IDIAP RESEARCH REPORT

Neural Network based Regression for Robust Overlapping Speech Recognition using Microphone Arrays

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

IDIAP–RR 08-09

April 2008

submitted for publication

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

IDIAP Research Report 08-09

Neural Network based Regression for Robust Overlapping Speech Recognition using Microphone Arrays

Weifeng Li John Dines Mathew Magimai.-Doss Herve Bourlard

April 2008

submitted for publication

Abstract. This paper investigates a neural network based acoustic feature mapping to extract robust features for automatic speech recognition (ASR) of overlapping speech. In our preliminary studies, we trained neural networks 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. In this paper, we explore the mapping of higher order mel-filterbank cepstral coefficients (MFCC) to lower order coefficients. We also investigate the mapping of features from both target and interfering distant sound sources to the clean target features. This is achieved by using the microphone array to extract features from both the direction of the target and interfering sound sources. We demonstrate the effectiveness of the proposed approach through extensive evaluations on the MONC corpus, which includes both non-overlapping single speaker and overlapping multi-speaker conditions.

1 Introduction

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 pariticular location by combining the individual microphone signals. The simplest technique is the *delay-and-sum* (DS) 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 superdirctive beamformer [1] and Generalized Sidelobe Canceller (GSC) [2], calculate the filter coeffients to optimise a particular criterion.

It is important to note that the motivation behind 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. 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 [4]. This provides ample motivation for the investigation of distant microphone processing techniques that specifically target improvements to ASR performance.

One such approach that has received considerable interest and is the focus of our work uses a neural network (NN), based mapping to obtain 'enhanced' or 'clean' features from the 'noisy' features extracted from the distant microphone recordings. In pointing to previous work on non-linear feature mapping using neural networks for robust distant microphone ASR [9, 10, 11], we note that this work has concentrated solely on the mapping of distant features to clean features. We distinguish our approach by exploiting additional sources of information to improve the effectiveness of the mapping. In our preliminary studies [5], we achieved encouraging results, in particular in the presence of overlapping speech, thus motivating further investigation of this research direction.

In this paper we investigate two additional sources of information that can be exploited in the feature mapping. Firstly, we explore the mapping of higher order mel-filterbank cepstral coefficients (MFCC) to lower order coefficients used in ASR. Typically, such higher order coefficeints are not used in ASR as they contain largely redundant or irrelevant information. We propose that such redundancy can be exploited by the NN to provide added robustness in the presence of overlapping speech. Secondly, we investigate the mapping of features from both target and interfering distant sound sources to the clean target features. This is achieved by using the microphone array to extract features from both the direction of the target and interfering sound sources. Our work is evaluated on the multi-channel numbers corpus (MONC) [6], showing a significant improvement in performance for overlapping speaker conditions.

The paper is organized as follows. In Section 2, we describe briefly the neural network based mapping approach. In Section 3, we describe the experimental setup. In Sections 4 and 5, respectively, we present the experimental studies on the higher order MFCC-based mapping and on the mapping from multiple sound sources. In Section 6, we summarize with main conclusions.

2 Mapping approach

In our mapping approach we take input features extracted from 'noisy' distant microphone recordings (either directly or after microphone array beamforming) and map these to 'clean' recordings. The mapping need not be performed between equivalent features. During training this requires that parallel recordings of clean and noisy data are available while only the noisy features are required during testing. For the non-linear mapping we employ a multilayer perceptron (MLP) with one hidden layer.

Formally, let $\mathbf{d}(n)$ denote the input feature vector of the distant audio at frame n, respectively. At

n-th frame the feature vector of the clean speech $\mathbf{c}(n)$ can be estimated using the MLP:

$$\begin{aligned}\hat{\mathbf{c}}(n) &= f(\mathbf{d}(n)) \\ &= \sum_{p=1}^{P} \left(w_p \cdot g\left(b_p + \mathbf{w}_p^T \mathbf{d}(n)\right)\right) + b\end{aligned} \tag{1}$$

where $g(\cdot)$ and P are the sigmoidal activation function and the number of the neurons employed in the hidden layer. The parameters $\Theta = \{w_p, b_p, \mathbf{w}_p, b\}$ are obtained by minimizing the mean squared error:

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

over the training examples. Here, N denotes the number of training examples (frames). The optimal parameters can be found through the error back-propagaton algorithm [8].

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 array. All 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. Systems were trained using HTK [7] with their respective audio processing and feature extraction front-ends. The training data is equivalent to condition S1 of the development and evaluation sets. MAP adaptation was also performed on these models using the development set for each scenario (thus, each adapted system comprises a set of four models, one adapted to each of the recording scenarios).

The corpus is divided into training data (6049 utterances) and per-condition data sets for development/adaptation (2026 utterances) and testing (2061 utterances). In the feature mapping methods, the MLP is trained from data drawn from the development data set which consists of 2,000 utterances (500 utterances of each recording scenario in the development/adaptation set). The total number of training examples (frames) are 371,543. . A diagram of the model training and feature estimation is given in Figure 1.

4 Mapping of higher order MFCCs

In our previous studies [5], we have proposed to estimate the log mel-filterbank energy (MFBE) vectors of clean speech by mapping those of distant speech. Our investigations of these experiments suggested that in the higher-order filterbanks the estimated log spectral energies could approximate the clean

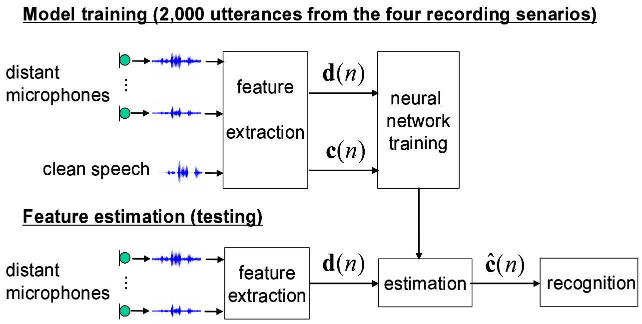

Figure 1: Diagram of the mapping-based speech recognition.

features more accurately than those in the lower-order filterbanks, as shown in Figure 2. This maybe be caused by the more variable properties of lower-order filterbank energies due to the narrower band in the Mel-filterbank design.

An alternative mapping may be performed in the cepstral domain, which is more straightforward in the context of MFCC-based recognizer. For ASR, the MFCCs are usually truncated (e.g. only first 12 MFCCs are used) since it is generally accepted that the higher-order MFCCs provide the detailed harmonics information of log spectrum and yield few benefits for speech recognition. In the context of the neural network learning, however, the detailed information may contribute to the improvement of the performance of the mapping to the clean features, since the MLP may be trained to exploit redundancy in the higher order coefficients while ignoring irrelavent information.

In this section, we perform the studies on the feature mapping in the MFCC domain, and investigate the ASR perfomance when mapping higher order noisy MFCC vectors using input MFCC dimensions of 12, 16, 20 and 23 to clean truncated MFCC vectors of dimension 12. We also compare our results with the best system from our earlier studies in [5], thus the system setup is maintained equivalent; that is, the input to the MLP comprises the feature vectors exctracted from both the delay-sum-beamformer plus the centre microphone. The size of the MLPs across the different ASR experiments were kept the same.

Table 1 shows the recognition results in terms of recognition accuracies for the different experiments described above. 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[1], and S123 with the three speaker overlap condition S123 having the worst performance.
- Expected results which have also been earlier observed in the literature [13, 3] such as model level adaptation improves performance. Additionally we note that adaptation has the greatest impact where there is the greatest mismatch between train and test conditions (ie. condition S123).
- From 12 to 20, recognition accuracies increase as the dimensionalities of MFCCs increase, which illustrates that more detailed information in MFCCs is helpful for neural network learning. However ASR performance drops when the full MFCC vector is used, which indicates thatthe highest order coffecients remain harmful to ASR performance.

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

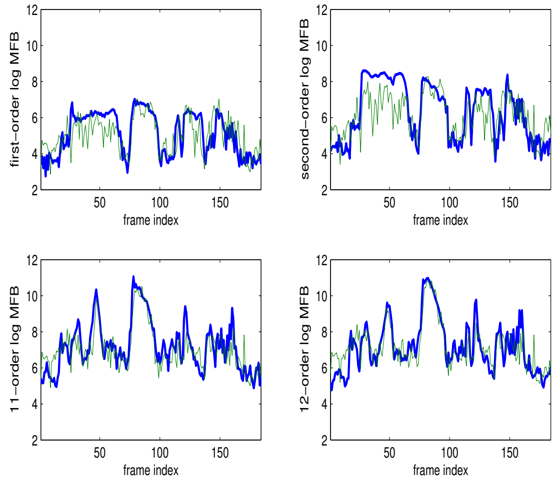

Figure 2: Effect of the mapping method in S12 recording scenario. Upper: the first (left) and second (right) log MFB trajectories of the clean speech signal (bold line) and log MFB based mapping (thin line); Lower: the 11 (left) and 12 (right) log MFB trajectories of the clean speech signal (bold line) and log MFB based mapping (thin line);

- Although MFCC12 does not performs as well as our earlier log MFB experiments, the MFCC20 gives better results, especially for non-adaptation cases. This indicates that the mapping features in MFCC domain with more detailed information is more robust in our speech separation task.

Therefore, MFCC20 based mapping is employed in the following studies.

Table 1: Recognition accuracies (as percentages) of different systems. 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. *log MFBE* refers to the best system from our studies in [5]. Systems are denoted MFCC*XX* where *XX* specifies the dimensionality of the input MFCCs.

	S1	S12	S13	S123	Average
log MFBE	88.6	78.9	83.8	72.5	80.9
MFCC12	88.2	77.5	82.6	71.2	79.9
MFCC16	89.1	79.2	83.6	72.7	81.2
MFCC20	89.2	80.8	84.3	73.6	**82.0**
MFCC23	89.1	80.0	84.1	73.0	81.6
log MFBE	89.7	81.9	84.6	75.8	83.0
MFCC12	89.7	80.4	84.1	74.0	82.1
MFCC16	89.8	81.1	84.5	75.0	82.6
MFCC20	89.9	81.8	85.1	75.9	**83.2**
MFCC23	89.8	81.3	84.5	75.1	82.7

5 Mapping from multiple sound sources

In our preliminary studies [5] we found that augmenting the features to be mapped from the DS beamformer with features extracted from the centre microphone of the array could improve the mapping. We pursue this idea further in these studies by mapping both target and interfering sound sources (the centre microphone signal could be considered a mixture of target and interfering noise). This is achieved in two ways. Firstly, target and interfering audio signals are obtained by directing the DS beamformer in the direction of these sound sources. However, there still remains considerable undesired signal components in the DS outputs, thus, we further process them using a frequency domain maksing post-filter [14] that is applied to the beamformer output to eliminate unwanted sounds. The frequency-domain maksing post-filter is formulated as follows:

- If $b_i(f)$ is the frequency-domain output of beamformer i, the post-filtered output $p_i(f)$ is obtained as:

$$p_i(f) = h_i(f)b_i(f), \tag{3}$$

where the frequency response of the post-filter is estimated by

$$h_i(f) = \begin{cases} 1 & \text{if} \;\; i = \text{argmax}_{i'}|b_{i'}(f)|, i' = 1, ..., I \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and I is the number of beamformers.

In our studies, two beamformers are designed corresponding to the target speech and the interfering speech (In S123 scenario, one beamformer is directed to the target speech and the other directed to the middle position of the two interfering speech.). Note that in S1 scenario (only one speaker), the output of one beamformer is noise-like. Once the two beamformed speech signals are obtained, the 20-order MFCCs are extracted and used as inputs in our mapping method.

We performed the following ASR experiments:

1. *DS*: MFCCs extracted from DS-enhanced speech;

2. *DSmask*: MFCCs extracted from the speech enhanced by the DS and subsequent masking post-filter.

3. *M2DS*: MFCCs estimated by the mapping of the two DS enhanced speech;

4. *M2DSmask*: MFCCs estimated by the mapping of the two DS+masking enhanced speech.

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

- The frequency-domain masking post-filter is very effective at improving the quality of the separated speech (verified by informal listening). In all of the scenarios, the ASR performance is greatly improved by the frequency-domain masking post-filter. Model adaptation is also benefical to increase the recognition accuracies.

- The mapping of the two DS-enhance speech yields significant improvement of ASR performance compared with DS and DSmask (especially without model adaptation), indicating estimating the interfering speech is important for the mapping method. Compared with MFCC20 in the Table 1, the recognition accuracies are improved by 5.0% and 4.3% corresponding to without and with model adaptation, respectively. This suggests that estimating the interfering speech more accurately is helpful for the mapping method.

Table 2: Recognition accuracies (as percentages) of different methods. 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
DS	89.0	57.0	67.7	48.5	65.6
DSmask	89.8	81.7	82.4	69.3	80.8
M2DS	90.6	86.2	88.2	83.1	87.0
M2DSmask	90.3	88.1	88.6	84.1	**87.8**
DS	90.3	59.3	69.5	50.2	67.3
DSmask	90.1	83.0	85.3	74.2	83.2
M2DS	90.8	87.2	88.2	83.8	87.5
M2DSmask	90.5	88.4	88.6	84.1	**87.9**

- Except for the S1 condition, M2DSmask yields the best recognition system on overlap speech conditions illustrating that the frequency-domain masking post-filter is also helpful for the mapping method. However, compared to M2DS the improvement of M2DSmask is marginal, which can be explained by the hypothesis that the MLP is performing a similar role to the masking postfilter, both being provided with essentially the same information as input (source and interfering speech) through in different representation (FFT versus MFCC).

- For M2DS and M2DSmask, the gains of model adaptation is marginal. This may be explained by the fact that the best mapping method evaluated is very effective at suppressing the influence of intefering speakers on the extracted features. Hence, there is much reduced mismatch between the four recording, thus, obviating the need for adaptation to each scenario. scenarios.

6 Conclusions and future works

We have presented our approach to further improve the recognition performance of overlapping speech. The proposed approach achieves large improvements in recognition accuracy in our evaluations on the MONC corpus.

There are several areas in which the further investigations are needed. In the MONC corpus, the clean speech is available, however in real applications the real clean speech is not availalbe, and instead close-talking microphones are usually employed. It is worth investigation the mapping method using CTM speech. We plan to extend this work to more realistic environments, e.g. overlapping speech encounted in meeting scenarios. Future works also lie in the incorporation of more advanced beamforming techniques. On the other hand, the mapping method need not necessarily performed between equivalent features, and the investigation of the data-driven post-filter (in FFT domain) is another furture direction.

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

[4] 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.

[5] Weifeng Li, Mathew Magimai.-Doss, John Dines, and Hervé Bourlard, "MLP-based log spectral energy mapping for robust overlapping speech recognition," IDIAP Technical Report, 07-54, 2007.

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

[7] The Hidden Markov Model Toolkit, http://htk.eng.cam.ac.uk/

[8] S. Haykin, Neural Networks - A Comprehensive Foundation, 2nd edition, Prentice-Hall, 1998.

[9] H. B. D. Sorensen, "A cepstral noise reductionmulti-layer neural network," in Proc. ICASSP, vol. 2, pp. 933âĂŞ936, 1991.

[10] D. Yuk and J. Flanagan, "Telephone speech recognition using neural networks and hiddenMarkov models," in Proc. ICASSP vol. 1, pp. 157âĂŞ160, 1999.

[11] C. Che, Q. Lin, J. Pearson, B. de Vries, and J. Flanagan, " Microphone arrays and neural networks for robust speech recognition," Proceedings of the workshop on Human Language Technology, pp. 342-347, 1994.

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

[13] 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.

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