A German distant speech recognizer based on 3D beamforming and harmonic missing data mask

Juan A. Morales-Cordovilla, Hannes Pessentheiner, Martin Hagmüller,
Pejman Mowlaee, Franz Pernkopf, Gernot Kubin

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
{moralescordovilla,hannes.pessentheiner,hagmueller,
pejman.mowlaee,pernkopf,gernot.kubin}@tugraz.at

Abstract

This paper addresses the problem of distant speech recognition in reverberant noise conditions applying a star-shaped microphone array and missing data techniques. The performance of the system is evaluated over a German database, which has been contaminated with noise of an apartment of the DIRHA (Distant Speech Interaction for Robust Home Applications) project. The proposed system is composed of three blocks. First, a beamformer yields an enhanced single-channel signal by filtering multi-channel signals and summing up all signals afterwards. To optimize the filter weights, we apply convex (CVX) optimization over three spatial dimensions given the spatio-temporal position of the target speaker as prior knowledge. Second, the beamformer output is exploited to extract pitch and estimate the stationary part of the background noise. Third, the system produces a final noise estimate by combining both, the stationary noise part as well as the harmonic noise estimate obtained from the pitch. Finally, the filter-bank representation of the enhanced signal and its corresponding missing data mask obtained from this final noise estimate are sent to the speech recognition back-end. The purpose of this paper is to analyze the impact of employing a beamformer followed by a missing data technique.

Index Terms: distant speech recognition, cvx-optimized beamforming, missing data imputation, star-shaped microphone array, reverberant and noisy environment, natural mixing, German database.

1. Introduction

The distant interaction of a speaker with a dialogue system, which controls some mechanisms of a house, is a difficult challenge because of many reasons: the wake-up of the system (distinction between simple conversations and commands), the change of the user accent in the automatic speech recognition (ASR), and the degradation of the speech signal due to background noise, reverberation, or the speaker position. Different projects such as CHIL, DICIT, and the recently finalized CHiME [1] have been proposed to solve this challenge, but the Distant-speech Interaction for Robust Home Applications (DIRHA) European project [2] (for people with disabilities) is different from the others in the use of the microphone array technology.

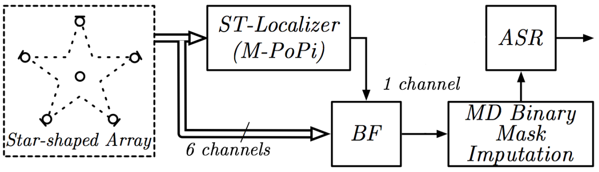

Figure 1: Block diagram of the proposed system for distant speech recognition, which consists of a 6-element star-shaped microphone array, a spatio-temporal localizer (ST-Localizer) of the speaker utterance, a beamformer (BF), a missing data (MD) binary mask imputation, and an automatic speech recognition (ASR) system.

To address the problems mentioned above, we propose the enhancement framework depicted in Fig. 1, which is an improved version of the one presented in [10]. It consists of a spatio-temporal localizer (ST-Localizer), which determines the user's position and speech activity. Later, a novel convex (CVX)-optimization–based beamformer (BF) attenuates the interference signals from directions different from the user's one. Finally, a missing data (MD) binary mask imputation method further increases the robustness of the ASR on the still degraded signal provided by the beamformer. In this paper, we avoid the problem of the spatio-temporal localization and focus on the beamformer and the compensation method justifying their proposed configuration with experimental results.

This paper also introduces a new and more realistic German speech database than presented in the previous work [10] to evaluate the proposed enhancement framework. In particular, we present a medium-vocabulary German database for a microphone array configuration, which contains embedded clean signals contaminated with real room impulse responses and mixed with real noises in a 'natural' way [1].

The paper is structured as follows: sections 2 and 3 describe the CVX beamforming and MD imputation methods, respectively. Section 4 explains the proposed BAS-embedded database and the ASR configuration. Section 5 presents and analyzes the experimental results, and in section 6 we summarize the most important ideas presented in the paper together with some future works.

2. Convex-optimization–based Beamformer

In our experiments, we employ a novel CVX-optimization–based beamformer. The beamformer design, first reported in

This work was partially funded by the European project DIRHA (FP7-ICT-2011-7-288121) as well as the K-Projects AAP (Advanced Audio Processing) and ASD (Acoustic Sensing and Design) in the COMET programme (Competence Centers for Excellent Technologies) with joint support from Speech Processing Solutions GmbH Vienna, BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), and the Government of Styria (A8: Wissenschaft und Gesundheit as well as A12: Wirtschaft, Tourismus, Sport). The programme COMET is conducted by the Austrian Research Promotion Agency (FFG).

[10], exhibits an improved extension of the design mentioned in [6]. The remarkable improvements of our modified beamformer are null-steering, the compatibility with different array geometries, and an optimization to three spatial dimensions. The last one is a prerequisite to enable beamforming in three spatial dimensions and to reduce the influence of reflections from the ceiling and the floor discussed in [9]. The CVX constrains the white noise gain to be larger than a lower limit γ. It considers the three-dimensional undistorted capturing response with steering direction (φ_s, θ_s) and nulls placed in different directions as constraints. The beamformer design is based on least squares computations that approximate a desired three-dimensional directivity pattern

$$\hat{b}(\omega, \varphi, \theta) = \sum_{n=1}^{N} w_n(f) e^{i \frac{\omega}{c} r_n \cdot \eta(\varphi, \theta, \varphi_n, \theta_n)}$$

with

$$\eta(\varphi, \theta, \varphi_n, \theta_n) = \sin(\theta)\sin(\theta_n)\cos(\varphi - \varphi_n) + \cos(\theta)\cos(\theta_n),$$

or, in vector notation,

$$\hat{\mathbf{B}}(\omega) = \mathbf{G}(\omega) \cdot [\mathbf{w}(\omega) \otimes \mathbf{I}],$$

where f and ω represent the linear and angular frequency, φ and θ are steering-direction–dependent azimuthal and elevation angles, φ_n and θ_n are the angles of a microphone with index n, N is the number of microphones, c is the sound velocity, r_n is the distance between a microphone and the center of the coordinate system, and $\mathbf{w}(\omega) = (w_1(\omega), w_2(\omega), ..., w_N(\omega))^T$ is the beamformer coefficient vector. Moreover, $\mathbf{I}$ is the identity matrix, $\otimes$ denotes the Kronecker product, and $\mathbf{G}(\omega)$ is an $(N_\theta \times [N \cdot N_\varphi])$ capturing response matrix according to $G_{l,m,n}(\omega) = e^{i \frac{\omega}{c} r_n \cdot \eta(\varphi_m, \theta_l, \varphi_n, \theta_n)}$, where N_φ is the number of discretized azimuthal angles φ_m, and N_θ is the number of discretized elevation angles θ_l. The beamformer assumes the same desired response for all frequencies, i.e. $\hat{\mathbf{B}}(\omega) = \hat{\mathbf{B}}$, and

$$\arg\min_{\mathbf{w}(\omega)} \| \mathbf{G}(\omega) \cdot [\mathbf{w}(\omega) \otimes \mathbf{I}] - \hat{\mathbf{B}} \|_F$$

subjected to the white noise gain (WNG), the undistorted capturing response with steering direction (φ_s, θ_s), and the optional null-placement constraints

$$\frac{|\mathbf{w}^T(\omega)\mathbf{d}(\omega)|^2}{\mathbf{w}^H(\omega)\mathbf{w}(\omega)} \geq \gamma, \;\; \mathbf{w}^H(\omega)\mathbf{d}(\omega) = 1, \;\; \mathbf{w}^H(\omega)\mathbf{V}(\omega) = \mathbf{0},$$

where $\mathbf{d}(\omega) = (d_1(\omega), d_2(\omega), ..., d_N(\omega))^T$ represents the capturing response with steering direction (φ_s, θ_s), and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_S]$ is a matrix which consists of vectors $\mathbf{v}(\omega) = (v_1(\omega), v_2(\omega), ..., v_{M-1}(\omega))^T$ that describe the capturing response of, e.g., competing speakers or other noise sources, S is the number of nulls, $(\cdot)^T$ is the transpose, $(\cdot)^H$ is the Hermitian-transpose, and $\| \cdot \|_F$ is the Frobenius norm. We set the lower limit γ and the desired response $\hat{\mathbf{B}}$ in a way that we are able to distribute the narrow null-lobe marked in Fig. 2 over frequencies below 1000 Hz. This yields a decreased main-lobe width at lower frequencies without increasing the width at higher ones. Although null-steering is one of the beamformer's big improvements, we did not consider it due to the assumption of unknown noise source positions in our experiments.

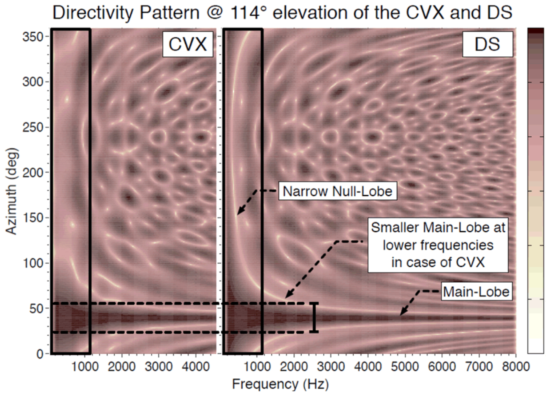

Figure 2: The directivity patterns of the CVX without null-steering and the DS (delay-and-sum, [12]) based on a 6-element star-shaped array with steering direction $\phi_s = 40°$ and $\theta_s = 114°$.

3. Missing Data Imputation

A missing data (MD) imputation mechanism based on binary mask [3] further de-noise the noisy single-channel signal yielded by the beamformer. After applying this technique, we obtain the log-mel spectrogram representation (Sec. 4.2) of the noisy signal and its corresponding MD binary mask. This mask is obtained by means of a threshold determined by comparing the noisy spectrogram with the noise spectrogram estimate. Then, using a Gaussian mixture model (GMM) trained with clean-reverberated speech, we replace the spectro-temporal elements of the noisy spectrogram dominated by the noise with a imputed estimation. The reason for using imputation rather than other methods, e.g., MD marginalization, is that it keeps the final representation of the clean estimated signal in the cepstral domain. This is a more appropriate representation for a medium or large vocabulary task.

The noise estimate used in this paper is a First-Last-Frames (FLF) noise revised by a harmonic tunnelling (Tun) noise in the pitch frames [8]. This estimation assumes that the first and last 20 frames of the cut signal correspond to noise, and these frames are used to estimate the log-Mel noise by means of a linear interpolation to the remaining frames. Taking into account that Tun noise is a celling estimate of the noise [8], we replace the FLF noise by the Tun where FLF is higher than Tun. We employ the pitch extractor described in [7] and a MD binary mask threshold of -3dB (computed empirically from a small development set).

4. Experimental Framework

4.1. Embedded-BAS Database

Due to a lack of suitable resources in German to evaluate the proposed enhancement framework, this paper also introduces a new German database for a star-shaped microphone array. More precisely, this array consists of 6 microphones (1 at the center and 5 on the circle) placed on the ceiling of the living room of the ITEA apartment used by Fondazione Bruno Kessler (FBK) for the DIRHA project [2] (see Fig. 3).

4.1.1. Embedded noisy signals

Each multi-channel test signal of this database represents what the microphone array would record: a speaker, in the presence

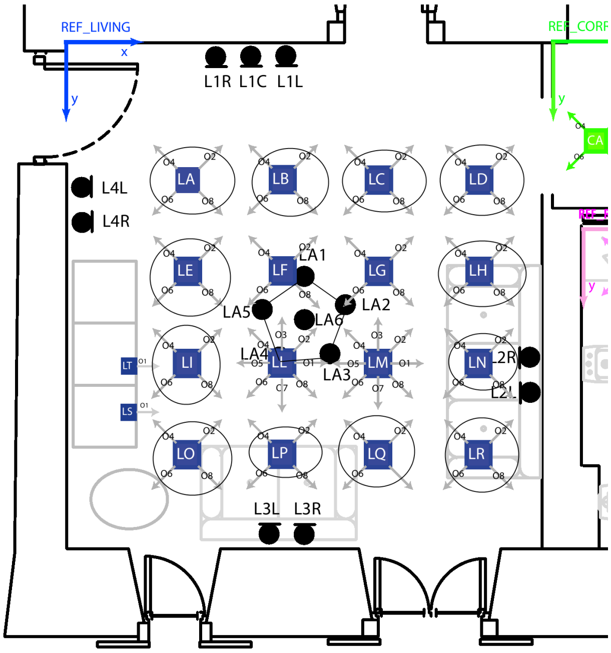

Figure 3: Living room of the ITEA apartment of Fondazione Bruno Kessler (FBK) with the microphone array at the center and the 12 speaker position/directions employed in this work [provided by FBK].

of noise, repeats the action of pronouncing an isolated utterance at a specific position in the room and later moved to another position to pronounce another utterance. We call this connection of utterances with continuous background noise and with different reverberations, which depend on the speaker position, *embedded noisy signal*.

For the controllability of the experiments, 12 speaker positions/directions—they are marked with a circle in Fig. 3—are used: LA/O8, LC/O6, LE/O8, LI/O2, LO/O2, LQ/O4, LB/O8, LD/O6, LH/O6, LN/O4, LP/O2 and LR/O4. To simulate the different SNR noisy conditions in the most possible 'natural' way, we follow the indications of SNR mixture of the CHiME corpus [1] by employing around 3-hours of real noise, recorded by the FBK group with this microphone array. The way to obtain an embedded noisy signal for a target SNR is summarized in the following steps:

1. We randomly select 7 isolated monaural clean (without reverberation) utterances of one speaker, convolve them with the corresponding impulse responses (obtained by the FBK group) of 7 random speaker positions/directions and obtain a 6-channel embedded clean-reverberant signal by connecting them with a time gap in the middle. These gaps are randomly selected between 0.5 and 5 seconds.

2. We randomly select a segment from all available segments of the 3-hours of noise, which yields the target SNR within an error of 1.5 dB. The following formula is used for the SNR:

$$SNR = 10log_{10}\frac{Ex_{central}}{En_{central}}(dB) \qquad (1)$$

   where $Ex_{central}$ and $En_{central}$ represent the whole energy of the central microphone of the embedded clean-reverberant signal and of the noise segment, respectively. If no noise segment is found that yields the target SNR, all channels of the embedded clean-reverberant signal are multiplied by a gain (which depends on the closest found SNR to the target SNR) to find at least an appropriate noise segment.

3. The final *embedded noisy signal* is the sum of this embedded clean-reverberant signal with the selected noise segment. In addition, sometimes this sum can produce a saturated signal in some of the channels. In order to avoid this problem we multiply all the channels of both, the embedded clean-reverberant signal and the noise, by a second factor, which avoids this problem.

4.1.2. Database description

The proposed *Embedded-BAS* database exhibits a sampling frequency of 16 kHz and employs the clean sentences of the Bavarian Archive for Speech Signals (BAS) PHONDAT-1 database [11]. The database consists of the training and test sets. The training set contains 4999 clean-reverberant isolated utterances corresponding to 50 different-gender speakers (around 100 sentences per speaker) with a reverberation that corresponds to position LA/O8 indicated in Fig. 3. The inclusion of the reverberation in the training set reduces the mismatch with the test set. The test set consists of 100 embedded clean-reverberant signals (700 isolated utterances, Sec. 4.1.1) corresponding to 100 different speakers (half of them are in the training set) contaminated at 10 and 0 dB. Both, the training and test sets, share the same medium-vocabulary lexicon and grammar and consist of 1504 words, which belong to around 500 different phrases.

4.2. ASR system

Both, the front-end and the back-end, have been derived from the standard recognizer employed in Aurora-4 database [5].

The front-end takes the enhanced signal and obtains mel frequency cepstrum coefficients (MFCCs) using 16 kHz sampling frequency, frame shift and length of 10 and 32 ms, 1024 frequency bins, 26 Mel channels and 13 cepstral coefficients. Then we apply cepstral mean normalization to the MFCCs. Delta and delta-delta features are also appended, obtaining a final feature vector with 39 components.

The back-end employs a transcription of the training corpus based on 34 monophones to train triphone-HMMs. This transcription has been derived from a more detailed transcription (based on 44 SAMPA-monophones) by means of a careful clustering of the less common monophones. Each triphone is modeled by a HMM of 6 states and 8 Gaussian-mixtures/state. By means of a monophone classification (created with the help of a linguistic) a tree-based clustering of the states is also applied to reduce the complexity and a lack of training data. Tree-based clustering also allows to create triphones models for the test stage that have not been observed in the training stage. We train a bigram using the training word transcription. By means of an expansion based on the grammar, the triphone transcription of the test lexicon and the triphones, we obtain the final macro HMMs for the test stage. It is important to point out that only the central microphone of the clean-reverberant training set without any enhancement (beamforming and imputation) is used to train our HMMs-models and the imputation GMM (Sec. 3).

Table 1: Word accuracies obtained by different configurations of the proposed systems tested over the presented Embedded-BAS database for different SNR values.

Systems	Clean	10 dB	0 dB	Average (clean pitch)
Baseline (central microphone)	93.24	79.34	43.69	72.09
DS Beamforming	94.73	83.61	51.73	76.69
CVX Beamforming	95.34	83.65	51.98	76.99
Baseline + Imputation (FLF+Tun noise)	91.75	80.46	40.84	71.02 (74.31)
DS Beamf. + Imputation (FLF+Tun noise)	93.79	84.19	48.21	75.40 (77.72)
CVX Beamf. + Imputation (FLF+Tun noise)	94.00	85.15	49.94	76.36 (78.69)
Baseline + Imputation (oracle mask)	93.24	93.06	77.92	88.07
DS Beamf. + Imputation (oracle mask)	94.73	94.67	81.58	90.33
CVX Beamf. + Imputation (oracle mask)	95.34	95.06	82.92	91.10

5. Experimental Results

Tab. 1 shows the different Word Accuracies (WAcc, %) achieved by different configurations of the proposed systems tested over the presented Embedded-BAS database for different SNR values.

The *Baseline (central microphone)* results are obtained by considering no enhancement and by using the center-microphone channel as our monaural signal channel. *DS Beamforming* and *CVX Beamforming* are the results achieved by delay-and-sum [12] and convex-optimization beamformers (Sec. 2). As mentioned in Sec. 1 , we assume that the ST-Localizer of Fig. 1 provides the oracle spatial and temporal localization of the speaker, i. e., we cut the embedded noisy signal in pieces which correspond to the isolated utterances, then each of these pieces together with its spatial position are sent directly to the beamformer. In the remaining part of this section, we can see the results of the three previous configurations considering imputation (Sec. 3) with FLF+Tun noise (Sec. 3).

The most significant conclusions which can be drawn from the table are the follows:

1. Using beamformers, especially the CVX, always improves the recognition results (compare the 72.09 of the *Baseline* with the 76.99 % of the *CVX Beamforming*).
2. Considering imputation after applying beamformers does not improve the results (compare the 76.99 of the *CVX Beamforming* with the 76.36 % of the *CVX Beamf. + Imputation (FLF+Tun noise)*).

The imputation technique is sensitive to MD mask errors, which is the reason for the lack of improvement. These errors are due to a bad estimation of the pitch, which produces an erroneous tunnelling noise and, as a result, mask errors. The parenthesized results with clean pitch show that, with a better pitch estimation, the addition of the imputation to the beamforming can be useful. Compare the 74.31 of *Baseline + Imputation* with the 78.69 of *CVX Beamf. + Imputation*. The results with oracle MD mask are only displayed to show the upper performance of this framework. We can see that we should further improve the noise estimation at 0 dB.

6. Conclusion and Future Work

This paper presented a system for distant speech recognition in reverberant and noisy conditions, intended to control a room with commands. The proposed system is an improved version of the system presented in [10]. The improvement consists of a recently presented beamformer based on convex optimization, the application of a single-channel enhancement algorithm based on MD imputation with FLF+Tun noise estimate, and the presentation of a more realistic database for evaluations. The database consists of embedded noisy signals, which represent, with 'natural' noise mixing, what the microphone array would record if the speaker emits German commands at different positions of the room. This database is a very suitable challenge for the spatio-temporal localization algorithm of the utterance which is our next future objective. To do it we plan to make use of the pitch information provided by the M-PoPi algorithm [4].

7. References

[1] H. Christensen, J. Barker, N. Ma, and P. Green. The chime corpus: A resource and a challenge for computational hearing in multisource environments. In *Interspeech*, 2010.

[2] European project FP7. Distant-speech interaction for robust home applications (dirha). In *http://dirha.fbk.eu*, 2012-2015.

[3] J. A. Gonzlez, A. M. Peinado, N. Ma, A. M. Gomez, and J. Barker. Mmse based missing-feature reconstruction with temporal modeling for robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2013.

[4] T. Habib and H. Romsdorfer. Concurrent speaker localization using multi-band position-pitch (m-popi) algorithm with spectro-temporal pre-processing. In *Interspeech*, 2010.

[5] H. G. Hirsch. Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task. Technical report, STQ AURORA DSR, Working Group, 2002.

[6] E. Mabande, A. Schad, and W. Kellermann. Design of robust superdirective beamformers as a convex optimization problem. In *ICASSP*, 2009.

[7] Juan A. Morales-Cordovilla, Pablo Cabaas-Molero, Victoria Snchez, and Antonio M. Peinado. A robust pitch extractor based on dtw lines and casa with application in noisy speech recognition. In *Iberspeech*, 2012.

[8] Juan A. Morales-Cordovilla, Ning Ma, Victoria Sánchez, José L. Carmona, Antonio M. Peinado, and Jon Barker. A pitch based noise estimation technique for robust speech recognition with missing data. In *ICASSP*, 2011.

[9] H. Pessentheiner, G. Kubin, and H. Romsdorfer. Improving beamforming for distant speech recognition in reverberant environments using a genetic algorithm for planar array synthesis. In *10th ITG Symposium on Speech Communication*, 2012.

[10] H. Pessentheiner, S. Petrik, and H. Romsdorfer. Beamforming using uniform circular arrays for distant speech recognition in reverberant environments and double-talk scenarios. In *Interspeech*, 2012.

[11] F. Schiel and A. Baumann. Phondat 1, corpus version 3.4. Technical report, Bavarian Archive for Speech Signals (BAS). http://www.bas.uni-muenchen.de/Bas/BasFormatseng.html, 2006.

[12] I. Tashev. *Sound Capture and Processing: Practical Approaches*. John Wiley and Sons, 2009.