THE 2ND 'CHIME' SPEECH SEPARATION AND RECOGNITION CHALLENGE: APPROACHES ON SINGLE-CHANNEL SOURCE SEPARATION AND MODEL-DRIVEN SPEECH ENHANCEMENT

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

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

ABSTRACT

In this paper, we address the small vocabulary track (track 1) described in the CHiME 2 challenge dedicated to recognize utterances of a target speaker with small head movements. The utterances are recorded in a reverberant room acoustics corrupted with highly non-stationary noise sources. Such adverse noise scenario imposes a challenge to state-of-the-art automatic speech recognition systems. We developed two individual front ends for the output of the delay-and-sum beamformer: (i) a model-driven single-channel speech enhancement stage which combines the knowledge of the speaker identity modeled by a trained vector quantizer with a minimum statistics based noise tracker, and (ii) a single-channel source separation stage which employs models of the target speaker as well as the background noise as codebooks. Our perceived signal quality and separation results averaged on the CHiME 2 development set justify the effectiveness of both strategies in terms of recovering the target speech signal. Also, our best results on keyword recognition accuracy show 20% improvement over the provided baseline results on the development and test sets.

Index Terms— Single-channel source separation, Model-driven speech enhancement, Automatic speech recognition.

1. INTRODUCTION

The 2nd 'CHiME' Speech Separation and Recognition Challenge is the third challenge on automatic speech recognition (ASR) robustness contests. The first round of the speech separation and recognition challenge presented in [1] focused on separating co-channel speech mixture composed of two speakers without background noise or reverberation. The second challenge in [2, 3], called 1st CHiME challenge addressed the problem of recognizing speech of a target speaker in reverberation corrupted with realistic noise sources recorded in a domestic room. The 2nd CHiME challenge tries to be more realistic and proposes two different tracks: the small vocabulary and the medium vocabulary. In this paper, we focus on the small vocabulary track (track 1), which is very similar to the first CHiME challenge, but now small movements of the head are allowed.

To improve the recognition accuracy of the target speech utterance, it is important to remove signal impairments in spatial domain as well as in spectral-temporal domain. For spatial filtering, in this paper, we apply a simple delay-and-sum (DS) beamformer. On the other hand, to address spectral-temporal filtering, we employ a single-channel enhancement and a single-channel separation algorithm to remove the undesired signal components from the noise corrupted speech signal.

The single-channel signal enhancement or separation solutions are limited in their performance mainly due to the following reasons: To estimate the additive noise, state-of-the-art speech enhancement methods (e.g. [4–7]) rely on the assumption that the corrupting noise is slowly varying in its second order statistics compared to speech. For the scenario in this challenge, however, the speech signal undergoes several impairments including room reverberation and corruption by highly non-stationary and unpredictable background noise or even other competing speakers. Therefore, noise estimation alone is not able to track such fast changes accurately. Furthermore reverberation will introduce some biases, leading to over- or under-estimation of noise. Some preliminary enhancement results justifying the limited performance in such adverse noise scenarios are provided in [8].

On the other hand, single-channel source separation (SCSS) techniques have been employed to improve the speech recognition accuracy of co-channel speech mixtures [1, 9–13]. Model-driven SCSS techniques achieve large improvements in terms of ASR [1, 9] as well as perceived signal quality [10, 11]. However, their successful performance is usually reported for co-channel speech mixtures without background noise or reverberation. Furthermore, sufficient source-specific data is required for model training as these methods rely on a pre-trained model of the interfering sources, which is applied during separation; In the CHiME 2 challenge, however, both background noise and reverberation exist and a variety of noise types may occur.

In this paper, we study two systems to address the small vocabulary track in the CHiME 2 challenge: (1) a model-driven single-channel speech enhancement (MD-SCSE) system, which combines a vector quantization (VQ)-based target speaker model together with a minimum statistic noise tracker, and (2) a SCSS algorithm presented in [14], which employs a speaker-dependent model of the target speaker and a general model for noise. Both models are trained in the log-mel spectral domain, in contrast

This work was partially funded by the European project DIRHA (FP7-ICT-2011-7-288121) and by the K-Project AAP in the COMET (Competence Centers for Excellent Technologies) programme with joint support from speech processing solutions Vienna, BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), and the Government of Styria ("Abt. 3: Wissenschaft und Forschung" as well as "Abt. 14: Wirtschaft und Innovation"). The programme COMET is managed by the Austrian Research Promotion Agency (FFG).

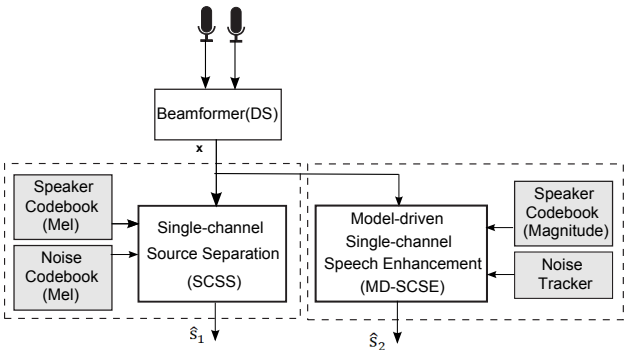

Fig. 1. Block diagram of the proposed front-ends used for robust speech recognition. An initial step of beamforming is combined with single-channel source separation and single-channel speech enhancement

to commonly used short-time Fourier transform (STFT) features [1, 10]. The two approaches are individually used as front-ends to provide enhanced signals. Mel-frequency cepstral features (MFCCs) are then extracted and ASR is performed.

The paper is structured as follows: In the following section we present the signal processing front-end composed of a delay-and-sum (DS) beamformer, single-channel source separation, and a model-driven single-channel speech enhancement. Section 3 provides details about the experimental setup, the database, feature extraction and about the automatic speech recognition engine. Section 4 reports separation and enhancement results obtained by the proposed front-ends averaged on the development set. Furthermore, we report ASR results on the development and test sets. Section 5 concludes on the paper.

2. FRONT-END SIGNAL PROCESSING

The strategies used as front-ends to enhance the binaural noisy input speech signal are displayed in Figure 1. The proposed system block diagram is composed of three steps: 1) DS beamformer, 2) single-channel speech separation, and 3) model-driven single-channel speech enhancement.

2.1. Delay-and-Sum Beamformer

For spatial filtering, the delay-and-sum beamformer is used, which provides the sum of the left and right microphone signals. Similar to previous CHiME challenge, the time-delay was set to zero to maintain the simplicity. The beamformer output is given by:

$$\mathbf{x} = \frac{\mathbf{x}_L + \mathbf{x}_R}{2}, \tag{1}$$

where $\mathbf{x}_L = \{x_L^t\}_{t=1}^T$ and $\mathbf{x}_R = \{x_R^t\}_{t=1}^T$ are the left and right time-domain microphone signals is the time frame index and T as the total number of samples. The DS beamformer output is passed either to MD-SCSE or SCSS, described in the following.

2.2. Single-channel Source Separation

For the rest of the paper, we neglect the time index t in our notation, as we process the signal at a frame-level basis. Lowercase letters are used for time-domain samples. Capital boldface letters denote the magnitude spectrum of the STFT vectors at each time index. Symbols $\tilde{\cdot}$ and $\hat{\cdot}$ are used to refer to actual mel-spectra and estimated magnitude, respectively. The unknown clean target speech signal and interfering noise signal are represented by **s** and **n**, respectively.

We use model-based SCSS for separating the observed noisy speech to find estimates for the unknown speech and noise components. In particular, we use the factorial VQ model proposed by Roweis [14]. This model is based on the MIXMAX model [15], i.e., the log-magnitude DFTs of two sources can be approximated by the element-wise maximum of their respective single-source log-magnitude DFT, i.e., $\log(\mathbf{X}) \approx \max(\log(\mathbf{S}), \log(\mathbf{N}))$, where $\mathbf{X} = \{X\}_{d=1}^D$ denotes the short-time magnitude spectrum of the signal mixture at each time frame t, d is the frequency index, and D denotes the total number of frequency bins. This approximation is based on the sparse nature of speech in time-frequency representations where each bin of a mixture spectrogram is dominated by a single source.

The factorial max-VQ model requires a codebook for each source. The sample vectors for learning the codebooks of each speaker and the noise are extracted from the STFT of both the clean reverb speaker signals and the noise signals, i.e., we perform SCSS in a speaker dependent manner. The 1024-point STFT is computed for time frames of 32 ms using a Hamming window. The frame-overlap is 10ms and zero padding is applied. We use the max-VQ model in the log mel-frequency domain, i.e., the magnitude spectra calculated at different frames are transformed into the log mel-domain. The log mel-domain speech and noise frames are denoted by $\tilde{\mathbf{S}}$ and $\tilde{\mathbf{N}}$, respectively. Separating speech signals in the log mel-domain rather than in the log STFT domain, reduces the size of the binary mask while the influence on the separation performance is limited. In [10], it was shown that a transformation of higher resolution at low frequency on the selected features achieves a better separation performance. The codebooks for speaker i, $\mathscr{S}_i$, and the noise $\mathscr{N}$ are obtained by the K-means algorithm [16] using the log-mel spectrogram data (i.e. reverberated single speaker data and noise data). Each codebook $\mathscr{S}_i$ consists of K codevectors, i.e., $\mathscr{S}_i = \left\{\tilde{\mathbf{S}}_k^i\right\}_{k=1}^{K=500}$, where each codeword entry is composed of 26 mel-filter-bank coefficients [17].

For separation of the two sources the best combination of codebook vectors is determined leading to the minimal ℓ^2-norm for each frame and speaker i, i.e.

$$\{k_1^*, k_2^*\} = \arg\min_{k_1,k_2} \left\| \tilde{\mathbf{X}} - \max\left(\tilde{\mathbf{S}}_{k_1}^i, \tilde{\mathbf{N}}_{k_2}\right) \right\|_2^2, \tag{2}$$

where $\tilde{\mathbf{X}}$ is the actual mel-spectra of the output of the beamformer **x**. The search space for separation of two sources is $\mathscr{O}\left(K^2\right)$. In [13], we introduced the iterated conditional modes (ICM) algorithm to dramatically reduce the computational costs for codebook vector selection by almost two orders of magnitude. Once we have found the optimal indices $\{k_1^*, k_2^*\}$ for all t, we use the corresponding codevectors as approximation of the log-mel speaker and noise spectrum denoted by $\tilde{\mathbf{S}}_{k_1^*}$ and $\tilde{\mathbf{N}}_{k_2^*}$, respectively. These approximations enable to compute a continuous mask in mel-domain. However, we convert $\tilde{\mathbf{S}}_{k_1^*}$ and $\tilde{\mathbf{N}}_{k_2^*}$ to magnitude spectrum representations $\hat{\mathbf{S}}_{k_1^*}$ and $\hat{\mathbf{N}}_{k_2^*}$, using a conversion from mel-domain to frequency domain [18] and derive a softmask in frequency domain (more details on softmask estimation is given in Section 2.4). This mask is applied to the noisy signal for recovery of the target speech signal.

2.3. Model-driven Single-Channel Speech enhancement

Here, we combine the noise estimate obtained by the minimum statistics (MS) noise tracker [19], with the minimum mean square error (MMSE) speech estimate derived from a pre-trained codebook for the target speaker. The idea to use these components together has already been presented in [8] under the name of model-driven speech enhancement and was shown to achieve effective noise reduction performance on the SiSEC 2011 challenge (as a subset of CHiME 1 challenge) [20]. More recently, the extension of the idea for binaural scenario was shown to achieve reasonable performance on the full development set of CHiME 1 compared with several other participants [21].

We train codebooks on the clean reverberated speech dataset of each speaker, in order to capture the spectral characteristics of the target speaker in the noisy signal. For each speaker, 600 utterances of the training set, were used to train the speaker models in the amplitude spectrum domain. In the following, the speaker model is denoted by $\mathscr{C}_S = \{\mathbf{S}_r\}_{r=1}^{R}$, where r is the codebook index. A VQ with $R = 2048$ centroids is used to model the target speaker.

For noise estimation The minimum statistics (MS) noise tracker [19] is used to obtain estimates for the power spectral density of the stationary part of the background noise denoted by $E\{\hat{N}^2_{MS,d}\}$. This noise estimate is then used to perform the voice activity detection (VAD) based on the a posteriori SNR estimate $\gamma_d = \frac{X_d^2}{E\{\hat{N}^2_{MS,d}\}}$, with a decision threshold of 0 decibels. For the speech absence region, we select the attenuation level of $G_{d,min} = -15$ decibels as often used in speech enhancement [22]. The minimum statistics approach has the advantage of getting updated even in speech presence regions [19]. The first six frames in each sentence are assumed to be noise only, to initialize the noise estimate. This choice achieved the best results over the CHiME scenario.

For MMSE speech estimation, the speech frames detected by the VAD are passed for further analysis, by the speaker codebook. To this end, we select a subset of M most likely codevectors after sorting their ℓ^2-norm distortion measure in the magnitude spectrum domain defined as

$$r^* = \arg\min_r \underbrace{E\{\|\mathbf{W}(\mathbf{X} - \hat{\mathbf{S}}_r)\|_2^2\}}_{\mathbf{d}}, \tag{3}$$

where $\hat{\mathbf{S}}_r$ refers to the r-th codevector in the speaker codebook. We define $\mathbf{d} = \{e_r\}_{r=1}^{R}$ where $e_r = \sum_{d=1}^{D} W_d (X_d - S_{r,d})^2$ as the distance metric used to find the optimal codevector r^*, and $W_d = \sqrt{\gamma_d}$ is the weighting function based on the a posteriori SNR. For each r-th codevector, we also define $\alpha_r = P(\mathbf{S}_r|\mathbf{X})$ as the probability of selecting the r-th codevector, which is calculated as $\alpha_r = \frac{e_r}{\sum_{r=1}^{R} e_r}$, where $\sum_r \alpha_r = 1$.

These M codevectors are considered as potential candidates to reconstruct the unknown speech spectrum. The role of γ_k is to emphasize on the dominance of the target speech spectral components according to the estimated a posteriori SNR. This weighting essentially converts the traditionally used least squares (LS) distance metric to a weighted least squares (WLS) metric, in the context of model-based speech enhancement. Previous study in [23] shows that a WLS distance metric is more robust to noise, as it finds better codewords entries compared to an unweighted metric here, a correct codevector refers to the one calculated for the clean signal scenario which provides the upper-bound of a model-driven speech enhancement method [23]. The MMSE speech estimate is approximated by the truncated weighted sum of the most likely ($M < R$) speech spectra,

$$\hat{\mathbf{S}}_{MMSE} = \sum_{m=1}^{M} \alpha_{r_m} \tilde{\mathbf{S}}_{r_m}, \tag{4}$$

where r_m with $m \in [1, M]$ denotes the index of the M most likely speech spectra in the speech codebook where $\alpha_{r_m} = \frac{e_{r_m}}{\sum_{r=1}^{M} e_{r_m}}$. Here, we found $M = 5$ suffices.

2.4. Softmask Signal Reconstruction

The speech and noise estimates obtained from either the single-channel source separation or the model-driven single-channel speech enhancement algorithms are used to form a square root Wiener filter as an estimate for the a priori SNR

$$\hat{G}_d^W = \frac{\hat{S}_d}{\sqrt{\hat{S}_d^2 + \tau \hat{N}_d^2}}, \tag{5}$$

where τ is the over subtraction factor. For SCSS, we select $\hat{\mathbf{S}} = \hat{\mathbf{S}}^i_{k_1^*}$ and $\hat{\mathbf{N}} = \hat{\mathbf{N}}_{k_2^*}$, while for the enhancement scenario we select $\hat{\mathbf{S}} = \hat{\mathbf{S}}_{MMSE}$ and $\hat{\mathbf{N}} = \hat{\mathbf{N}}_{MS}$. Previous studies show improved speech enhancement and robust speech recognition by choosing $1.3 < \tau < 2$ for low SNR scenarios [6]. $\tau = 1.7$ for the enhancement algorithm leads to the best result while $\tau = 1$ is used for SCSS. To re-synthesize time signals, the softmask mask is multiplied with the original noisy spectrogram and the inverse STFT followed by an overlap-and-add procedure is applied. The phase of the noisy signal is used for reconstruction.

3. EXPERIMENTAL SETUP

The recognition system have been evaluated using the 2012 2nd CHiME challenge track 1 (small vocabulary) database [3]. The challenge consists of recognizing the keywords *digits* and *letters*, from the GRID sentences uttered by a target speaker in a reverberant noisy environment. In the test stage, only the *Isolated-Test* (and not the *Embedded*) utterances of the database has been considered. In the training stage both the *Reverberated* and the *Noisy* (and not the *Clean*) set with 17000 utterances from 34 different speakers (18 males and 16 females) have been employed to train our models. Also we have employed the *Isolated-Development* and *Noise* data provided by the organizers to tune some parameters of our system.

Both the front end (FE) and the back-end (BE) have been derived from the recognition system provided by the organizers. The FE takes the single-channel enhanced signal of the two proposed methods presented in Section 2, and obtains mel frequency cepstrum coefficients (MFCCs) using 16 kHz sampling frequency, frame shift and length of 10 and 32 ms, D=1024 frequency bins, 26 Mel channels and 13 cepstral coefficients. The same parameter setup has been used in the enhancement methods described in Section 2. Delta and delta-delta features with a window length of 5 (half length 2) are also appended, obtaining a final feature vector with 39 components. The same parametrization is used by the organizers. We applied cepstral mean normalization (CMN) to obtain MFCC feature vectors. The BE model employs word level left-to-right HMMs with the following parametrization: 7 component-gaussian-mixtures/state with diagonal covariance matrices and the same language model as

Input SNR (dB)	-6	-3	0	3	6	9
PESQ						
Noisy (Baseline)	1.58	1.76	2.03	2.26	2.50	2.74
MMSE-LSA [4]	1.04	1.27	1.55	1.83	2.08	2.34
Cepstral smoothing [5]	1.25	1.27	1.85	2.18	2.46	2.71
MD-SCSE	1.69	1.82	**2.10**	**2.46**	**2.53**	**2.75**
SCSS	**1.79**	**1.90**	2.06	2.25	2.42	2.58
SDR (dB)						
Noisy (Baseline)	-6.28	-3.93	-1.56	1.04	3.65	5.97
MD-SCSE	-1.71	-0.36	1.62	3.05	4.34	5.36
SCSS	**-0.48**	**0.64**	**2.46**	**3.94**	**5.02**	**5.79**
SIR (dB)						
Noisy (Baseline)	-6.28	-3.93	-1.56	1.04	3.65	5.91
MD-SCSE	2.39	3.78	6.57	8.86	11.17	13.98
SCSS	**3.43**	**4.46**	**6.99**	**9.45**	**11.65**	**14.12**

Table 1. PESQ and BSS EVAL results results obtained by the proposed FEs. The results are averaged on the development set of CHiME 2 and are grouped according to six SNR levels. Bold face numbers highlight the best performance achieved for each SNR condition.

the organizers. The number of states per word is selected as described by the organizers. Each speaker-dependent (SD) model is created as the organization explains, by retraining an initial speaker-independent (SI) model by using only the 500 utterances corresponding to each speaker.

It is important to point out two modifications made to the baseline system of the organization which have helped to improve the performance of the results. First, we introduced a floor value on the log-mel representation in the FE. Second, we use a floor value for the state-variance of the Gaussian mixtures in the BE. Retraining our models is only possible when this variance floor is established. By means of a small subset of the development set, we have set these two floors to -15.00 log-mel-units and to 0.01 times the global mean variance of the training set, respectively.

4. RESULTS

4.1. Quality Metrics

We first report the speech quality results obtained by our FEs. The evaluation criteria used are perceptual evaluation of speech quality (PESQ) [24] and the metrics in the blind source separation evaluation (BSS EVAL) toolkit [25], often used to measure the perceived signal quality and blind source separation performance. The PESQ and the BSS EVAL results are reported in Table 1. The results in terms of signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR) are reported in decibels. All results are averaged over the development set grouped according to the signal-to-noise ratios. For a fair comparison, the beamformer output is used as the input signal to the speech enhancement methods studied here. From these results, the following observations are made: Higher SIR and SDR results are obtained by SCSS compared to MD-SCSE, while improved PESQ results are observed for the MD-SCSE method, especially for high SNRs. Some audio wave files are available[1]. Both strategies, achieve a significant improvement in terms of noise reduction, compared to the selected state-of-the-art speech enhancement methods [4, 5]. Further comparison of the proposed MD-SCSE with the standard noise reduction technique as vector Taylor series (VTS) based approach [26] showed an average improvement of 1.24 decibels in SDR.

[1] `http://www2.spsc.tugraz.at/people/pmowlaee`

4.2. ASR Results

Tables 2 and 3 show the keyword recognition accuracy (WAcc) in percent for the proposed systems for the development and test sets of track 1 (small vocabulary) in the CHiME 2 challenge. A full description of the challenge is provided in [27]. The results are reported for clean reverberated and noisy training and for speaker-independent (SI) and speaker-dependent (SD) recognition. In Table 3, comparing the *Organizers Baseline* (56.89 % on average) with the proposed *TUGraz Baseline* (63.26) result, we can see an improvement of 6.39 %. This improvement is mainly due to the log-mel representation floor (applied in the FE) and the state-variances floor (applied in the BE).

The next improvements are obtained by applying the front ends: MD-SCSE and SCSS. For SI and SD recognizers, with clean reverberated training, MD-SCSE improves the performance 64.15% and 65.51%. Other improvement is obtained by the *SCSS (SI) system and later by the (SD)* system which reaches up to 69.84% and then to 70.31% with clean reverberated training and with SI and SD models, respectively. This improvement is due to the capacity of the *SCSS* system to enhance the reverberated-noisy signal at different SNRs. Especially, at low SNRs, the single-channel source separation strategy shows large improvements with respect to the baseline results. This can be explained by the previous findings on the high capability of model-driven SCSS algorithms in separating co-channel speech mixtures [1]. We conjecture that one reason is that the noise at low SNRs mainly consists of another competing speaker or has in general some harmonic structure, and therefore a model-driven SCSS can improve ASR performance by canceling the spectral components of the interfering signal. The previous *SCSS* result can be even more improved, if noisy training and speaker dependent (SD) models are used. In this case the system reaches a performance of 74.00% for MD-SCSE and 77.66% for SCSS. This last result achieved by SCSS is our proposed best results for the CHiME 2 Challenge.

5. CONCLUSION

In this paper, we addressed the issue of robust speech recognition in multisource reverberant environments as described in the CHiME 2 recording setup. Here, we studied the performance of two front ends applied on the output signal of a delay-and-sum beamformer: a single-channel source separation and a model-driven single-channel speech enhancement method. The proposed single-channel front-end processing methods demonstrate promising improvements across different input signal-to-noise ratios, both in signal quality related measures and in ASR performance. On average, our best proposed system led to 8.5 dB improvement in signal-to-interference ratio with 3 dB improvement in signal-to-distortion ratios, compared to the noisy input signal. Our best recognition performance improved the baseline results from 56.89 % to 77.66 % word recognition accuracy averaged over all SNRs.

The two ideas presented in this work were earlier developed in [8, 12, 21]. However, in this study we we particularly presents the performance of such systems for a multisource reverberant scenario. In our proposed approaches, we have neither employed more sophisticated features such as relative spectral transform-perceptual linear prediction (RASTA-PLP) and cepstrum third-order normalization (CTN) nor any important tuning of our parameters, using the development set. Further improvements of current ASR results are expected by combining the proposed ap-

Final results (Development set)	-6dB	-3dB	0dB	3dB	6dB	9dB	Average
Clean reverberated training							
Baseline (Organizers)	32.08	36.33	50.33	64.00	75.08	83.50	56.88
TUGraz Baseline (SI)	40.92	46.17	57.75	66.58	75.58	81.92	61.48
MD-SCSE (SI)	45.33	47.08	59.17	70.17	78.00	83.17	63.82
SCSS (SI)	53.08	54.92	65.50	74.17	79.83	85.33	68.80
TUGraz Baseline (SD)	38.92	47.25	57.33	69.83	81.08	87.33	63.62
MD-SCSE (SD)	42.00	50.42	61.08	73.83	81.17	88.67	66.20
SCSS (SD)	48.25	52.83	66.67	76.33	83.67	88.67	69.40
Noisy training							
Baseline (Organizers)	49.67	57.92	67.83	73.67	80.75	82.67	68.75
TUGraz (SI)	54.17	59.42	67.08	76.25	79.50	82.58	69.83
MD-SCSE (SI)	56.33	61.50	68.92	76.50	80.58	83.58	71.23
SCSS (SI)	59.17	65.00	70.75	75.58	80.17	82.75	72.23
TUGraz Baseline (SD)	53.92	61.83	71.92	78.25	83.17	85.58	72.44
MD-SCSE (SD)	54.92	63.08	71.92	77.42	84.08	85.33	72.79
SCSS (SD)	**62.00**	**66.83**	**75.25**	**80.83**	**85.00**	**86.25**	**76.02**

Table 2. Keyword recognition accuracy obtained by the best methods for speaker-dependent (SD) and speaker-independent (SI) recognition and for clean reverberated and for noisy training, reported on the development set. Bold face highlights the best performance achieved for each SNR condition.

Final results (Test set)	-6dB	-3dB	0dB	3dB	6dB	9dB	Average
Clean reverberated training							
Baseline (Organizers)	32.17	36.33	50.33	64.00	75.08	83.50	56.89
TUGraz Baseline (SI)	39.33	44.58	54.42	66.75	75.83	82.92	60.63
MD-SCSE (SI)	41.92	50.62	58.42	69.67	77.05	84.25	64.15
SCSS (SI)	53.08	56.83	66.42	76.25	80.83	85.67	69.84
TUGraz Baseline (SD)	37.92	44.17	59.50	70.42	80.42	87.17	63.26
MD-SCSE (SD)	38.58	48.17	61.50	73.08	82.75	89.00	65.51
SCSS (SD)	48.25	54.75	67.25	77.83	84.33	89.50	70.31
Noisy training							
TUGraz Baseline (SI)	54.33	60.75	68.42	75.77	79.83	83.42	70.42
MD-SCSE (SI)	55.67	60.42	68.92	75.42	80.42	82.58	70.57
SCSS (SI)	58.83	64.00	71.75	77.75	81.67	84.58	73.09
TUGraz Baseline (SD)	55.0	64.00	71.5	77.58	81.08	85.08	72.37
MD-SCSE (SD)	55.83	66.17	73.50	79.17	83.75	85.42	74.00
SCSS (SD)	**63.67**	**68.92**	**76.00**	**83.92**	**85.25**	**88.25**	**77.66**

Table 3. Keyword recognition accuracy obtained by the best methods for speaker-dependent (SD) and speaker-independent (SI) recognition for clean reverberated and for noisy training reported on the test set. Bold face highlights the best performance achieved for each SNR condition.

proaches with ROVER fusion [28], integrating more robust features, improving the noise estimation by means of the pitch [29, 30], and employing a better tuning of some of the parameters (e.g., τ in softmask estimation for signal reconstruction).

6. REFERENCES

[1] M. Cooke, J. R. Hershey, and S. J. Rennie, "Monaural speech separation and recognition challenge," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 1–15, 2010.

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

[3] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The pascal chime speech separation and recognition challenge," *Computer Speech and Language, Special Issue on Multisource Environments*, Jan. 2012.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443 – 445, Apr 1985.

[5] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 45 –48.

[6] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*, John Wiley & Sons, 2006.

[7] P. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, 2007.

[8] P. Mowlaee, R. Saeidi, and R. Martin, "Model-driven Speech Enhancement for Multisource Reverberant Environment: Signal Separation Evaluation Campaign (SiSEC 2011)," in *Proc. the 10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA)*, 2012, pp. 454–461.

[9] J. R. Hershey, S. J. Rennie, P. A. Olsen, and T. T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 45–66, Jan. 2010.

[10] P. Mowlaee, R. Saeidi, M. Christensen, Z. Tan, T. Kinnunen, P. Fränti, and S. Jensen, "A joint approach for single-channel speaker identification and speech separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 9, pp. 2586 – 2601, 2012.

[11] P. Mowlaee, *New Stategies for Single-channel Speech Separation*, Ph.D. thesis, Institut for Elektroniske Systemer, Aalborg Universitet, 2010.

[12] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source-filter based single channel speech separation using pitch information," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 242–255, Feb. 2011.

[13] M. Stark and F. Pernkopf, "On optimizing the computational complexity for VQ-based single channel source separation," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 237 –240.

[14] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, 2003, pp. 1009–1012.

[15] A. Nadas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, 1989.

[16] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, The Springer International Series in Engineering and Computer Science Series. Kluwer, 1992.

[17] "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms, etsi es 201 108 v1.1.3," 2003.

[18] L.E. Boucheron, P.L. De Leon, and S. Sandoval, "Low bit-rate speech coding through quantization of mel-frequency cepstral coefficients," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 610 –619, Feb. 2012.

[19] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[20] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Audio source separation -," in *10th Int. Conf. on Latent Variable Analysis and Signal Separation*, 2012, pp. 414–422.

[21] P. Mowlaee and R. Saeidi, "Target speaker separation in a multisource environment using speaker-dependent postfilter and noise estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[22] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[23] M. G. Christensen and P. Mowlaee, "A new metric for VQ-based speech enhancement and separation," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, 2011, pp. 4764–4767.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *speech communication*, vol. 2, pp. 749–752, Aug. 2001.

[25] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462 –1469, July 2006.

[26] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, may 1996, vol. 2, pp. 733 –736.

[27] E. Vincent, J. Barker, S. Watanabe, J. Le roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 45 –48.

[28] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, dec 1997, pp. 347 –354.

[29] J. A. Morales-Cordovilla, Ning Ma, V. Sanchez, J. L. Carmona, A.M. Peinado, and J. Barker, "A pitch based noise estimation technique for robust speech recognition with missing data," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, may 2011, pp. 4808 –4811.

[30] J. A. Morales-Cordovilla, A. M. Peinado, V. Sanchez, and J. A. Gonzalez, "Feature extraction based on pitch-synchronous averaging for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 640 –651, march 2011.