NON-LINEAR MAPPING FOR MUTLI-CHANNEL SPEECH SEPARATION AND ROBUST OVERLAPPING SPEECH RECOGNITION

Weifeng Li, John Dines, Mathew Magimai.-Doss, and Hervé Bourlard

Idiap Research Institute, CH-1920, Martigny, Switzerland

{wli, dines, mathew, bourlard}@idiap.ch

ABSTRACT

This paper investigates a non-linear mapping approach to extract robust features for ASR and speech separation of overlapping speech. Based on our previous studies, we continue to use two additional sound sources, namely from the target and interfering speakers. The focuses of this work are: 1) We investigate the feature mapping between different domains with the consideration of MMSE criterion and regression optimizations, demonstrating the mapping of log mel-filterbank energies to MFCC can be exploited to improve the effectiveness of the regression; 2) We investigate the data-driven filtering for the speech separation by using the mapping method, which can be viewed as a generalized log spectral subtraction and results in better separation performance. 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.

Index Terms— microphone array, speech separation, binary masking, overlapping speech recognition, neural network

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 particular 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 superdirective beamformer [1] and Generalized Sidelobe Canceller (GSC), 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, which is more serious in the presence of overlapping speech. It is important to note that the motivation behind microphone array techniques such as the beamforming described above 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 signals 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 [2].

While the beamforming methods generally result in a linear transformation, non-linear feature mapping approach using neural networks has received considerable interest for robust automatic speech recognition (ASR). The idea of the feature mapping method is to obtain 'enhanced' or 'clean' features from the 'noisy' features extracted from the distant microphone recordings. In pointing to previous works on multi-channel feature mapping using neural networks for robust ASR (e.g. [3, 4]), we note that a microphone array is used and feature mapping of a delay-and-sum (DS) enhanced speech signal to clean speech signal is performed in MFCC domain. In their mapping framework, a multi-layer perceptron (MLP) was trained for each MFCC component. We distinguish our approach by exploiting the redundant or irrelevant information in a full-vector based mapping and by using additional sources of information to improve the effectiveness of the mapping.

In our previous studies [5, 6], we explored the redundancy of the higher order MFCC vectors to improve the effectiveness of the mapping. We also investigated the mapping of features from both target and interfering distant sound sources, obtained by using the microphone array techniques, to the clean target features. We achieved encouraging results, in particular in the presence of overlapping speech, thus motivating further investigation of this research direction. So far we performed the feature mapping between equivalent domains (e.g. log mel-filterbank energy (MFBE), MFCC). In theory the mapping need not be performed between equivalent domains. In this paper, we firstly investigate the feature mapping between different domains with the consideration of MMSE criterion and regression optimizations. Secondly we investigate the data-driven filtering for the speech separation by using the neural network based mapping method.

2. MAPPING APPROACH

Let's assume that we have both the direction of the target and interfering sound sources, s_t and s_i respectively, through the use of microphone array. Let $\mathbf{s}_t(n)$ and $\mathbf{s}_i(n)$ denote the feature vectors extracted from the target and interfering sound sources at frame n, respectively. In our mapping approach, we take those two input features, $\mathbf{s}_t(n)$ and $\mathbf{s}_i(n)$, and map them to "clean" recordings. To allow nonlinear mapping, we used a generic multilayer perceptron (MLP) with one hidden layer, estimating the feature vector of the clean speech $\mathbf{c}(n)$ associated with the n-th input frame:

$$\begin{aligned}\hat{\mathbf{c}}(n) &= f(\mathbf{s}_t(n), \mathbf{s}_i(n)) \qquad (1)\\ &= \sum_{p=1}^{P} \left(w_p \cdot \text{sig}\left(b_p + \mathbf{w}_{p,t}^T \mathbf{s}_t(n) + \mathbf{w}_{p,i}^T \mathbf{s}_i(n) \right) \right) + b\end{aligned}$$

where $\text{sig}(\cdot)$ 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:

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

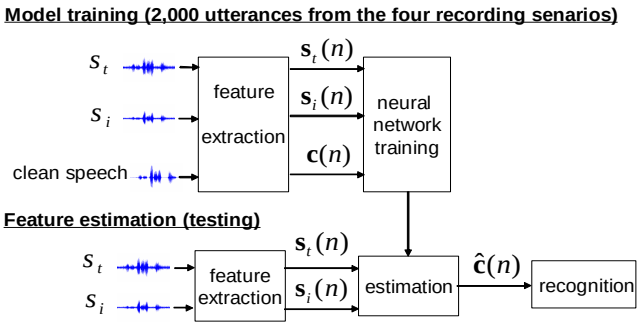

Fig. 1. Diagram of the regression-based speech recognition.

over the training examples. Here, N denotes the number of training examples (frames). The optimal parameters can be found through the error back-propagation algorithm [7]. Note that during training this requires that parallel recordings of clean and noisy data are available while only the noisy features are required for the estimation of clean data during testing.

With the assumption that the distribution of the target data is Gaussian-distributed, minimizing the mean square error in (2) is the result of the principle of maximum likelihood [8]. From the perspective of Blind Source Separation (BSS) and Independent Component Analysis (ICA), the principle of maximum likelihood, which is highly related to the minimization of mutual information between clean sources, can be also employed to estimated the clean sources [9]. Their methods, however, lead to a linear transformation, and the probability densities of the sources must be estimated correctly, while our mapping method is highly non-linear and does not require the information concerning the probability densities of the sources.

3. EXPERIMENTAL DATA AND SETUP

The Multichannel Overlapping Numbers Corpus (MONC) was used to perform speech recognition experiments. In this corpus, there are four recording scenarios [10]: 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 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. In this paper, two delay-and-sum (DS) beamformer enhanced speech signals[1] are used, although DS enhanced speech with a subsequent binary masking post-filter yield a marginal improvement in ASR performance [6]. The ASR frontend generated 12 MFCCs and log-energy with corresponding delta and acceleration coefficients. The more detailed descriptions of recording configurations and speech recognition system can be found in [10] and [6].

[1] 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 speakers.). Note that in S1 scenario (only one active speaker), the output of one beamformer is noise-like.

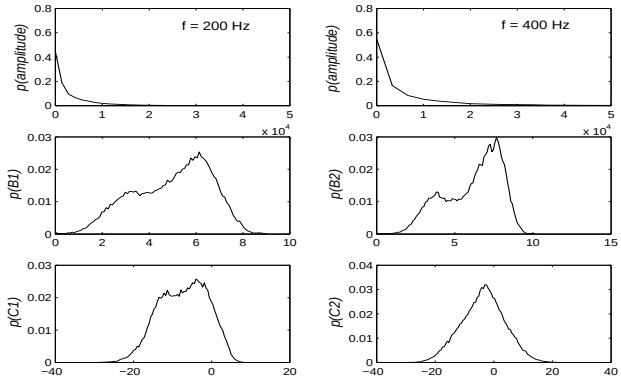

Fig. 2. Probability density functions (pdf) of the different representations of the clean speech. Upper: pdf of the amplitudes at 200 Hz and 400 Hz; Middle: pdf of the first and second order log MFBEs; Lower: pdf of the first and second order MFCCs.

4. FEATURE MAPPING BETWEEN DIFFERENT DOMAINS

In this section, we investigate the mapping method between different domains. We selected the following three domains: spectral amplitude after Fourier transformation, log mel-filterbank energies (log MFBE), and mel-frequency cepstral coefficients (MFCC). As we pointed out in the above section, the target data $\mathbf{c}(n)$ with a Gaussian distribution is optimal from the point view of the minimum mean square error [8]. We investigated the histograms of the features of clean speech as shown 2. It was found that: (1) the PDFs of the amplitudes of the clean speech are far from being Gaussian, as previously widely reported in the literature, (2) the PDFs of the log MFBEs are bi-modal (the lower modal may be due to the low SNR segments), and (3) the PDFs of MFCCs have approximative Gaussian distributions. Therefore we selected the MFCCs as the target domain in our mapping method.

In fact, the mapping to MFCCs is more straightforward in the context of the ASR system,in which MFCCs are used as the features. Furthermore, delta and acceleration MFCCs are usually used in the recognizer. The delta coefficient at frame n are computed using the following regression formula [11]:

$$\mathbf{c}_d(n) = \frac{\sum_\theta \theta[\mathbf{c}(n+\theta) - \mathbf{c}(n-\theta)]}{2\sum_\theta \theta^2} \tag{3}$$

where $\mathbf{c}(n+\theta)$ and $\mathbf{c}(n-\theta)$ denote the corresponding static MFCC vectors at frame $(n+\theta)$ and $(n-\theta)$, respectively. The MMSE of the delta MFCC vectors can be formulated as

$$\begin{aligned}
\mathcal{E}_d &= \sum_n ||\mathbf{c}_d(n) - \hat{\mathbf{c}}_d(n)||^2 \\
&= \frac{\sum_n ||\sum_\theta \theta\{[\mathbf{c}(n+\theta) - \hat{\mathbf{c}}(n+\theta)] + [\mathbf{c}(n-\theta) - \hat{\mathbf{c}}(n-\theta)]\}||^2}{(2\sum_\theta \theta^2)^2} \\
&\approx \frac{\sum_n \sum_\theta \theta^2\{||\mathbf{c}(n+\theta) - \hat{\mathbf{c}}(n+\theta)||^2 + ||\mathbf{c}(n-\theta) - \hat{\mathbf{c}}(n-\theta)||^2\}}{(2\sum_\theta \theta^2)^2} \\
&\approx \frac{2\sum_\theta \theta^2 \cdot \sum_n\{||\mathbf{c}(n+\theta) - \hat{\mathbf{c}}(n+\theta)||^2\}}{(2\sum_\theta \theta^2)^2} \\
&= \frac{\mathcal{E}}{2\sum_\theta \theta^2}
\end{aligned}$$

where we assume $[c(n+\theta) - \hat{c}(n+\theta)]$ and $[c(n-\theta) - \hat{c}(n-\theta)]$ are uncorrelated. Therefore, MMSE in the MFCCs in (2) also

Table 1. Recognition accuracies (as percentages) of the mapping method between different domains. 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
amplitude	90.2	85.8	87.7	83.1	86.7
log MFBE	90.6	87.2	88.9	84.0	**87.7**
MFCC	90.6	86.2	88.2	83.1	87.0
amplitude	90.5	86.4	87.8	83.9	87.2
log MFBE	90.8	88.1	89.2	85.0	**88.3**
MFCC	90.8	87.2	88.2	83.8	87.5

results in MMSE in the delta coefficients (likewise for acceleration coefficients), which can help the ASR performance.

We performed the three corresponding ASR experiments by mapping of amplitudes, log MFBE, and MFCCs of the two DS enhanced speech to MFCCs [2]. Table 1 shows the recognition results in terms of recognition accuracies for the different experiments described above. adaption of acoustic models, respectively. It is found that ASR performance drops when going from single non-overlapping speaker condition S1 to overlapping speaker conditions S13, S12[3], and S123 having the worst performance.

The mapping of the log MFBEs from two DS enhanced speech to MFCCs yields the best ASR performance, especially in overlapping speech scenarios. This may be explained by the fact that the smaller dynamic range of the log MFBE vectors as shown in Figure 2 is advantageous for regression optimization [12]. The gains from model adaptation are marginal. This may be explained by the fact that the mapping methods evaluated are already very effective at suppressing the influence of interfering speakers on the extracted features. Hence, there is much reduced mismatch between the four recording, obviating the need for adaptation to each scenario. Figure 3 shows an example of the mapped MFCC trajectories compared with the ones of the clean speech in S12 recording scenario. It can be seen that the mapping method results in good approximation to the clean speech.

5. REGRESSION-BASED SPEECH SEPARATION

Let s_t and s_i respectively denote the beamformer-enhanced target speech and interfering speech, and let c denote the reference clean speech. By applying a window function and analysis using short-time Fourier transform (STFT), in the time-frequency domain we have complex vector $\mathbf{S}_t(n)$, $\mathbf{S}_i(n)$, and $\mathbf{C}(n)$, where n denotes frame indexes. In order to reduce the dynamic ranges of direct spectral amplitudes, we instead obtain the log amplitudes:

$$\mathbf{S}_t^{(L)}(n) = \log|\mathbf{S}_t(n)|,$$

$$\mathbf{S}_i^{(L)}(n) = \log|\mathbf{S}_i(n)|,$$

$$\mathbf{C}^{(L)}(n) = \log|\mathbf{C}(n)|.$$

[2]When mapping MFCCs to MFCCs, the 20-order MFCCs were used [6]. Through this paper, the size of the MLPs across the different ASR experiments were kept same in this paper. In other words, the total number of parameters in the MLP was set up experimentally to be equal to 10% of the training examples/frames.

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

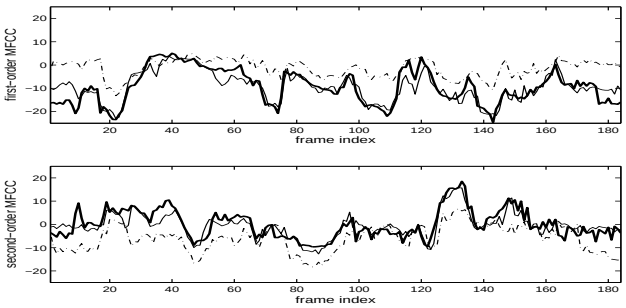

Fig. 3. Effect of the mapping method on the first and second MFCC trajectories in S12 recording scenario. bold solid line: MFCC trajectories of the clean speech; dash-dot line: MFCC trajectories of beamformed speech. thin solid line: the mapped MFCC trajectories from log mel-filterbank energies (log MFBE).

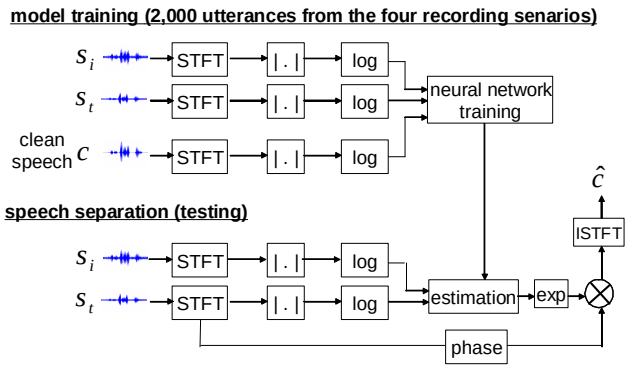

Fig. 4. Diagram of regression-based speech enhancement.

By employing a *multi-layer perceptron* (MLP) regression method, we can obtain the estimated version of clean speech as:

$$\hat{\mathbf{C}}^{(L)}(n) = f(\mathbf{S}_t^{(L)}(n), \mathbf{S}_i^{(L)}(n)) \tag{4}$$

where $f(\cdot)$ is the non-linear function as in (1), which parameters can be optimized according to (2). Once $\hat{C}^{(L)}(n)$ is obtained, separated target speech can be generated by taking the exponential operation and performing inverse short-time Fourier transform (ISTFT) with the combination of the phase of the DS enhanced speech as shown in Figure 4.

The use of MMSE in the log spectral domain is also consistent with the fact that log spectral measure is more related to the physiological interpretation of the log spectral energies and that some better speech enhancement results have been reported with log spectral distortion measures [13]. In [14], a multichannel MMSE estimator of the speech spectral amplitudes was derived for the reduction of uncorrelated noise. However, it can handle additive noise only and has to makes the assumptions regarding the distributions of the speech and noise spectra. The proposed mapping method can be viewed as a generalized log spectral subtraction, and makes no assumptions regarding the additive noise (or interfering speech), nor about any distributions of he speech and noise spectra.

We evaluated the mapped speech signals using source-to-distortion ratio (SDR), which is defined as

$$\text{SDR [dB]} = 10\log_{10}\frac{\sum_{n=1}^{N}\|\mathbf{C}^{(L)}(n)\|^2}{\sum_{n=1}^{N}\|\mathbf{C}^{(L)}(n) - \hat{\mathbf{C}}^{(L)}(n)\|^2}, \tag{5}$$

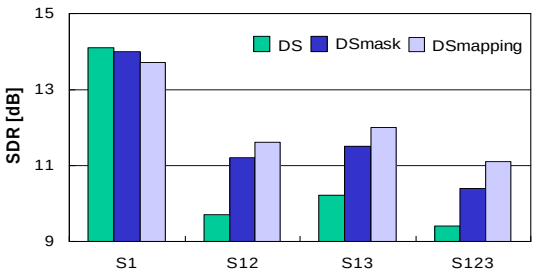

Fig. 5. Diagram of regression-based speech separation.

where $\mathbf{C}^{(L)}(n)$ is the log spectral amplitude vector from the clean speech and $\hat{\mathbf{C}}^{(L)}(n)$ is the estimated version. Here N denotes the number of frames during one utterance. The SDR is averaged over the number of utterances. For comparison we also evaluate the DS-enhanced target speech signals ("DS") and the ones with a subsequent binary masking post-filte ("DSmask") [15]. Note that while "DSmask" provides a hard-decision based on spatial filtering, the proposed method "DSmapping" yields a regression-based soft-decision. Figure 5 shows the average SDR for different methods. Firstly, it can been seen that SDR drops as the amount of overlap increases. Secondly, expect for S1 condition "DSmask" yields significantly higher SDR than "DS", and "DSmapping" obtains highest SDR values, which demonstrates that the mapped speech results in best approximation to the clean speech. Finally, informal listenings show that the interfering speech is well suppressed by using the non-linear regression method. We also evaluated the ASR performance of different speech. As shown in Table 2, in overlapping speech conditions the mapped speech give the better recognition accuracies compared with binary-masked version, especially in S123 conditions. Note that the differences of mapping method between Table 1 and Table 2.

Table 2. Recognition accuracies (as percentages) of different types of speech signals.

	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
DSmapping	88.3	83.5	85.5	80.2	**84.4**
DS	90.3	59.3	69.5	50.2	67.3
DSmask	90.1	83.0	85.3	74.2	83.2
DSmapping	88.6	84.4	86.0	80.6	**84.9**

6. CONCLUSIONS AND FUTURE WORKS

We have presented the mapping approach for the further improvement the recognition performance of overlapping speech and for speech separation. The proposed approach achieves higher recognition accuracies on overlapping multi-speaker conditions while keeping the performance of the ASR system on single non-overlapping condition intact. The separated speech involves less distortions while well suppressing the interfering speech.

Because the MONC database was recorded with loudspeakers, we don't get the natural movements of humans. The training and test data was recorded in the same room and the positions of the speakers (loundspeaker) during the training and test were kept the same. In future we plan to extend this work to more realistic environments (e.g. overlapping speech in meeting scenarios; with different recording rooms and speaker locations) to explore the generalization of our proposed method. On the other hand, we plan to develop the soft masking filter for the speech separation by constraining the regression weights.

7. ACKNOWLEDGEMENTS

This work was supported by the European Union 6th FWP IST Integrated Project on "Augmented Multi-party Interaction with Distant Access" (AMIDA, FP6-033812, www.amida.ch) and the Swiss National Science Foundation through the Swiss National Center of Competence in Research (NCCR) on "Interactive Multi-modal Information Management" (IM2, www.im2.ch).

8. REFERENCES

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

[2] O. Cetin and L. Shriberg, "Speaker Overlaps and ASR Errors in Meetings: Effects Before, During, and After the Overlap", Proc. ICASSP, pp. 357-360, 2006.

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

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

[5] W. Li, M. Magimai.-Doss, J. Dines, and H. Bourlard, "MLP-based log spectral energy mapping for robust overlapping speech recognition," Proc. EUSIPCO, 2008.

[6] W. Li, J. Dines, M. Magimai.-Doss, and H. Bourlard, "Neural network based regression for robust overlapping speech recognition using microphone arrays," Proc. Interspeech, 2008.

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

[8] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[9] A. Hyvärinen and E. Oja. "Independent Component Analysis: Algorithms and Applications," Neural Networks, 13(4-5):411-430, 2000.

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

[11] The HTK Book, http://htk.eng.cam.ac.uk/

[12] M. L. Seltzer, Microphone Array Processing for Robust Speech Recognition, PhD Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, July, 2003.

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

[14] Thomas Lotter, Christian Benien, and Peter Vary, "Multichannel direction-independent speech enhancement using spectral amplitude estimation," EURASIP Journal on Applied Signal Processing, Volume 2003, Issue 11, pp. 1147-1156, 2003.

[15] 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.