COMBINING CEPSTRAL NORMALIZATION AND COCHLEAR IMPLANT-LIKE SPEECH PROCESSING FOR MICROPHONE ARRAY-BASED SPEECH RECOGNITION

Cong-Thanh Do[1], *Mohammad J. Taghizadeh*[2] *and Philip N. Garner*[2]

[1]LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
[2]Idiap Research Institute, CH-1920 Martigny, Switzerland

E-mail: cong-thanh.do@limsi.fr, {mohammad.taghizadeh, phil.garner}@idiap.ch

ABSTRACT

This paper investigates the combination of cepstral normalization and cochlear implant-like speech processing for microphone array-based speech recognition. Testing speech signals are recorded by a circular microphone array and are subsequently processed with superdirective beamforming and McCowan post-filtering. Training speech signals, from the multichannel overlapping Number corpus (MONC), are clean and not overlapping. Cochlear implant-like speech processing, which is inspired from the speech processing strategy in cochlear implants, is applied on the training and testing speech signals. Cepstral normalization, including cepstral mean and variance normalization (CMN and CVN), are applied on the training and testing cepstra. Experiments show that implementing either cepstral normalization or cochlear implant-like speech processing helps in reducing the WERs of microphone array-based speech recognition. Combining cepstral normalization and cochlear implant-like speech processing reduces further the WERs, when there is overlapping speech. Train/test mismatches are measured using the Kullback-Leibler divergence (KLD), between the global probability density functions (PDFs) of training and testing cepstral vectors. This measure reveals a train/test mismatch reduction when either cepstral normalization or cochlear implant-like speech processing is used. It reveals also that combining these two processing reduces further the train/test mismatches as well as the WERs.

Index Terms— Cepstral normalization, Cochlear implant-like speech processing, Kullback-Leibler divergence, Microphone array speech recognition, Overlapping speech

1. INTRODUCTION

Automatic speech recognition (ASR) in adverse environments is difficult because of several variabilities, for instance environmental noises, reverberations, etc. which affect the input speech signal. Assume that the ASR systems use a hidden Markov model (HMM) framework with short-term cepstral vectors; these variabilities would create mismatches between training and testing of ASR. Train/test mismatch can be measured in signal, feature, or model spaces [1]. In [2], train/test mismatch was measured by calculating the Kullback-Leibler divergence (KLD) between the global probability density functions (PDFs), $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$, of cepstral vectors $\mathbf{x}$ extracted from training and testing speech signals, respectively:

$$D(f_{Tr}(\mathbf{x}), f_T(\mathbf{x})) = \int f_{Tr}(\mathbf{x}) \log \frac{f_{Tr}(\mathbf{x})}{f_T(\mathbf{x})} \mathrm{d}\mathbf{x} \qquad (1)$$

Improving speech feature vectors to reduce train/test mismatch is one of the fundamentals to achieve noise robust ASR. In fact, the train/test mismatch, in general, and the KLD between $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$, in particular, could be reduced by several techniques [1]. Cepstral normalization, including ceptral mean normalization (CMN) [3] and cepstral variance normalization (CVN) [4], is such a technique that processes cepstral vectors to reduce the train/test mismatch and, therefore, improve ASR performance. CMN subtracts means of the cepstral coefficients and, therefore, removes the effect of transmission channel on the input speech. Meanwhile, CVN normalizes and scales the variances of testing cepstral coefficients to the variances of the training ones [4, 5]. Hence, it is straightforward that the processing, performed by CMN and CVN, would reduce the KLD between $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$.

In [2], it has been shown that noise robust ASR could be achieved with a framework using cochlear implant-like processing of speech signals. In this framework, the original training and testing speech signals are processed by the cochlear implant-like signal processing algorithm. Original training speech signals were clean and original testing speech signals were noisy. The cochlear implant-like signal processing algorithm synthesizes spectrally reduced speech (SRS) from original training and testing speech signals. The noise robustness of the ASR system using SRS, in both train and test, is better than the baseline system, which uses original speech signals for train and test. Measurements of train/test mismatches, using the KLD between $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$, revealed that the train/test mismatches in the SRS-based ASR systems were smaller than those in the baseline ASR systems. That is, the use of cochlear implant-like processing of speech signals helps in reducing train/test mismatch, measured via the KLD between $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$ [2].

In fact, the combination of cepstral normalization and cochlear implant-like speech signal processing might reduce the train/test mismatch, measured via the KLD between $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$, further than using separately either cepstral normalization or cochlear implant-like speech signal processing. The motivation is as follows. Cochlear implant-like signal processing treats the speech signal in the temporal domain (signal space) whereas cepstral normalization techniques process cepstral vectors in the cepstral domain (feature space). Therefore, the train/test mismatch reduction, performed by the combination of the two techniques, could be complementary.

In this paper, we investigate the combination of cepstral normalization techniques (CMN and CVN) and the cochlear implant-like speech signal processing in order to improve the noise robustness of ASR, more specifically, to improve speech recognition with a microphone array. Microphone arrays have been widely used to capture and pre-process speech signals, especially overlapping speech, which happens in several real situations [6], as a first step to improve speech signal quality for ASR. In our experiments, a microphone array is used to capture speech signals for recognition. The cochlear implant-like speech signal processing and the cepstral normalization are performed afterwards.

The paper is organized as follows. Section 2 introduces the signal processing strategy, which is used to process overlapping speech

captured by a circular microphone array. Section 3 describes the cochlear implant-like speech signal processing algorithm. In section 4, cepstral normalization techniques, as well as their implementation, are described. Section 5 presents the ASR experimental setup, including the ASR system training on the MONC (multichannel overlapping numbers corpus) speech database [7]. After that, experimental results, including the WERs and train/test mismatch measure using the KLDs, are introduced in section 6. Finally, section 7 discusses and section 8 concludes the paper.

2. MICROPHONE ARRAY SIGNAL PROCESSING

Using a microphone array to capture speech signals is practical and helpful in noise and reverberation reduction. This is the first step to enhance the quality of an input speech signal. In fact, it has been shown that using a microphone array is helpful to improve speech recognition performance when there is overlapping speech, compared to the performance of lapel microphone speech recognition [6]. An enhancement-based approach [8] is widely used in microphone array speech recognition. In this approach, either a fixed or adaptive beamforming algorithm [9] is applied to the multi-channel captured speech, and then, a post-filtering operation is applied on the resulting output speech signal.

Assume that the microphone array captures K signals $y_1(n)$, $y_2(n), \ldots, y_K(n)$; the purpose of a beamforming algorithm is to form a beam and point it to a desired direction to extract the speech of interest from these K signals, which are often corrupted by noise, reverberation, and competing sources. This formation is done by, basically, delaying-and-weighting or, more generally, filtering each microphone output $y_k(n)$, and then, summing the delayed-and-weighted or the filtered signals together. In case of filter-and-sum beamformers, which are flexible in controlling the beam pattern as a function of frequency [10], the beamformers output $\widehat{y}(n)$ is calculated as

$$\widehat{y}(n) = \sum_{k=0}^{K-1} \sum_{m=0}^{M-1} y_k(n-m) h_k(m) \qquad (2)$$

where the filters h_k, $k = 1, \ldots, K$, can be either fixed or adaptively determined. Without lost of generality, h_k, $k = 1, \ldots, K$, can be assumed to be FIR filters of length M. In the current paper, we are interested only in fixed beamformers since the locations of the speakers are known beforehand. On the other hand, an advantage of fixed beamformers, for instance the superdirective beamformer, is their low computational complexity and robustness with respect to room reverberation and sensor mismatch, etc. [11]. Furthermore, in terms of ASR performance improvement, fixed beamforming algorithms are comparable to adaptive beamforming algorithms [8].

The directional discrimination time-space filtering of multichannel acquisition results in suppression of the interference sources and thus improves the signal-to-noise ratio (SNR). Beamforming filters are designed based on requirements of the application. The optimal beamforming for maximizing the array-gain is known as the superdirective beamformer. The array-gain is defined as the SNR improvement of the beamformer output with respect to the single channel. Figure 1 illustrates the beam-pattern of a superdirective beamformer in frequencies 250, 500, 1000 and 2246 Hz for a microphone array set-up of our recordings [7]. A speaker is located at azimuth and elevation 135 and 25 degrees with respect to the array center. As the figure shows, the beam pattern is adjusted towards the desired speaker and it is kept the same for all recording scenarios.

The distance between the microphones is equal to the half of the wavelength of the maximum frequency to suppress the majority of the grating lobes. The average SNR of the recordings is 9 dB. In fact, applying a post-filter on the beamformer output signal improves the output signal quality as well as the performance of the ASR system recognizing this signal [12]. In this work, the dominated noise has diffuse characteristics so we use a microphone array post-filter, which was introduced in [12], to improve the output signal of the superdirective beamformer [13]. This post-filter, known as the McCowan post-filter, builds upon the existing Zelinski array post-filter [14] by replacing the assumption of incoherent noise with the assumption of a known noise field coherence function. Based on the knowledge of the noise field coherence function, a more accurate estimate of the signal power density is obtained by solving a set of relevant equations [12]. The signal power density is then used in a Wiener transfer function.

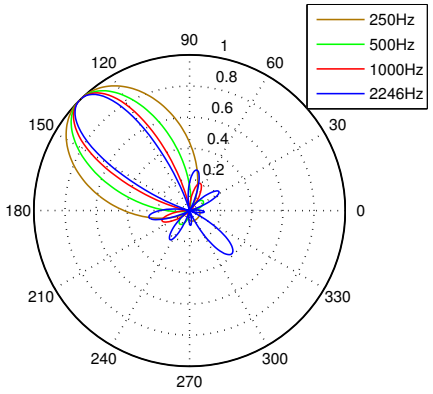

Fig. 1. Beam patterns for a superdirective beamformer in frequencies 250, 500, 1000 and 2246 Hz with a circular microphone array. A speaker is located at azimuth and elevation 135 and 25 degrees with respect to the array center. The beam pattern is adjusted towards the speaking person and it is kept the same for all recording scenarios.

3. COCHLEAR IMPLANT-LIKE SPEECH SIGNAL PROCESSING

It has been shown that processing, simultaneously, enhanced testing speech and clean training speech signals, with the cochlear implant-like speech processing algorithm, helps in improving ASR performance compared to enhancing noisy speech alone [2]. This processing, which results in the re-synthesis of cochlear implant-like spectrally reduced speech (SRS) from subband temporal envelopes of the original speech signal, reduces the train/test mismatch measured via the KLD, as in equation (1) [2]. In the following, we describe the cochlear implant-like speech processing algorithm that has been used to process testing speech, captured by the microphone array, as well as clean training speech. This algorithm is similar to that in [15, 16] and is inspired from the algorithm introduced in [17]. The major difference between the algorithms used in [17] and in [15, 16] lies in the type of carrier signals; subband temporal envelopes in [17] were used to modulate white noise whereas in [15, 16], they were used to modulate sinusoids.

A speech signal $s(n)$ is first decomposed into S subband signals $s_i(n), i = 1, \ldots, S$ by using a perceptually-motivated analysis filterbank consisting of S bandpass filters. The filterbank consists of nonuniform bandwidth bandpass filters which are linearly spaced on the Bark scale in order to simulate the motion of the basilar membrane [18]. In this paper, each bandpass filter in the filterbank is a second-order elliptic bandpass filter having a minimum stopband attenuation of 50 dB and a 2-dB peak-to-peak ripple in the passband. The lower, upper, and central frequencies of the bandpass filters are calculated as in [19]. Figure 2 shows an example of an analysis filterbank consisting of 16 bandpass filters.

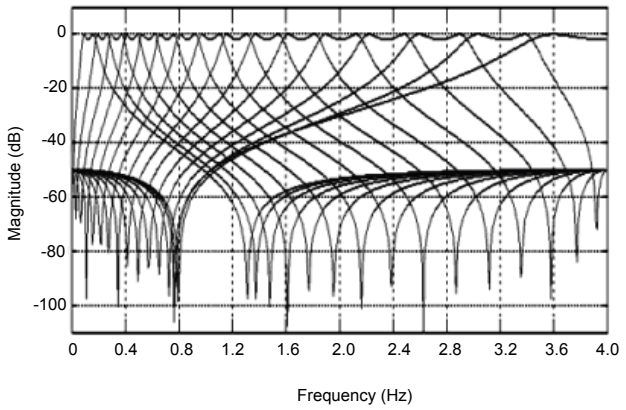

Fig. 2. Frequency response of an analysis filterbank consisting of 16 second-order elliptic bandpass filters. The bandpass filters are linearly spaced on the Bark scale.

The subband temporal envelopes $e_i(n)$ of the subband signals $s_i(n), i = 1, \ldots, S$ are then extracted by, first, full-wave rectification of the outputs of the bandpass filters and, subsequently, lowpass filtering of the resulting signals. These envelopes have the same sampling rate (8 kHz) as that of the subband signal. In this work, the filter that was used to limit the bandwidth of the subband temporal envelopes is a fourth-order elliptic lowpass filter with 2-dB of peak-to-peak ripple and a minimum stop-band attenuation of 50-dB. The subband temporal envelope $e_i(n)$ is then used to modulate a sinusoid whose frequency f_{ci} equals the central frequency of the corresponding analysis bandpass filter of that subband. The subband modulated signal is then filtered again by the same bandpass filter used for the original analysis subband [17]. Finally, all the processed subband signals are summed to synthesize the SRS. The mathematical formula of the SRS $\widehat{s}(n)$ can be expressed as follows:

$$\widehat{s}(n) = \sum_{i=1}^{S} e_i(n) \cos(2\pi f_{ci} n) \tag{3}$$

As reported in [2], using $S = 16$ subbands in the cochlear implant-like processing of training and testing speech signals is relevant to gain noise robust ASR. Indeed, 16 is the spectral resolution from which the SRS signal contains sufficient spectral information compared to the original speech signal [15, 16]. To this end, in this paper, we use $S = 16$ subbands in the cochlear implant-like speech processing algorithm. In addition, for the subband temporal envelopes extraction filter, a 50 Hz cut-off frequency has been used since this cut-off frequency ensures reasonable subband temporal envelope bandwidths, henceforth denoted as W, for human and machine speech recognition with short-term speech features [17, 15, 16].

4. CEPSTRAL NORMALIZATION

Cepstral normalization is a basic and efficient technique to normalize cepstral vectors for gaining ASR noise robustness [5]. In this section, we describe the implementation of cepstral normalization techniques, including cepstral mean normalization (CMN) and cepstral variance normalization (CVN), in the front-end of our ASR systems.

Assume that $\mathbf{X} = [\mathbf{x}_1^T \ \mathbf{x}_2^T \ \ldots \ \mathbf{x}_N^T]$ is a sequence of cepstral vectors extracted from a speech utterance, where $\mathbf{x}_i = [x_{1i}, x_{2i}, \ldots, x_{Li}]^T$ is a L-dimensional cepstral vector extracted from the i-th speech frame, the sequence of CMN vectors $\widetilde{\mathbf{X}}_M$ is calculated as $\widetilde{\mathbf{X}}_M = [\mathbf{x}_1^T - \mu^T \ \mathbf{x}_2^T - \mu^T \ \ldots \ \mathbf{x}_N^T - \mu^T]$, where $\mu = [\mu_1, \mu_2, \ldots, \mu_L]^T$ is a L-dimensional mean vector calculated from all the cepstral vectors extracted from the utterances. In fact, the mean vector μ might be calculated on all the data from a speaker to gain more noise robustness [20]. However, in the current work, the mean vector μ was simply calculated from each speech utterance.

To further normalize the CMN cepstral vectors sequence $\widetilde{\mathbf{X}}_M$ with CVN, the variance of the cepstral coefficients are, first, normalized to 1 based on the local variance $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_L]^T$, and then, scaled to the target variance $\widetilde{\sigma} = [\widetilde{\sigma}_1, \widetilde{\sigma}_2, \ldots, \widetilde{\sigma}_L]^T$ [4, 20]. The matrix form of the sequence of cepstral mean and variance normalization vectors $\widetilde{\mathbf{X}}_{M+V}$ is as follows:

$$\widetilde{\mathbf{X}}_{M+V} = \begin{bmatrix} \frac{(x_{11}-\mu_1)}{\sigma_1}\widetilde{\sigma}_1 & \ldots & \frac{(x_{1i}-\mu_1)}{\sigma_1}\widetilde{\sigma}_1 & \ldots & \frac{(x_{1L}-\mu_1)}{\sigma_1}\widetilde{\sigma}_1 \\ \vdots & & \vdots & & \vdots \\ \frac{(x_{L1}-\mu_L)}{\sigma_L}\widetilde{\sigma}_L & \ldots & \frac{(x_{Li}-\mu_L)}{\sigma_L}\widetilde{\sigma}_L & \ldots & \frac{(x_{LL}-\mu_L)}{\sigma_L}\widetilde{\sigma}_L \end{bmatrix}$$

where the local variance σ is calculated from the cesptral vectors extracted from the given utterance. In our implementation, the target variance $\widetilde{\sigma}$ is the global variance of all the cepstral vectors used in training. Indeed, the variance of the individual cepstra in training and testing are, first, normalized to 1, and then, scaled to the global variance of the training cepstra. Therefore, the Kullback-Leibler divergence (KLD), calculated as in (1), should be reduced.

5. SPEECH RECOGNITION EXPERIMENTS

5.1. Speech database

The multichannel overlapping Numbers corpus (MONC) database [7, 6] was used for the experiments in this paper. The utterances in the Number corpus (30-word vocabulary), which were collected over telephone lines, include isolated digit strings, continuous digit strings, and ordinal/cardinal numbers. To acquire the MONC database, the utterances of the Numbers corpus were played back on one or more loudspeakers, and the resulting sound field was recorded with lapel microphones, a single tabletop microphone, and a tabletop microphone array. The recordings were made in a moderately reverberant 8.2m× 3.6m×2.4m rectangular room. Background noise was made mainly by the PC power supply fan. The loudspeakers were positioned around a circular meeting room table to simulate the presence of 3 competing speakers in the meeting room. The angular spacing between them was 90^o and the distance from table surface to the centre of the main speaker element was 35cm. Lapel microphones were attached to t-shirts hanging below each loudspeaker. The microphone array includes 8 microphones, which were distributed circularly on a 20-cm diameter circle, and was placed in the centre of the table. An additional microphone was placed at the centre of the array. A graphical description of the room arrangement can be found in [7].

Speech signals captured by the microphone array, in three recording scenarios, were used for speech recognition experiments. The three recording scenarios included S_1: there was one speaker (no overlapping speech), $S_{1,2}$: there was one desired speaker and one competing speaker, and $S_{1,2,3}$: there was one desired speaker and two competing speakers. The clean training set was used for training an ASR system. Cochlear implant-like speech signal processing and cepstral normalization were then applied in the front-end, on the speech signals that were captured by the microphone array and pre-processed with superdirective beamforming and post-filtering.

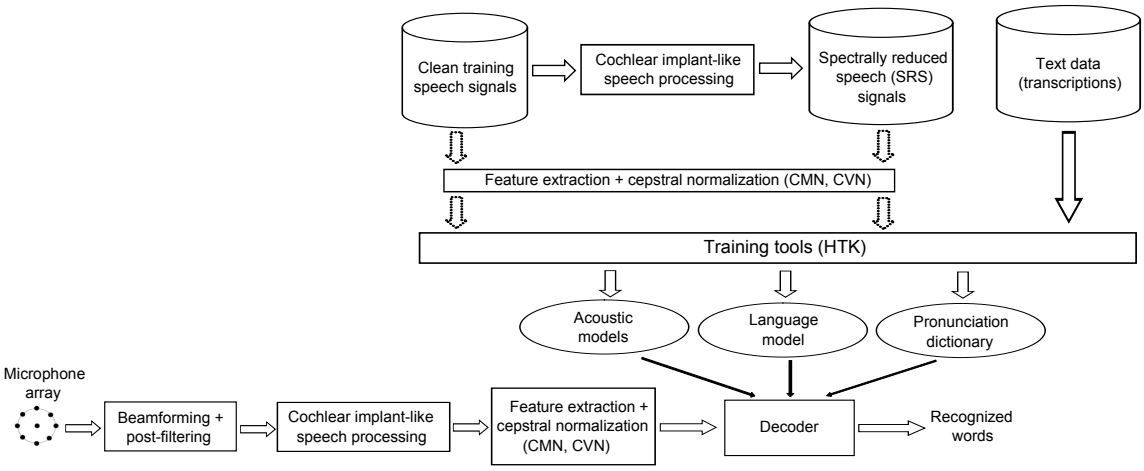

Fig. 3. Training and testing protocols for the microphone array-based speech recognition experiments. Clean speech signals were used for training. In testing, microphone array speech signals were processed with superdirective beamforming and post-filtering. Cepstral normalization techniques (CMN, CVN) were implemented separately or in combination with cochlear implant-like speech signal processing.

5.2. Experimental protocols

An ASR system was trained on the clean training set, including 6049 utterances, of the original Numbers corpus, using the HTK toolkit [20]. The system consists of acoustic models that are tied-state triphone hidden Markov models (HMMs). The triphone HMMs are standard with 3 emitting states per triphone and 12 Gaussian mixtures per state. The system uses 39-dimensional ($L = 39$) speech feature vectors, which consist of 13 Mel frequency cepstral coefficients (MFCCs) (including the 0-th coefficient) along with their delta and acceleration coefficients. This system gave a word error rate (WER) of 6.45% using the clean test set, including 2061 utterances, from the original Numbers corpus.

We implemented cepstral normalization techniques (CMN, CVN) separately or in combination with cochlear implant-like speech signal processing, in the front-end of ASR system, to evaluate the effectiveness of their combination in reducing further the WER of microphone array-based ASR. CMN can be implemented independently or in combination with CVN and cochlear implant-like speech processing. However, CVN was always implemented with CMN or with both CMN and cochlear implant-like speech processing. Testing speech signals were taken from the output of the post-filtering, after the superdirective beamforming processing. Otherwise, training speech signals were clean speech. If either CMN, CVN, cochlear implant-like speech processing or their combinations had been implemented in training, they were implemented in testing, and vice versa. The systems used 39-dimensional MFCCs and HMM-based ASR architecture as mentioned previously (tied-state triphone HMMs with 12 Gaussian mixtures per state). Figure 3 displays the training and testing protocols that were used in our experiments.

5.3. Train/test mismatch measure

The Kullback-Leibler divergence (KLD) between the probability density functions (PDFs), $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$, of cepstral vectors $\mathbf{x}$, extracted globally from training and testing speech, was used to measure the train/test mismatch (see section 1). It should be also noted that this is inherently a "coarse" measure since, in this case, the PDFs are estimated from MFCC vectors that were extracted globally from training and testing speech utterances. However, this measure could partially reveal the behavior of an HMM-based ASR system with a testing set whenever this set is modified [2]. To this end, the KLDs were calculated between every pair of PDFs of the MFCC vectors, extracted from the corresponding training and testing speech. This measure was used to evaluate the train/test mismatch reduction effect when combining cepstral normalization techniques with the cochlear implant-like speech signal processing.

6. EXPERIMENTAL RESULTS

6.1. Word error rate (WER)

In the experiments, either cepstral normalization, cochlear implant-like speech processing or their combinations were implemented in the front-end. The effectiveness of combining cepstral normalization and cochlear implant-like speech processing, in reducing the WER of microphone array-based ASR, was evaluated. WERs obtained from the speech recognition experiments are shown in figure 4. In all three recording scenarios, it can be observed from figure 4 that implementing only CMN or cochlear implant-like speech processing, in the front-end, helps in reducing the WER of microphone array-based ASR. The reduction gain obtained with CMN is better than that obtained with cochlear implant-like speech processing.

On the other hand, combining CMN with cochlear implant-like speech processing is better than using CMN alone, in terms of WER reduction. Combination of CMN and CVN also helps in reducing the WER, compared to using CMN alone, but the reduction gain is smaller than that obtained with the combination of CMN and cochlear implant-like speech processing, in all three scenarios. Furthermore, the combination of cepstral normalization (CMN and CVN) with cochlear implant-like speech processing makes it possible to reduce further the WER, in all three scenarios, compared to the combination of CMN and CVN.

6.2. Kullback-Leibler divergence (KLD)

Numerical values of the KLDs, calculated between the global PDFs of training and testing cepstral vectors (MFCCs), are shown in figure 5. For each KLD calculation, the PDFs $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$, which were 12-Gaussian mixtures, were estimated from the training and testing cepstral vectors of the corresponding speech recognition test.

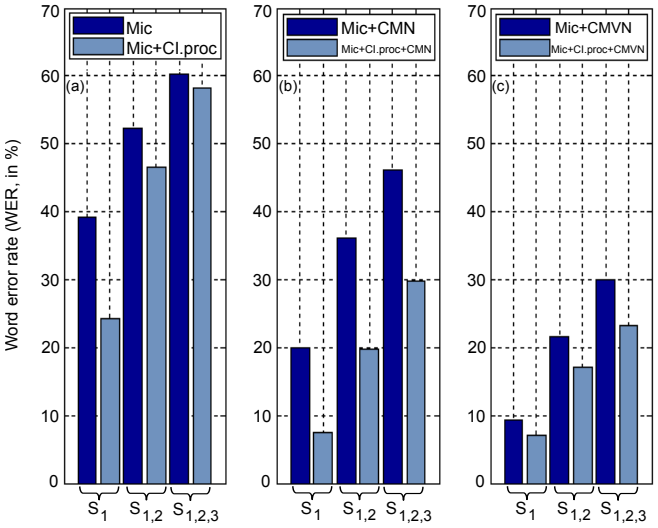

Fig. 4. Word error rates (WERs), in %, from microphone array-based ASR experiments. In testing, speech signals recorded by microphone array in three recording scenarios, S_1, $S_{1,2}$ and $S_{1,2,3}$, were evaluated. S_1: there is one desired speaker. $S_{1,2}$: there is one desired speaker and one competing speaker. $S_{1,2,3}$: there is one desired speaker and two competing speakers. "Mic" means microphone array and "CI.proc" means cochlear implant-like speech processing. CMN: cepstral mean normalization, CMVN: cepstral mean and variance normalization.

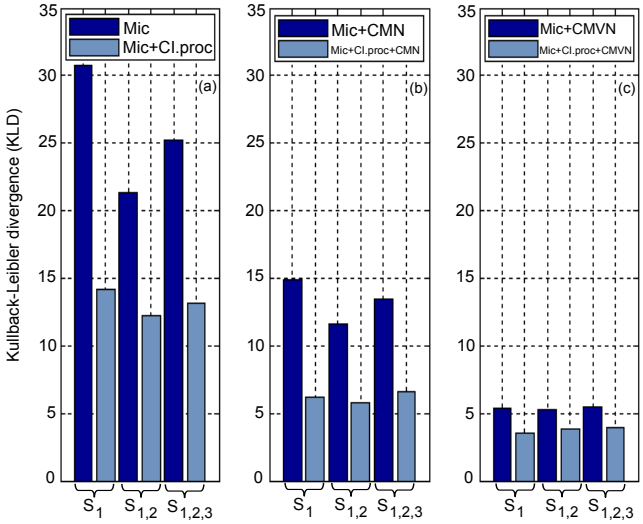

Fig. 5. Numerical values of the Kullback-Leibler divergences (KLDs), calculated between probability density functions (PDFs), $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$, of cepstral vectors (MFCCs) $\mathbf{x}$, extracted globally from training and testing speech utterances, respectively. The KLDs, between $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$ which are 12-Gaussian mixtures, were calculated by the Monte Carlo sampling method [21]. For each KLD calculation, the PDFs were estimated from the training and testing cepstral vectors of the corresponding speech recognition test (see figure 4).

The KLDs between the Gaussian mixtures were calculated by using the Monte Carlo sampling method [21]. In all three scenarios, it can be observed that implementing either cochlear implant-like speech processing, cepstral normalization (CMN and CVN) or their combinations helps in reducing the train/test mismatch, measured via the KLD. Furthermore, combining cepstral normalization (CMN or CMN+CVN) and cochlear implant-like speech processing reduces further the train/test mismatch, compared to using only cepstral normalization (see figures 5(b) and 5(c)). This train/test mismatch reduction is, therefore, consistent with the WER reduction, gained when combining cepstral normalization (CMN or CMN+CVN) and cochlear implant-like speech signal processing (see figure 4).

7. DISCUSSION

When combining CMN with cochlear implant-like speech processing, the WER relative reductions achieved in three testing scenarios, S_1, $S_{1,2}$ and $S_{1,2,3}$, are 62.43%, 45.13% and 35.45%, respectively, compared to using CMN only (see figure 4(b)). On the other hand, the combination of cepstral normalization (CMN+CVN) with cochlear implant-like speech processing makes it possible to reduce relatively 22.76%, 20.85% and 22.26% of the WER, compared to using CMN+CVN only, in three testing scenarios S_1, $S_{1,2}$ and $S_{1,2,3}$, respectively (see figure 4(c)). In three testing scenarios S_1, $S_{1,2}$ and $S_{1,2,3}$, the application of only cochlear implant-like speech processing makes it possible to reduce relatively 38.16%, 10.81% and 3.45% of the WER, respectively, compared to when there is only the microphone array signal processing (see figure 4(a)).

The WERs achieved with the combination of cepstral normalization (CMN+CVN) and cochlear implant-like speech processing, in three scenarios S_1, $S_{1,2}$ and $S_{1,2,3}$, are 7.16%, 17.16% and 23.29%, respectively. In the $S_{1,2}$ and $S_{1,2,3}$ scenarios where there is overlapping speech from 1 and 2 competing speakers, respectively, these WERs are lower than the latest WERs, 19.37% and 26.64%, reported on the same testing sets with the same training conditions, but using maximum a posteriori (MAP) adaptation to compensate noise [6]. When there is no overlapping speech in the scenario S_1, the 7.16% WER, achieved with the combination of CMN+CVN and cochlear implant-like speech processing, is comparable with that obtained with MAP adaptation, 7.00%, as reported in [6].

In fact, applying cochlear implant-like speech signal processing, which operates in the temporal domain, in training and testing, reduces the train/test mismatch, measured via the KLD between $f_{Tr}(\mathbf{x})$ and $f_T(\mathbf{x})$ [2]. On the other hand, cepstral normalization techniques, which operate in the cepstral domain, could also reduce the train/test mismatch, measured via the KLD. This reduction is straightforward since the variances of testing cepstral vectors are normalized and scaled to the variances of the training ones. As a consequence, combining such temporal and cepstral domain processing should have a complementary effect in terms of train/test mismatch reduction as well as WER reduction. The experimental results, displayed in figures 4 and 5, confirm these hypotheses.

Cepstral normalization might be combined also with other temporal domain speech processing that are different from cochlear implant-like speech processing. It might also be useful to compare the WER reduction gain, achieved by combining cepstral normalization and cochlear implant-like speech processing, with that achieved with the combination of cepstral normalization and other temporal domain processing, for instance the speech enhancement using linear prediction residual [22], etc. Moreover, combining cepstral normalization and cochlear implant-like speech processing might be beneficial in real situations where there is overlapping speech, e.g. in meeting speech recognition with a microphone array [23].

8. CONCLUSION

In this paper, we have investigated the combination of cepstral normalization and cochlear implant-like speech signal processing for

microphone array-based speech recognition. Speech signals have been recorded, in three scenarios, by a circular microphone array and have been processed with superdirective beamforming and McCowan post-filtering [12]. In testing, cochlear implant-like speech processing has been applied on the output of the post-filtering and, in training, this processing has been applied on clean training speech of the MONC database [7]. Cepstral normalization, including CMN and CVN, have been applied on the training and testing cepstral vectors (MFCCs). The CMN has been performed using local mean vector, calculated from cepstral vectors extracted from each speech utterance. In CVN, the variance of cepstral vectors has been normalized and scaled to the global variance of training cepstral vectors.

Speech recognition experiments have shown that implementing either cepstral normalization or cochlear implant-like speech processing helps in reducing the WERs of microphone array-based speech recognition. Furthermore, when there is overlapping speech, combining cepstral normalization with cochlear implant-like speech processing makes it possible to reduce further the WERs, lower than the previously published WERs, obtained with MAP adaptation, on the same training and testing data [6]. Train/test mismatch measures, via the KLDs between the global PDFs of training and testing cepstral vectors, have revealed that using either cepstral normalization or cochlear implant-like speech processing, on speech recognition with microphone array, helps in reducing the train/test mismatch. Numerical results have also revealed that combining these two processing, one in the cepstral domain and another in the temporal domain, reduces further the train/test mismatches, measure via the KLDs, as well as the WERs.

9. ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)". Cong-Thanh Do's postdoctoral research is currently supported by OSEO, the French State Agency for Innovation, under the Quaero program and by the ANR project QCOMPERE. The authors would like to thank Dr. Jean-Luc Gauvain (LIMSI-CNRS) for the guidance in the direction of the work.

10. REFERENCES

[1] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Communication*, vol. 25, pp. 29–47, 1998.

[2] C.-T. Do, D. Pastor, and A. Goalic, "A novel framework for noise robust ASR using cochlear implant-like spectrally reduced speech," *Speech Communication*, vol. 54, no. 1, pp. 119–133, Jan. 2012.

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[4] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[5] P. N. Garner, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Communication*, vol. 53, no. 8, pp. 991–1001, 2011.

[6] D. C. Moore and I. A. McCowan, "Microphone array speech recognition: experiments on overlapping speech in meetings," in *Proc. IEEE ICASSP, April 06 - 10, Hong Kong, China*, Apr. 2003, vol. 5, pp. 497–500.

[7] D. C. Moore and I. A. McCowan, "The multi-channel overlapping numbers corpus (MONC)," 2003, http://www.cslu.ogi.edu/corpora/monc.pdf.

[8] M. Seltzer, "Bridging the gap: towards a unified framework for hands-free speech recognition using microphones arrays," in *Proc. IEEE HSCMA Hands-free Speech Communication and Microphone Arrays Workshop*, May 2008, pp. 104–107.

[9] M. Brandstein and D. Ward, *Microphone array: signal processing techniques and applications*, Springer, Berlin, Germany, 2001.

[10] J. Benesty, M.M. Sondhi, and Y. Huang, *Springer handbook of speech processing*, Springer, Germany, 2008, Chapter 50: Microphone arrays (G.W. Elko and J. Meyer, pp. 1021-1041).

[11] H.W. Lollmann and P. Vary, "Post-filter design for superdirective beamformers with closely spaced microphones," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 291–294.

[12] I. A. McCowan and H. Bourlard, "Microphone-array post-filter based on noise field coherence," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[13] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *Proc. IEEE HSCMA Hands-free Speech Communication and Microphone Arrays Workshop*, May 2011, pp. 92–97.

[14] R. Zelinski, "A microphone array with adaptive post-filtering for noise reduction in reverberant rooms," in *Proc. IEEE ICASSP*, 1988, pp. 2578–2581.

[15] C.-T. Do, D. Pastor, and A. Goalic, "On the recognition of cochlear implant-like spectrally reduced speech with MFCC and HMM-based ASR," *IEEE Trans. Audio, Speech and Language. Process.*, vol. 18, no. 5, pp. 1065–1068, Jul. 2010.

[16] C.-T. Do, D. Pastor, G. Le Lan, and A. Goalic, "Recognizing cochlear implant-like spectrally reduced speech with HMM-based ASR: experiments with MFCCs and PLP coefficients," in *Proc. INTERSPEECH, Makuhari, Japan, September 26-30*, 2010, pp. 2634–2637.

[17] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[18] G. Kubin and W. B. Kleijn, "On speech coding in a perceptual domain," in *Proc. IEEE ICASSP, March 15 - 19, Phoenix, AZ, USA*, Mar. 1999, vol. 1, pp. 205–208.

[19] T. S. Gunawan and E. Ambikairajah, "Speech enhancement using temporal masking and fractional Bark gammatone filters," in *Proc. 10th Australian Intl. Conf. on Speech Sci. & Tech., Dec. 8 - 10, Sydney, Australia*, Dec. 2004, pp. 420–425.

[20] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (for HTK version 3.4)*, Cambridge University Engineering Department, Cambridge, UK.

[21] J. Hershey and P. Olsen, "Approximating the Kullback-Leibler divergence between Gaussian mixture models," in *Proc. IEEE ICASSP*, 2007, vol. 4, pp. 317–320.

[22] B. Yegnanarayana, C. Avendano, H. Hermansky, and P.S. Murthy, "Speech enhancement using linear prediction residual," *Speech Commun.*, vol. 1, no. 28, pp. 25–42, May. 1999.

[23] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings: the AMI and AMIDA projects," in *Proc. IEEE ASRU*, 2007, pp. 238–247.