IDIAP RESEARCH REPORT

Robust overlapping speech recognition based on neural networks

Weifeng Li [a] John Dines [a]
Mathew Magimai.-Doss [a]

IDIAP–RR 07-55

July 2007

revised in July 2007

[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 07-55

Robust overlapping speech recognition based on neural networks

Weifeng Li John Dines Mathew Magimai.-Doss

July 2007

revised in July 2007

Abstract. We address issues for improving hands-free speech recognition performance in the presence of multiple simultaneous speakers using multiple distant microphones. In this paper, a log spectral mapping is proposed to estimate the log mel-filterbank outputs of clean speech from multiple noisy speech using neural networks. Both the mapping of the far-field speech and combination of the enhanced speech and the estimated interfering speech are investigated. Our neural network based feature enhancement method incorporates the noise information and can be viewed as a non-linear log spectral subtraction. Experimental studies on MONC corpus showed that MLP-based mapping techniques yields a improvement in the recognition accuracy for the overlapping speech.

1 Introduction

Recognition of speech in the presence of multiple simultaneous speakers - the so-called 'cocktail party' condition - remains a challenging problem. In such circumstances, headset microphones positioned next to the speakers' mouths have, to date, provided the best recognition performance, however they have a number of disadvantages in terms of cost and ease of use. The alternative is to capture the speech from one or more distant microphones located in the far field, however, such remote microphone recordings generally result in significantly reduced ASR performance.

Recently a thrust of research has focused on techniques to efficiently integrate inputs from multiple distant microphones with the goal of improving ASR performance. The most fundamental and important multi-channel method is the microphone array beamformer method, which consists of enhancing signals coming from a particular location by filtering combining the individual microphone signals. The simplest technique is using the *delay-and-sum* beamformer, which compensates for delays to microphone inputs so that the target signal from a particular direction synchronizes while noises from different directions do not. Other more sophisticated beamforming methods, such as superdirective beamformer [1] and Generalized Sidelobe Canceller (GSC) [2], calculate the filter coefficients to optimize a particular criterion. Although some reports, e.g. [3], has shown that such microphone array techniques can provide improved ASR performance, they generally require a sensitive microphone arrangement, a strict synchronization between channels, and a reliable means of speaker localization. Other multi-channel methods based on *blind source separation* (BSS) [4] or *independent component analysis* (ICA) [5] relies on certain assumptions like statistical independence or de-correlated components, which cannot be guaranteed in most practical situations.

On the other hand, the motivation behind the microphone array and blind source separation techniques is to enhance or separate the speech signals, and they are not designed directly in the context of speech recognition. It is well known that the most widely used front-ends like MFCC in the state-of-the-art speech recognition systems are extracted based on the log mel-filterbank (MFB) outputs [6]. In this work, we will concentrate on multi-channel enhancement of the spectra of target speech for improving the ASR performance in the presence of multiple simultaneous speakers. More specifically, we propose to approximate the log spectral outputs of clean speech by a non-linear combination of the log spectra obtained from multiple distant microphones. In theory, the approach does not require a sensitive microphone arrangement, a strict synchronization between channels, and a reliable speaker-tracking system, and does not need any assumption concerning statistical independence or de-correlated components, either. We also propose to estimate the log spectral outputs of clean speech from those of the multi-channel enhanced speech and the interfering reference, which can be viewed as a highly non-linear log spectral subtraction. The effectiveness of the proposed method is demonstrated in the improvement of word recognition accuracies in different overlapping speech scenarios..

The organization of this paper is as follows: In Section 2, we describe briefly the neural network based mapping approach. Section 3 describes the experimental setup. Section 4 provides the experiments on the mapping of array speech to clean speech. Section 5 presents the experimental studies of the mapping of the enhanced speech to clean speech. In Section 6, we summarize with main conclusions.

2 Algorithms

The idea of the log spectral mapping is to approximate the log mel-filterbank (MFB) vector of clean speech by the non-linear combination of several input speech. Fig. 1 shows the concept of the proposed method. Let $\mathbf{x}_i$ denote the feature vector for the ith distant microphone and $x_{i,k}(n)$ denote the kth element for the frame n. Let $\mathbf{s}$ denote the feature vector obtained from the clean speech and $s_k(n)$ denote the kth element for the frame n. Let $\hat{\mathbf{s}}$ denote the estimated feature vector obtained from the feature vectors of five distant microphones. Each element of $\mathbf{s}$ is approximated independently.

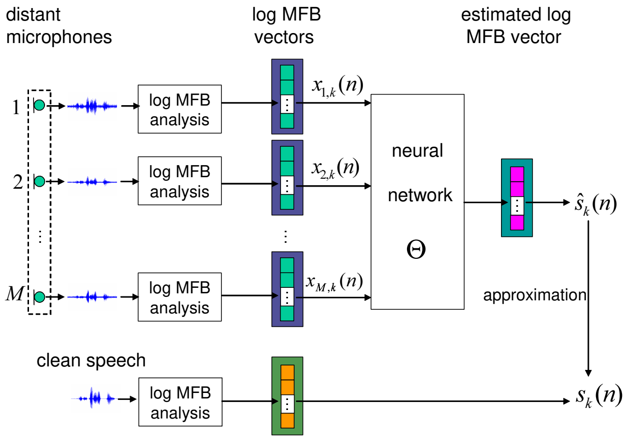

Figure 1: Concept of the log spectral mapping.

Let us introduce an input vector:

$$\mathbf{x}_k(n) = [x_{1,k}(n), x_{2,k}(n), \cdots, x_{M,k}(n)]^T. \tag{1}$$

For multi-layer perceptron (MLP) regression, the network with one hidden layer with P neurons is used. The kth element of the feature vector is estimated by

$$\begin{aligned} \hat{s}_k(n) &= f(\mathbf{x}_k(n)) \\ &= \sum_{p=1}^{P} \left(w_p \tanh\left(b_p + \mathbf{w}^T \mathbf{x}_k(n)\right)\right) + b, \end{aligned} \tag{2}$$

where $\tanh(\cdot)$ is the tangent hyperbolic activation function. Here $\mathbf{w} = [w_{1p}, w_{2p}, \cdots, w_{Mp}]^T$ is the weight vector attached to the pth neurons in the hidden layer. The parameters $\Theta = \{w_p, \mathbf{w}, b_p, b\}$ are found by minimizing the mean squared error:

$$\mathcal{E} = \sum_{n=1}^{N} [s_k(n) - \hat{s}_k(n)]^2, \tag{3}$$

over the training examples. Here, N denotes the number of training examples (frames). The parameters can be updated using a gradient descent algorithm [7]:

$$\Theta = \Theta - \eta \frac{\partial \mathcal{E}}{\partial \Theta}, \tag{4}$$

where η is the learning rate and is set as 0.001 experimentally.

Note that clean speech is required for finding the optimal parameters in the regression training, while in the test phase the clean speech is no longer required. Multiple regression means that regression is performed for each Mel-filter bank. The use of minimum mean squared error (MMSE) in the log spectral domain is motivated by the fact that log spectral measure is more related to the subjective quality of speech [8] and that some better results have been reported with log distortion measures [9][1].

[1] In [9], Porter and Boll found that for speech recognition, minimizing the mean squared errors in the log $|DFT|$ is superior to using all other DFT functions and to spectral magnitude subtraction.

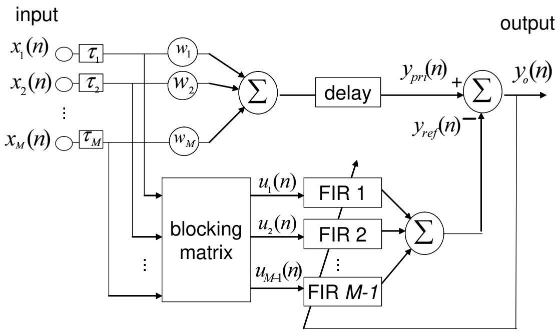

Figure 2: Block diagram of Generalized Sidelobe Canceller.

3 Experimental setup

In order to evaluate the proposed algorithms, the Multichannel Overlapping Numbers Corpus (MONC) [10] was used to perform speech recognition experiments. This database comprises a task for continuous digit recognition in the presence of overlapping speech. The database was collected in a moderately reverberant, 8.2m×3.6m×2.4m rectangular room. Three loudspeakers (L1, L2, L3) were placed at 90deg spacings around the circumference of a 1.2m diameter circular table at an elevation of 35cm. The placement of the loudspeakers simulated the presence of a desired speaker (L1) and two competing speakers (L2 and L3) in a realistic meeting room configuration. An 8-element, equally spaced, circular array of 20cm diameter was placed in the middle of the table, and an additional microphone was placed at the centre of the table. All subsequent discussion will refer to the recording scenarios as S1 (no overlapping speech), S12 (with 1 competing speaker L2), S13 (with 1 competing speaker L3), and S123 (with 2 competing speakers L2 and L3).

The speech recognition experiments were carried out using whole-word HMMs. Each number HMM had 18 states with 16 output distributions. 'sil' had five states with three distributions, and 'sp' had three states with one distribution. Each distribution of a number HMM had 20 Gaussians and that of 'sil' or 'sp' had 36 Gaussians. The duration of analysis window is 20 milliseconds with a frame shift of 10 milliseconds. 23-channel MFB analysis is applied, and the logarithmic outputs of the filterbanks are computed. The estimated log MFB outputs are transformed into 12 mel-frequency cepstral coefficients (MFCCs). The feature vector consisted of 12 MFCCs and log-energy with their corresponding delta and acceleration coefficients. A baseline speech recognition system was trained using HTK on the clean training set from the original Numbers corpus. MAP adaptation was performed on the baseline models using the cross-validation set for each scenario pair, and then the speech recognition performance of the adapted models was assessed using the corresponding recorded test set.

4 Preliminary experiments on the regression-based method

For comparison, we performed the following experiments:

centre recognition of the speech captured by the centre microphone;

DS recognition of the speech enhanced by using delay-and-sum beamformer;

GSC recognition of the speech enhanced by using generalized sidelobe canceller (GSC);

MA recognition of the neural network processed features by mapping 8-channel array speech to clean speech.

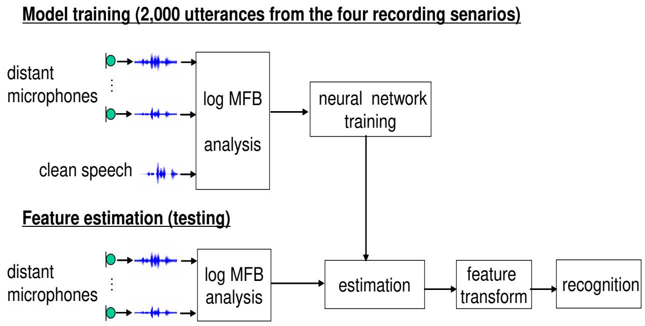

Figure 3: Diagram of the mapping-based speech recognition.

Table 1: Recognition accuracies (as percentages) for different methods.

	S1	S12	S13	S123	Average
centre	78.0	34.5	40.8	24.3	44.4
DS	73.8	46.3	54.7	39.8	53.7
GSC	74.0	49.1	54.2	41.4	54.7
MA	80.0	56.0	65.6	48.2	**62.5**
centre	89.0	38.7	46.9	27.6	50.6
DS	90.4	61.9	70.2	52.8	68.8
GSC	88.3	63.4	68.4	56.7	69.2
MA	84.7	64.9	73.0	54.7	**69.3**

For the "DS" and "GSC", the 8-element circular microphone array was used and the delay is calculated using geometric information of the placement of louderspeaker and microphone array. The architecture of the GSC used is shown in Fig. 2. It comprises a fixed beamformer (top branch), a blocking matrix and three adaptive FIR filters (bottom branch). The top branch produces the beamformed signal which is used as the primary signal. In our experiments, The delay is chosen as half of the adaptive filter order to ensure that the component in the middle of each of the adaptive filters at time n corresponds to $y_{pri}(n)$. The blocking matrix in the bottom branch is used to block out the target signal with which takes the difference between the signals at the adjacent microphones. The three FIR filters are adapted sample-by-sample to generate replicas of the noise or interfering sources involved in the beamformed signal by using Normalized Least Mean Square (NLMS) method [11]. The output $y_o(n)$ takes the form of the subtraction of the interfering replicas $y_{ref}(n)$ from $y_{pri}(n)$. As a result, the target signal is enhanced and the detrimental signals such as ambient noise and interferences are suppressed. In our experiments the number of taps and step-size of adaptation in adaptive beamformer are set as 100 and 0.01 experimentally.

For "MC" and "MA", the training data for neural network consists of 2,000 utterances (500 utterances of each recording scenario in the cross-validation set). The total number of training examples (frames) are 371,543. For a test utterance, the log MFB outputs were first estimated, and then were converted into MFCCs for recognition by using the Discrete cosine transformation (DCT). A diagram of the model training and feature estimation is given in Fig. 3.

Table 1 shows recognition results in terms of recognition accuracies for all channel-scenario pairs. The upper and lower parts of this table depict recognition results without and with the adaption of acoustic models. This table reveals that speech recognition performance degrade significantly in the presence of the interfering speech. For example, with the baseline recognition system using the centre microphone, S12 and S13 result in dramatical reduction of recognition accuracies than S1,

Table 2: Recognition accuracies (as percentages) for different methods.

	S1	S12	S13	S123	Average
DSM	82.5	57.0	69.1	49.7	64.6
GSCM	82.3	60.5	68.7	56.3	66.9
PRM	84.6	63.8	72.1	56.2	**69.2**
PCM	85.6	63.3	73.2	54.4	69.1
DSM	88.8	63.5	73.6	55.8	70.4
GSCM	86.8	66.2	73.2	62.1	72.1
PRM	87.6	69.8	76.1	62.4	74.0
PCM	88.1	70.6	77.4	62.7	**74.7**

and S12 performs worse than S13 because the location of interfering speaker is closer; S123 performs worst as the number of interfering speakers increases. The multi-channel methods (DS and GSC) are effective and outperform the original noisy speech. The mapping of the array speech to the clean speech contributes to the improve the recognition performance, especially in the case of without the adaptation of the acoustic models.

5 Mapping the enhanced speech to clean speech

From the preliminary experiment, we found that the enhanced speech (DS or GSC) based on multi-channel method helps to improve the speech quality which results in the higher recognition performance, as shown in Table 1. An reasonable motivation of further improving of the recognition performance is to map the enhanced speech to clean speech. On the other hand, the output the GSC is obtained by

$$y_o(n) = y_{pri}(n) - y_{ref}(n), \tag{5}$$

where $y_{pri}(n)$ and $y_{ref}(n)$ represent the delay-and-sum (DS) enhanced speech and the reference interfering speech from other directions, respectively. In the frequency domain, it is a simple spectral subtraction [12] which is highly linear. By using our proposed regression method, the estimated log MFB energy of clean speech can be obtained by

$$\hat{s}(n) = f(Y_{pri}(n), Y_{ref}(n)), \tag{6}$$

where $Y_{pri}(n)$ and $Y_{ref}(n)$ represent the corresponding log MFB energies for the DS-enhanced speech and the reference interfering speech, respectively. Here $f(\cdot)$ denotes the mapping function learned by multi-layer perceptron (MLP) networks, which is highly non-linear. In the non-linear mapping, even the $Y_{ref}(n)$ can be replaced by another recorded speech using a distant microphone. In other words, the estimated log MFB energy of clean speech can be obtained by

$$\hat{s}(n) = f(Y_{pri}(n), Y(n)), \tag{7}$$

where $Y(n)$ represent the log MFB energies of the far-field speech. In this way, our neural network based feature enhancement can be viewed as a generalized log spectral subtraction.

For comparison, we performed the following experiments:

DSM recognition of the DS-enhanced speech;

GSCM recognition of the GSC-enhanced speech;

PRM recognition of the neural network processed features by mapping y_{pri} and y_{ref} to clean speech using Equation (6); and

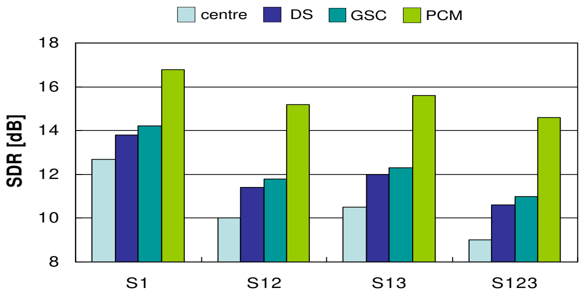

Figure 4: SDR values of different methods.

PCM recognition of the neural network processed features by mapping y_{pri} and the centre microphone speech to clean speech using Equation (7).

Table 2 shows recognition results in terms of recognition accuracies for all channel-scenario pairs. The upper and lower parts of this table depict recognition results without and with the adaption of acoustic models. Mapping the enhanced speech results in considerable improvements compared to "MA" in Table 1. By mapping the enhanced speech and the estimated interfering speech (or another far-field speech), we obtained a further improvement in recognition performance. Compared with GSC as shown in Table 1, PRM results in significant improvement due to the non-linear regression method.

The effectiveness of the approximation is verified from the viewpoint of signal-to-deviation ratio (SDR), which is defined as

$$\mathrm{SDR}\ [\mathrm{dB}] = 10 \log_{10} \frac{\sum_{n=1}^{N} \|\mathbf{s}(n)\|^2}{\sum_{n=1}^{N} \|\mathbf{s}(n) - \hat{\mathbf{s}}(n)\|^2}, \tag{8}$$

where $\mathbf{s}(n)$ is the reference feature vector from the close-talking microphone and $\hat{\mathbf{s}}(n)$ is the estimated feature vector. Here N denotes the number of frames during one utterance. The SDR is averaged over the number of utterances. Fig. 4 shows the average SDR for different methods. First it can been seen that SDR drops as the amount of overlap increases. Secondly, the SDR values for the mapping method are significantly higher than all the non-mapping method. This means a better approximation to the clean speech is obtained, which contributes to the improvement of the recognition performance as shown Table 2.

6 Conclusions

In this work, we investigated the MLP-based feature mapping approach to extract robust MFCCs for multi-channel overlapping speaker speech recognition. We trained an MLP to learn the mapping from log MFBEs of distant microphones speech signal to log MFBEs of clean speech. Experimental studies on MONC corpus showed that MLP-based mapping techniques yields a improvement in the recognition accuracy for the overlapping speech. The future work 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.

Acknowledgements

This work was supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811) and the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2.

References

[1] M. S. Brandstein and D. B. Ward (Eds.), Microphone Arrays: Signal Processing Techniques and Applications, Springer-Verlag, Berlin, 2001.

[2] 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.

[3] D. Moore and I. McCowan, "Microphone array speech recognition: Experiments on overlapping speech in meetings. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. V-497V-500, April 2003.

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

[5] Te-Won Lee, Independent component analysis : theory and applications, Kluwer Academic Publishers (Boston), 1998.

[6] L. R. Rabiner, and B. H. Juang. Fundamental of Speech Recognition, Prentice-Hall, 1993.

[7] S. Haykin, Neural Networks - A Comprehensive Foundation, Prentice-Hall, 1999.

[8] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, Objective Measures of Speech Quality, Prentice-Hall, 1988.

[9] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 18.A.2.1-18.A.2.4, 1984.

[10] The Multichannel Overlapping Numbers Corpus.
www.idiap.ch/ mccowan/arrays/monc.pdf

[11] S. Haykin, Adaptive Filter theory, Prentice Hall, 2002.

[12] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoustics, Speech and Signal Processing, vol.ASSP-27, no.2, pp. 113-120, 1979.