IDIAP RESEARCH REPORT

MLP-based Log Spectral Energy Mapping for Robust Overlapping Speech Recognition

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

IDIAP–RR 07-54

October 2007

submitted for publication

[a] IDIAP Research Institute, Martigny, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), 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 07-54

MLP-based Log Spectral Energy Mapping for Robust Overlapping Speech Recognition

Weifeng Li Mathew Magimai.-Doss John Dines Herve Bourlard

October 2007

submitted for publication

Abstract. This paper investigates a multilayer perceptron (MLP) based acoustic feature mapping to extract robust features for automatic speech recognition (ASR) of overlapping speech. The MLP is trained to learn the mapping from log mel filter bank energies (MFBEs) extracted from the distant microphone recordings, including multiple overlapping speakers, to log MFBEs extracted from the clean speech signal. The outputs of the MLP are then used to generate mel filterbank cepstral coefficient (MFCC) acoustic features, that are subsequently used in acoustic model adaptation and system evaluation. The proposed approach is evaluated through extensive studies on the MONC corpus, which includes both non-overlapping single speaker and overlapping multi-speaker conditions. We demonstrate that by learning the mapping between log MFBEs extracted from noisy and clean signals the performance of ASR system can be significantly improved in overlapping multi-speaker condition compared a conventional delay-sum beamforming approach, while keeping the performance of the system on single non-overlapping speaker condition intact.

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). 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 [1] or the *Generalized Sidelobe Canceller* (GSC) [2], 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 [3] 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.

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 practise, 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 [4]. 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) [5]. 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 the literature, non-linear feature mapping using neural networks has been extensively studied for reducing noise [7], noise robust ASR [8, 9, 12], speaker normalization [10], channel robust ASR [12, 13], robust distant-talking microphone ASR [11, 12]. In these approaches, a neural network is trained to learn the mapping between acoustic features of "noisy" speech to acoustic features of clean speech. In previous work, the mapping was typically performed on cepstral domain features for ASR studies.

In this paper, we investigate the use of neural network based acoustic feature mapping to extract features for robust speech recognition on multiple overlapping speaker distant microphone recordings. In our work, ASR is performed on mel frequency cepstral coefficient (MFCCs) acoustic features, but the mapping is performed on the log mel filter bank (MFBEs) energies, to obtain noise robust estimation of the MFCCs. Thus, we train a multilayer perceptron (MLP) that learns the mapping from log mel filter bank (MFBEs) energies of noisy speech (speech from distant microphones) to the log MFBEs of clean speech.

We have performed our investigations on the Multichannel Overlapping Numbers Corpus (MONC) corpus. Our studies show that by learning the mapping between noisy and clean log MFBEs significant improvement in the ASR performance can be achieved on speaker overlap conditions when compared to MFCCs generated from the DS beamformed speech signal. While we have tried to maintain similar evaluation method as used in previously published results on the MONC corpus [3, 14], we avoid making a direct comparison since any differences in system configuration may unfairly favour one over the other. We do note however that overall our proposed approach does compare favourably with published results.

The paper is organized as follows. In Section 2, we describe briefly the neural network based mapping approach. Section 3 describes the experimental setup. Section 4 provides the experimental results and analysis. In Section 5, we summarize with main conclusions.

2 Neural Network-Based Feature Mapping

The basic idea of feature mapping approach is that given a sequence/set of a pair of feature vectors $(\mathbf{x}_n, \mathbf{s}_n)$, where $n = 1, \cdots, N$ and N is number of pairs, learn a mapping function $f(\cdot)$ such that:

$$\hat{\mathbf{s}}_n = f(\mathbf{x}_n) \tag{1}$$

where $\hat{\mathbf{s}}_n$ is an estimate of $\mathbf{s}_n$. In the neural network-based mapping approach the learning of the mapping function $f(\cdot)$ amounts to training a neural network with $\mathbf{x}_n$ as the input and $\mathbf{s}_n$ as the target output. In our case, the input feature vector $\mathbf{x}_n$ corresponds to the log MFBEs of noisy speech signal (speech from distant microphones) and the target output feature vector corresponds to the log MFBEs of clean speech signal. The neural network is multilayer perceptron (MLP).

Unlike previous approaches, where the mapping of cepstral features have been mainly investigated, we investigate mapping of log MFBEs. We can motivate this from a physiological interpretation of the log spectral energies, but stronger justification can be gained by considering the key properties of the log MFBEs. In particular, with respect to the truncated cepstral representation, the log MFBEs contain highly correlated and redundant information. Such redundancy may be useful when the spectrum contains low SNR in narrowband regions that only strongly affects some of the MFBEs. For the case of overlapping speech, such conditions may arise due to the formant peaks of competing speaker(s). In preliminary experiments we compared the mapping of log MFBEs against the mapping of MFCCs and found that in general the mapping approach shows improvement in ASR performance in both cases, but mapping the log MFBEs was observed to be consistently better. Thus, we train and evaluate our ASR acoustic model on MFCC features estimated from log MFBEs mapped from far-field microphone to clean recording conditions.

There are two possible approaches to learn the mapping between $\mathbf{x}_n$ and $\mathbf{s}_n$:

- Learn a mapping function $f_d(\cdot)$ for each feature component $d = 1, \cdots, D$. In other words, training a mapping neural network for each log MFBE. We refer to this approach as component independent mapping.
- Learn a single mapping function $f(\cdot)$. In other words, training a single neural network that maps all the log MFBEs. We refer to this approach as vector-based mapping.

In our experiments, multi-layer perceptron (MLP) with one hidden layer are used for learning the mapping function $f(\cdot)$ over the training examples with minimum mean squared error (MMSE). The use of MMSE in the log spectral domain is motivated by the fact that log spectral measure is more related to the subjective quality of speech [16] and that some better results have also been reported with log distortion measures [17][1]. 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 i.e. it is predicted from the input log MFBEs from the distant microphones speech.

3 Experimental data and setup

The Multichannel Overlapping Numbers Corpus (MONC) [6] 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

[1] In [17], 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.

subsequent discussions 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. The word models had 16 emitting states, each modelled by a GMM of 20 components. The 'sil' and 'sp' models had three and one emitting state, respectively, with 36 Gaussian mixture components. The duration of the feature analysis is 25 milliseconds with a frame shift of 10 milliseconds. 23-channel log-MFB analysis is applied, which is transformed into 12 mel-frequency cepstral coefficients (MFCCs). Thus, the feature vector comprises 12 MFCCs and log-energy with 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.

The corpus is divided into training data (6049 utterances) and per-condition data sets for development/adaptation (2026 utterances) and testing (2061 utterances). The MLP is trained from data drawn from the development data set (500 utterances per condition).

We performed two standard multichannel ASR experiments:

1. *centre*: Using the MFCCs extracted from the centre microphone speech signal.
2. *DS*: Using the MFCCs extracted from the delay-sum beamformer (DS) enhanced speech signal (standard approach).

When learning the mapping using MLP, the input feature to the MLP can be extracted from a single distant microphone, or all the distant microphones, or a enhanced speech signal or combination of them. We performed the following ASR experiments for the component independent method where an MLP corresponding to each log MFBE component and one MLP for frame level log energy is trained (i.e. 23 + 1 MLPs):

1. *MA*: MFCCs extracted using log MFBEs estimated by mapping MLP that takes log MFBEs extracted from all the 8-channel array speech as input.
2. *MDS*: MFCCs extracted using log MFBEs estimated by mapping MLP that takes log MFBEs extracted from DS-enhanced speech as input.
3. *MDSC*: MFCCs extracted using log MFBEs estimated by mapping MLP that takes log MFBEs of both DS-enhanced speech and centre microphone speech as input.

The size of the MLPs across the different ASR experiments were kept same. We then selected the best performing MLP-based mapping method and compared it against the standard *DS* method using vector-based mapping approach.

4 Results and Analysis

4.1 Component Independent Mapping

Table 1 shows recognition results in terms of recognition accuracies for *center*, *DS* and different methods of the component independent approach. The upper half and lower half of this table depict the recognition results without and with the adaption of acoustic models, respectively. Some of the major observations are:

- ASR performance drops when going from single non overlap speaker condition S1 to overlap speaker conditions S13, S12[2], and S123 with the three speaker overlap condition S123 having the worst performance.

[2]In S12 condition the speakers are more closer than S13 condition which can explain why S12 condition is having lower performance than S13 condition

- Irrespective of the method, mapping approach always yields better performance for all conditions when compared to *center*, and *DS* (except for the S1 condition after adaptation), with the improvements much pronunced in the overlap conditions.

- Straight forward not-so-surprising results which have also been earlier observed in the literature [11, 4] such as model level adaptation improves performance, *DS* is better than *center*, and *MDS* being better than *DS*.

- Among the mapping methods *MDCS* stands out as the best method indicating that while mapping features combining the features from different "versions" of speech signal at the input of the MLP is a good idea.

Table 1: Recognition accuracies (as percentages) of different systems for component independent mapping studies. Upper half of the table represents accuracies for no adaptation case and lower half of the table represents accuracies for adaptation case. The best system based upon average accuracy across all the conditions is in boldface fonts.

	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
MA	80.0	56.0	65.6	48.2	62.5
MDS	82.5	57.0	69.1	49.7	64.6
MDSC	85.6	63.3	73.2	54.4	**69.1**
centre	89.0	38.7	46.9	27.6	50.6
DS	90.4	61.9	70.2	52.8	68.8
MA	84.7	64.9	73.0	54.7	69.3
MDS	88.8	63.5	73.6	55.8	70.4
MDSC	88.1	70.6	77.4	62.7	**74.7**

The effectiveness of the MLP-based mapping approach can also been seen from the viewpoint of signal-to-deviation ratio (SDR), which is defined as

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

where $\mathbf{s}_n$ is the log MFBE feature vector from the clean speech 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. 1 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 all the mapping methods are higher than *DS* and *center*. The highest being for the best performing mapping method *MDSC*.

4.2 Vector-based Mapping

As mentioned earlier in Section 3, we picked the best method for component independent mapping i.e., *MDSC* and extended it to vector-based mapping approach. Figure 2 illustrates the effect of the vector-based mapping method for an utterance in S12 recording scenario. It can be seen that in non-speech segments (e.g., the first and last 15 frames) the interfering speech energies are suppressed by using the mapping method, compared to noisy speech. The vector-based mapping method results in better approximation to the clean speech than the component-based mapping method. In both non-speech and speech frames, vector-based mapping method closely follows the clean speech spectral envelope when compared with the component-based mapping method.

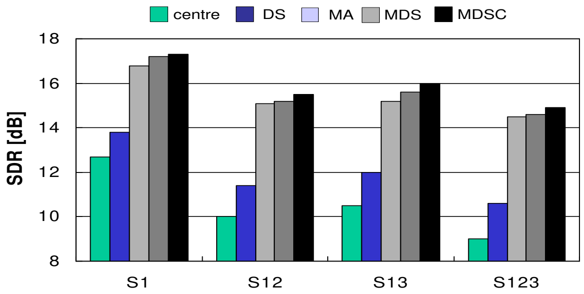

Figure 1: SDR values of different methods.

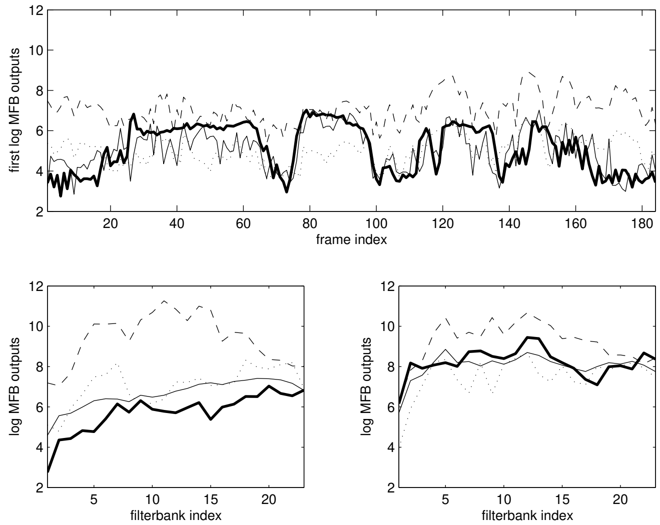

Figure 2: Effect of the mapping method in S12 recording scenario. Upper half: the first log MFBE trajectories of the clean speech signal (bold solid line), centre microphone speech signal (dashed line), component-based mapping (dotted line), and vector-based mapping (thin solid line); Lower half left: log MFBE outputs of different speech signal at the fifth frame (non-speech segment); Lower half right: log MFBE outputs of different speech at the 50th frame (speech segment).

When adapting the acoustic model on a development data the objective is to bring the emission distribution of the acoustic model closer to the "adapted" feature or target feature distribution. The main advantage of adaptation is that the models need not retrained from the scratch. However, when using the MLP-based mapping method it may be possible to improve the match between the trained emission distribution and the unseen test data distribution by extracting the feature for acoustic model training data using the MLP mapping. In order to check it we extracted the feature for the acoustic model clean speech (single speaker) training data by mapping the log MFBEs, followed by estimation of MFCC, and training of the acoustic model. We refer to this approach as *MDSC+FA*.

A similar approach can be applied for delay-sum beamforming *DS* system, where, DS beamforming is performed on S1 condition of acoustic model training data and then the acoustic model is trained. We refer to this approach as *DS2*.

In our MLP-based feature mapping approach, the feature that is being mapped is log MFBE. In the spectral domain since the spectral peaks have high signal-to-noise ratio, one can argue that MLP is learning only spectral peaks mapping very well or picking the spectral peaks. To validate this hypothesis, we used the *DS2* configuration with a modification where the MFCC feature is extracted after scaling the log MFBEs as done in [15]. The key idea of this approach is to emphasize spectral peaks so as to improve the robustness of ASR system towards noise. We refer to this approach as *PP*.

Table 2 shows the recognition performance of the different experiments described above. We can draw following inferences from the results:

- Vector-based mapping approach performs better than component independent mapping approach. This can be due to the fact that MFBEs are correlated and, the ability of MLP to model correlation effectively. It also can be seen that unlike the component independent mapping the performance of vector-based approach for S1 condition is slightly lower or on par with the *DS*.

- For no adaptation case, *DS2* yields a better system when compared to *DS* however, after adaptation the *DS* yields the better system. This can be probably attributed to the fact the *DS2* is only trained on S1 condition data.

- On S1 condition, *PP* yields a performance comparable to *MDSC*, however for rest of the conditions the performance is significantly poorer. This suggests that our approach of mapping log MFBEs using MLP is more than just matching the spectral peaks. Further work needs to be done to understand it very well.

- *MDSC+FA* yields the best system with significant improvement on overlap speech conditions.

Table 2: Recognition accuracies (as percentages) of different systems for vector-based mapping studies. Upper half of the table represents accuracies for no adaptation case and lower half of the table represents accuracies for adaptation case. The best system based upon average accuracy across all the conditions is in boldface fonts.

	S1	S12	S13	S123	Average
MDSC	88.0	76.1	79.4	66.2	77.4
MDSC+FA	88.6	78.9	83.8	72.5	**80.9**
DS2	89.0	57.0	67.7	48.5	65.6
PP	89.5	50.1	62.2	38.8	60.1
MDSC	90.2	76.6	80.1	64.8	77.9
MDSC+FA	89.7	81.9	84.6	75.8	**83.0**
DS2	90.3	59.3	69.5	50.2	67.3
PP	89.7	53.0	64.7	41.7	62.3

5 Summary and 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. We studied two variants of MLP-based mapping, namely, component independent mapping and vector-based mapping. Experimental studies on MONC corpus showed that MLP-based mapping techniques yields a system that is significantly better (particularly for overlap condition) than the one yielded through standard approach of adapting the acoustic model on features extracted from DS beamformed speech signal. The best performance was achieved by the vector-based mapping approach.

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. We will evaluate our method against and in combination with more advanced beamforming/post-filtering microphone array processing techniques.

Acknowledgements

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 (IM)2. The authors would like to thank Prof. B. Yegnanarayana for the helpful discusissions.

References

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

[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. ICASSP, pp. V:497–500, 2003.

[4] A. Stolcke et al., "The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System", To appear in Lecture Notes in Computer Science, 2007.

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

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

[7] S. Tamura and A. Waibel, "Noise reduction using connectionist models", In Proc. ICASSP, pp. 1:553-556, 1988.

[8] H. Sorensen, "A cepstral noise reduction multi-layer neural network", In Proc. ICASSP, pp. 1:933-936, 1991.

[9] L. Barbier and G. Chollet, "Robust speech parameters extraction for word recognition in noise using neural networks", In Proc. ICASSP, pp. 1:145-148, 1991.

[10] X. Huang, "Speaker normalization for speech recognition", In Proc. of ICASSP, pp. 1:465-468, 1992.

[11] 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.

[12] D. Yuk, C. Che, L. Jin and Q. Lin, "Environment-independent speech recognition using neural networks and hidden Markov models", In Proc. ICASSP, pp. 6:3358-3361, 1996.

[13] D. Yuk and J. Flanagan, "Telephone speech recognition using neural networks and hidden Markov models", In Proc. ICASSP, pp. 1:157-160, 1999.

[14] X. Zhao, Z. Ou, M. Chen, Z. Wang, "Closely coupled array processing and model based compensation for microphone array speech recognition", In Proc. of ICASSP, pp. 1:417-420, 2005.

[15] V. Tyagi and C. Wellekens, "On desensitizing the Mel-cepstrum to spurious spectral components for robust speech recognition", In Proc. ICASSP, pp. 529-532, 2005.

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

[17] J. E. Porter and S. F. Boll, “Optimal estimators for spectral restoration of noisy speech”, In Proc ICASSP, pp. 18.A.2.1-18.A.2.4, 1984.