for the differing time of arrival of the the desired sound source at each of the microphones in the array.

Other sophisticated beamforming techniques, such as those proposed by Frost [2] or the *Generalized Sidelobe Canceller* (GSC) [3], optimize the beamformer to produce a spatial pattern with a dominant response for the location of interest. The main limitation of these schemes is the issue of signal cancellation. In [4] a superdirective beamformer and a further post-filtering have also been proposed to suppress interfering speech. However, in the case of overlapping speech (with coherent noise), the estimation of coherence matrix is far from trivial, and inaccurate estimations may consequently introduce artifacts into the reconstructed signal. Such disadvantages to conventional beamforming have spurred the development of approaches such as the MMI beamforming criterion for beamforming [13] which alleviates the signal cancellation problem by ensuring the orthogonality of desired and interfering signals.

It is important to note that the motivation behind the microphone array techniques such as delay-sum beamforming is to enhance or separate the speech signals, and as such they are not designed directly in the context of ASR. In practice, it is common for meeting ASR that a well trained acoustic model is first obtained using clean speech data (conversational telephone speech, broadcast news), which is then adapted by using the meeting speech both from close talking microphone (nearfield) as well as distant microphone speech after enhancing the speech by delay-sum beamforming [5] or superdirective beamforming [7]. This approach has been shown to perform well. However, if one looks closely at the ASR errors, a considerable amount of errors occur at the places where speakers overlap (multiple speakers are active) [6]. Thus, improving the signal-to-noise ratio (SNR) of the signal captured through distant microphones may not necessarily be the best means of extracting features for robust ASR on distant microphone data, particularly during periods of speaker overlap.

In our previous work [8], we have proposed to estimate the log spectral energies or Mel-frequency cepstral coefficients (MFCC) of clean speech based on a mapping of delay-and-sum beamformed speech using neural networks. The mapping method could be viewed as a non-linear processing technique that aims to approximate the clean speech through the fusion of the target speech and interfering speech. If the qualities of the estimated target speech and interfering speech are improved, then it is highly possible that the clean speech can be approximated with greater precision. Therefore, we propose to first separate the target speech and the interfering speech using MMI beamforming techniques, followed by a Zelinski and binary-masking based postfilter, and then to perform the mapping method for estimating the MFCCs of the clean speech. Our studies on the PASCAL SSC2 corpus [11] show the effectiveness of the proposed methods.

Although non-linear feature mapping using neural networks has been studied for robust distant microphone ASR in the cepstral domain [17][18][19], the inputs used for the estimation of the clean features in their algorithms are the noisy features obtained from a single input from either distant microphone speech [17][18] or microphone array beamformed speech [19]. We distinguish our approach by exploiting additional sources of information to improve the effectiveness of the mapping. More specifically, we perform a mapping of features of target and interfering sound sources, that have been firstly separated using state of the art beamforming techniques.

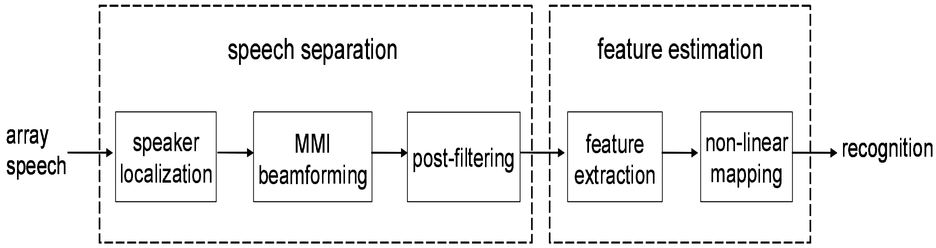

Figure 1: Diagram of system configuration.

This paper is organised as follows. Section 2 presents the system configuration as a whole. Section 3 and Section 4 give a detailed description of speech separation algorithms and non-linear feature estimation using neural networks, respectively. Section 5 then presents and discusses experimental results, and Section 6 gives conclusions and further work which may improve the performance of the system.

2 System Configuration

The system consists of two principal components as shown in Figure 1: speech separation followed by feature estimation. Initially we estimate the speaker's position with the speaker localization system. The speaker location information is used in a GSC-configured beamformer with a minimum mutual information (MMI) criterion to separate the speech of the two speakers, and the non-correlated noise and the competing speech are canceled by means of a Zelinski and binary-masking post-filter applied to the beamformer output. Then the features of the pre-separated speech are extracted and the features of the clean speech are estimated based on a non-linear regression. Finally the estimated features are recognised by the ASR system for evaluation. In the following two sections, the speech separation and feature estimation algorithms are described in details.

3 Speech Separation Algorithms

3.1 Speaker localization

The speaker tracking system we employed was based on [12]. New observations are associated with an active target or with the clutter model through the calculation of posterior probalilities. After the association step, the position of each speaker can be updated through the modified Kalman filter. In addition to the speeaker's position, the system is also capable of determining when each speaker is active.

3.2 MMI beamforming

The speaker location information is used in a GSC-configured beamformer with a minimum mutual information (MMI) criterion [13] to separate the speech of different speakers. Assuming there are two such beamformers aimed at different sources as shown in Figure 2, the output of the i-th beamformer for a given subband can be expressed as,

$$Y_i = \left(\mathbf{w}_{q,i} - \mathbf{B}_i \mathbf{w}_{a,i}\right)^H \mathbf{X} \tag{1}$$

where $\mathbf{w}_{q,i}$ is the *quiescent weight vector* for the i-th source, $\mathbf{B}_i$ is the *blocking matrix*, $\mathbf{w}_{a,i}$ is the *active weight vector*, and $\mathbf{X}$ is the input subband *snapshot vector*. In keeping with the GSC formalism, $\mathbf{w}_{q,i}$ is chosen to preserve a signal from the *look direction* and, at the same time, to suppress an

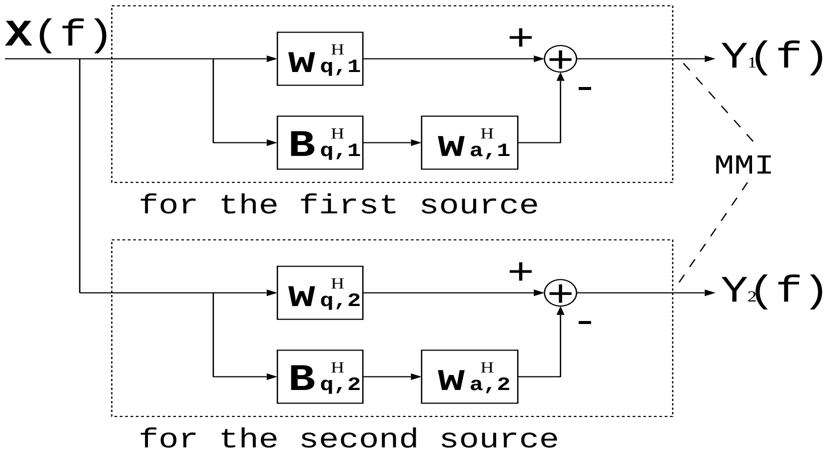

Figure 2: An MMI beamformer in GSC configuration.

interference [15, §6.3]. $\mathbf{B}_i$ is chosen such that $\mathbf{B}_i^H \mathbf{w}_{q,i} = \mathbf{0}$. $\mathbf{w}_{a,i}$ can be optimized by minimizing the mutual information $I(Y_1, Y_2)$ where Y_1 and Y_2 are the outputs of the two beamformers. The optimization procedure of finding that $\mathbf{w}_{a,i}$ under a minimum mutual information (MMI) criterion is described in Kumatani et al [13].

Minimizing a mutual information criterion yields a weight vector $\mathbf{w}_{a,i}$ capable of canceling interference that leaks through the sidelobes without the signal cancellation problems encountered in conventional beamforming. The GSC constraint solves the problems with source permutation and scaling ambiguity typically encountered in conventional blind source separation algorithms [14].

3.3 Post-filtering

In order to further alleviate the non-correlated noise on different channels, a fequency-domain Zelinski post-filter [20] is applied to the MMI-beamformed speech, which can be estimated by

$$\hat{g}_i(f) = \frac{\frac{2}{M(M-1)} \Re\{\sum_i^{M-1} \sum_{j=i+1}^{M} \phi_{x_i x_j}(f)\}}{\frac{1}{M} \sum_i^M \phi_{x_i x_i}(f)}, \tag{2}$$

Here $\Re\{\cdot\}$ and M denote the real operator and the number of channels, respectively. $\phi_{x_i x_i}$ and $\phi_{x_i x_j}$ represent the auto- and cross-spectral densities of the time-aligned inputs, respectively. Furthermore, a frequency-domain binary-masking filter [21]

$$\hat{h}_i(f) = \begin{cases} 1 & \text{if } \; i = \text{argmax}_{i'} |b_{i'}(f)|, i' = 1, ..., I \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $b_i(f)$ is Zelinsky filtered output (in this work $I = 2$), is used to eliminate the signal from competing speakers. Finally the frequency-domain post-filtered output $Z_i(f)$ is obtained by

$$Z_i(f) = \hat{g}_i(f) \hat{h}_i(f) Y_i(f). \tag{4}$$

4 Feature estimation

4.1 Feature extraction

The frequency-domain outputs are reconstructed into time-domain speech signals. The speech signals are extracted with a 25-millisecond window and a 10-millisecond frame shift. 26-channel Mel-filterbank

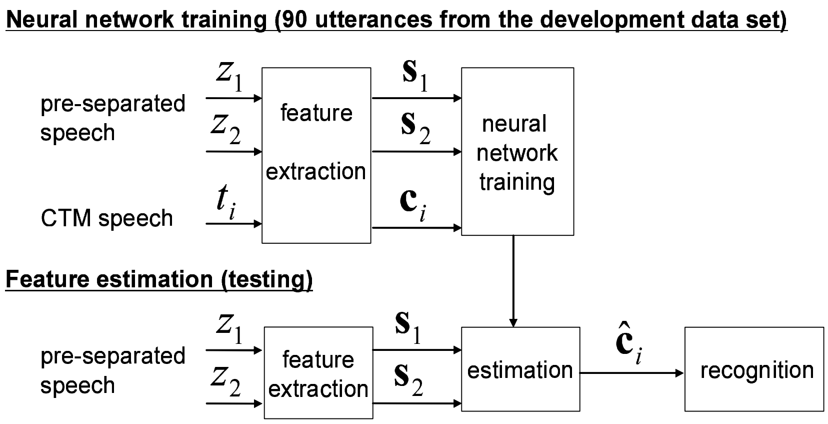

Figure 3: Diagram of the mapping-based speech recognition. CTM: close-talking microphone

analysis followed by the log operation is subsequently applied. Finally 12 Mel-frequency cepstral coefficients (MFCC) are obtained through the discrete cosine transformation (DCT) [9].

4.2 Non-linear mapping

The idea of the mapping method is to approximate the MFCC extracted from the speech signals captured by close-talking microphones through the non-linear combination of the MFCC from pre-separated speech signals, as shown in Figure 3. Let $\mathbf{s}_1(n)$ and $\mathbf{s}_2(n)$ denote the MFCC vectors extracted from the two pre-separated speech signals z_1 and z_2 at frame n, respectively. At the n-th frame the feature vector of the clean speech from the first speaker, $c_1(n)$, can be estimated using the neural network with one hidden layer:

$$\begin{aligned}\hat{\mathbf{c}}_1(n) &= f(\mathbf{s}_1(n), \mathbf{s}_2(n)) \\ &= \sum_{p=1}^{P} \left(w_p \cdot g\left(b_p + \mathbf{w}_{p1}^T \mathbf{s}_1(n) + \mathbf{w}_{p2}^T \mathbf{s}_2(n)\right)\right) + b \end{aligned} \tag{5}$$

where $g(\cdot)$ and P are the sigmoidal activation function and number of the neurons employed in the hidden layer. The clean speech from the second speaker can be estimated by swapping the inputs to the MLP, ie. $\hat{\mathbf{c}}_2(n) = f(\mathbf{s}_2(n), \mathbf{s}_1(n))$.

The parameters $\Theta = \{w_p, b_p, \mathbf{w}_{p1}, \mathbf{w}_{p2}, b\}$ are obtained by minimizing the mean squared error:

$$\mathcal{E}_i = \sum_{n=1}^{N} [\mathbf{c}_i(n) - \hat{\mathbf{c}}_i(n)]^2, \tag{6}$$

over the training examples. Here $\mathbf{c}_i(n), i \in \{1, \ldots, I\}$ denotes the MFCC vector from the ith close talking microphone where in this work $I = 2$. We denote the sample index as n coming from a total of N training examples. The optimal parameters can be found through the error back-propagaton algorithm [16].

Note that the clean speech is required for finding the optimal parameters in the neural network training, while in the test phase the clean speech is no longer required, i.e., it is predicted from the input feature vectors from the enhanced target speech and the interfering speech.

Note that before being fed into MLP, the two pre-separated speech inputs must be kept in a consistent order. We firstly normalize both the pre-separated speech and close-talking microphone speech,

Table 1: Recognition accuracies (as percentages) on the development data set.

	without adaptation	with adaptation
Close-talking microphone	80.6	88.0
Lapel microphone	38.5	67.5
Single distant microphone	0.7	9.4
Separated speech	10.6	35.8
Mapping of separated speech	46.9	58.9
Mapping of lapel microphone speech	70.1	78.8

and then find each of the pre-separated speech near to the corresponding close-talking microphone speech based on the minimum distance between their spectral envelopes. In our mapping method, the inputs of neural networks $\mathbf{s}_1$ and $\mathbf{s}_2$ are 21-dimensional MFCC vectors, while the dimensionality of the output $\mathbf{c}_i$ is 13. These settings are based on our previous studies [10].

5 Experiments and Results

We performed far-field automatic speech recognition experiments on the *PASCAL Speech Separation Challenge 2* (SSC2) [11] corpus. The data contain recordings of two speakers simultaneously and the uttrances is from the 5,000 word vocabulary Wall Street Journal (WSJ) task. The data were recorded with two circular, eight-channel microphone arrays. The diameter of each array was 20 cm, and the sampling rate of the recordings was 16 kHz. The database also contains speech recorded with close talking microphones (CTM). This is a challenging task for source separation algorithms given that the room is reverberant and some recordings include significant amounts of background noise.

Prior to beamforming, we first estimated the speaker's position with the speaker localization system described in [12]. In our beamforming system, the Gaussian pdf is used. The active weights for each subband were nitialized to zero. The system uses an HTK recognizer [9] with acoustic models trained on the WSJCAM0 database from close talking microphones. MLLR based transform is used to adapt the baseline acoustic models.

The corpus is divided into develop data set (178 utterances) and evaluation data set (143 utterances). For the development data set, a leave-one-out cross-validation approach is employed for the adaptation. For the evaluation data set, the development data is used for the adaptation. In our feature mapping method, 104 utterances from the development dataset are used for the neural networking training. The total number of training examples (frames) is 63,826.

Table 2: Recognition accuracies (as percentages) on the evaluation data set.

	without adaptation	with adaptation
Close-talking microphone	82.0	83.1
Lapel microphone	42.1	53.7
Single distant microphone	0.2	1.4
Separated speech	27.9	34.9
Mapping of separated speech	46.7	49.5
Mapping of lapel microphone speech	63.4	68.9

Tables 1 and 2 show recognition accuracies (as percentages) for the development and evaluation data sets for a number of different conditions. We can draw following observations from the results:

- ASR performance drops significantly when going from close-talking microphone, lapel microphone, and a single distant microphone. We also observe the expected results, which have also been earlier observed in the literature [23][5], that model level adaptation improves performance.

- The speech separation system gives much higher recognition accuracies than single distant microphone. However, these results are still much lower than those of the lapel and close-talking microphones.

- By the mapping the MFCCs to those of close-talking microphone, the recognition performance is further significantly increased. Note that without adaptation, the mapping system yields better recognition performance than the lapel microphones, which clearly demonstrates the effectiveness of the feature mapping process. With the mapping of separated speech, the recognition accuracies are higher on the development dataset than on the evaluation data set partly because the training data of neural networks is from one part of development data set.

- When the feature mapping method is applied to the lapel microphones, the recognition performance could also be increased.

6 Conclusions

We have presented our approach to automatically recognize simultaneous speech. Our system consisted of two principal components: a speech separation component which returns the separated speech as well as the locations of simultaneous speakers, and a feature estimation component in which we proposed to further enhance the feature vectors used for speech recognition. The technique achieves better performance to the lapel microphones without acoustic model adaptation, and shows large improvements in recognition accuracy compared with a single distant microphone case. In this work, the mapping was learned between distant microphones signal and clean speech signal. The future work in this direction is to detect speaker overlap and non-overlap regions in multiparty meetings and train/adapt the MLP directly using close-talking microphone speech as target speech.

Acknowledgments.

This work was supported by the European Union 6th FWP IST Integrated Project AMIDA (Augmented Multi-party Interaction with Distant Access, FP6-033812) and the Swiss National Science Foundation through the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2). The authors would like to thank Ivan Himawan, Dr. Iain McCowan, and Dr. Mike Lincoln for the helpful discussions.

References

[1] S. Haykin, Unsupervised adaptive filtering, volume 1, blind source seperation, New York: Wiley, 2000.

[2] O. L. Frost, “An algorithm for linearly constrained adaptive array processing”, Proc. IEEE, Vol. 60, No. 8, pp. 926-935, Aug. 1972

[3] L. J. Griffiths and C. W. Jim, “An Alternative Approach to Linearly Constrained Adaptive Beamforming”, IEEE Trans. on Antennas and Propagation, Vol. AP-30, No. 1, pp. 27-34, Jan. 1982.

[4] D. Moore and I. McCowan, “Microphone array speech recognition: Experiments on overlapping speech in meetings”, In Proc. ICASSP, pp. V:497–500, 2003.

[5] A. Stolcke et al., “The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System”, Lecture Notes in Computer Science, 2007.

[6] O. Cetin and E. Shriberg, “Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap”, In Proc. ICASSP, pp. 1:357-360, 2006.

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

[8] W. Li, M. Magimai.-Doss, J. Dines, and H. Bourlard, “MLP-based log spectral energy mapping for robust overlapping speech recognition”, IDIAP Technical Report, 07-54, 2007.

[9] Steve Young, et. al., The HTK Book, Version 3.4.
http://htk.eng.cam.ac.uk/docs/docs.shtml

[10] W. Li, J. Dines, M. Magimai.-Doss, and H. Bourlard, “Neural Network based Regression for Robust Overlapping Speech Recognition using Microphone Arrays”, IDIAP Technical Report 08-09, 2008.

[11] M. Lincoln, I. McCowan, I. Vepa, and H. K. Maganti, “The multichannel Wall Street Journal audio visual corpus (mc-wsj-av): Specification and initial experiments”, In Proc. ASRU, pp. 357-362, 2005.

[12] T. Gehrig and J. McDonough, “Tracking and far-field speech recognition for multiple simultaneous speakers”, In Proc. the Workshop on Machine Learning and Multimodal Interaction, September 2006.

[13] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wölfel, “Adaptive beamforming with a minimum mutual information criterion”, IEEE Transactions on Audio, Speech and Language Processing, vol. 15, pp. 2527-2541, 2007.

[14] H. Buchner, R. Aichner, and W. Kellermann, “Blind source seperation for convolutive mixtures: A unified treatment”, in Audio Signal Processing for Next-Generation Multimedia Communication Systems, pp. 255-289. Kluwer Academic, Boston, 2004.

[15] H. L. Van Trees, Optimum Array Processing, Wiley-Interscience, New York, 2002.

[16] P.J. Werbos, “Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences”, PhD Thesis, Harvard University, Cambridge, MA, 1974.

[17] H. B. D. Sorensen, “A cepstral noise reductionmulti-layer neural network”, In Proc. ICASSP, vol. 2, pp. 933-936, 1991.

[18] D. Yuk and J. Flanagan, “Telephone speech recognition using neural networks and hiddenMarkov models”, In Proc. ICASSP vol. 1, pp. 157-160, 1999.

[19] C. Che, Q. Lin, J. Pearson, B. de Vries, and J. Flanagan, “ Microphone arrays and neural networks for robust speech recognition”, In Proc. the workshop on Human Language Technology, pp. 342-347, 1994.

[20] K. Uwe Simmer, J. Bitzer, and C. Marro. Post-filtering techniques. In M. Brandstein and D. Ward, editors, Microphone Arrays, chapter 3, pages 39-60. Springer, 2001.

[21] I. McCowan, M. Hari-Krishna, D. Gatica-Perez, D. Moore, and S. Ba, “Speech Acquisition in Meetings with an Audio-Visual Sensor Array”, In Proc. the IEEE International Conference on Multimedia and Expo (ICME), July 2005.

[22] L. R. Rabiner, B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, 1993.

[23] Q. Lin, C. Che, D.-S Yuk, L. Jin, B. de Vries, J. Pearson, and J. Flanagan, “Robust distant-talking speech recognition”, In Proc. ICASSP, pp. 1:21-24 1996.