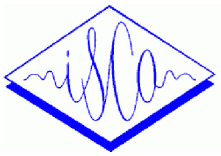

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Speaker Independent Single Channel Source Separation Using Sinusoidal Features

Shivesh Ranjan[†], *Karen L. Payton*[†], *Pejman Mowlaee*[‡]

[†] Electrical & Computer Engineering Dept., University of Massachusetts Dartmouth, North Dartmouth, MA 02747, USA

[‡] Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum, Germany

sranjan@umassd.edu, kpayton@umassd.edu, pejman.mowlaee@rub.de

10.21437/Interspeech.2012-430

Abstract

Model-based approaches to achieve Single Channel Source Separation (SCSS) have been reasonably successful at separating two sources. However, most of the currently used model-based approaches require pre-trained speaker specific models in order to perform the separation. Often, insufficient or no prior training data may be available to develop such speaker specific models, necessitating the use of a speaker independent approach to SCSS. This paper proposes a speaker independent approach to SCSS using sinusoidal features. The algorithm develops speaker models for novel speakers from the speech mixtures under test, using prior training data available from other speakers. An iterative scheme improves the models with respect to the novel speakers present in the test mixtures. Experimental results indicate improved separation performance as measured by the Perceptual Evaluation of Speech Quality (PESQ) scores of the separated sources.

Index Terms: single channel, source separation, speaker independent, sinusoidal features.

1. Introduction

Source Separation systems have become vital components of most computer audition systems. Without preprocessing by a source separation component, most speech recognition and speaker verification/recognition systems fail to work satisfactorily on mixtures of speech signals. Of all the different categories of source separation, the most difficult one is the so-called underdetermined case, which occurs when there are fewer observation channels than sources contributing to the speech mixtures. A particularly difficult case within this category is the situation when only one channel is available to observe the speech mixture, known as the Single Channel Source Separation (SCSS) problem. Even for the relatively simple case of instantaneous mixtures, this problem is ill-conditioned as the mixing matrix is non invertible.

The 2006 Speech Separation Challenge (SSC) resulted in an extensive comparison of different state-of-the-art SCSS systems in terms of automatic speech recognition accuracy, under similar testing conditions [1]. Several participants used model-based approaches to separate the binary mixtures [1–3]. One model-driven approach, using sinusoidal features [4, 5], was demonstrated by Mowlaee et al. [6–10]. They achieved better separation performance, as measured by Perceptual Evaluation of Speech Quality (PESQ) scores, by using speaker dependent sinusoidal features codebooks [10] than the model-based approaches of Weiss and Ellis [2]. The separation results reported in [10] were also comparable to Hershey et al. [3]. Relaxing the speaker-dependent codebook requirement, by using gender dependent codebooks, led to a decrease in separation performance [6, 7]. Weiss and Ellis demonstrated a speaker-independent approach to SCSS by creating speaker-specific models using STFT-based codebooks on a subset of speakers but then using the models to derive models of novel speakers whose utterances had been withheld [2].

In this report, a speaker-independent approach to SCSS is presented that builds on the strategy used by Weiss and Ellis. Rather than using HMMs trained with the log power spectrum though, sinusoidal features based Vector Quantized (VQ) codebooks are used to perform speaker independent SCSS. The initial speaker codebooks are generated using a best Eigenweights search, followed by construction of initial versions of the source utterances present in the test speech mixtures. An iterative scheme is then applied to the initially-developed codebooks to adapt them to the speakers present in the mixtures. The result is improved estimates of the individual speakers utterances. The improvements in separation are measured by the increase in PESQ scores of the reconstructed sources after the iterations. Further, building on the idea that a single channel recording is likely to contain a sequence of different mixtures of talkers, a novel Successive Mixtures Codebook Replacement (SMCR) algorithm is presented, which uses the codebooks developed after separating one mixture, as the initial codebooks for separating the next mixture, to obtain even better estimates of the sources.

In the next section, the use of sinusoidal features in source separation is introduced and the generation of sinusoidal codebooks is also briefly described. Section 3 provides a description of the proposed approach, and also discusses some modifications that have been incorporated to obtain better SCSS performance. The experimental results are presented in Section 4, and Section 5 summarizes what has been achieved and future work.

2. Sinusoidal Features for Source Separation

2.1. Sinusoidal signal representation

Exploiting the quasi-stationary nature of a speech frame, McAulay and Quatieri proposed that a linear combination of sinusoids could be used to approximate speech [4]. In the current application, within a frame, the speech of talker k, $s_k(n)$, is assumed to be composed of L sinusoidal components, plus

The work of Pejman Mowlaee is funded by the European Commission within the Marie Curie ITN AUDIS, grant PITNGA-2008-214699.

additive noise, $e_k(n)$:

$$s_k(n) = \sum_{i=1}^{L} A_{k,i} \cos(n\omega_{k,i} + \phi_{k,i}) + e_k(n), \quad (1)$$

where $n \in [0, N-1]$, and N is the frame length. Each cosine term is parameterized by amplitude $A_{k,i}$, frequency $\omega_{k,i}$ and phase, $\phi_{k,i}$. All three L-length parameters were extracted from the Short-Term Fourier Transform (STFT) of the speech frame. To account for the logarithmic sensitivity of human auditory perception, the sinusoidal features used in the present work were derived from frequency bins uniformly distributed along the Mel scale rather than linearly-spaced frequency bins [8]. Details of how these parameters were evaluated can be found in Mowlaee et al. [6].

Each sinusoid is then represented by a complex exponential sequence $\mathbf{v}_{k,i} = [1 \quad e^{j\omega_{k,i}} \quad \dots \quad e^{j\omega_{k,i}(N-1)}]^T$, where $i \in [1, L]$. An amplitude vector, $\mathbf{a}_k$, containing the products of the selected magnitude and phase vectors, for each bin was defined as $\mathbf{a}_k = [A_{k,1}e^{j\phi_{k,1}} \; A_{k,1}e^{-j\phi_{k,1}} \; \dots \; A_{k,L}e^{j\phi_{k,L}} \; A_{k,L}e^{-j\phi_{k,L}}]^T$. The kth speaker is represented in terms of the sinusoidal features as $\hat{s}_k = \mathbf{V}_k^T \mathbf{a}_k$, where $\mathbf{V}_k$ is a $2L \times N$ Vandermonde matrix given as $\mathbf{V}_k = [\mathbf{v}_{k,1} \quad \mathbf{v}^*_{k,1} \quad \mathbf{v}_{k,2} \quad \mathbf{v}^*_{k,2} \quad \dots \quad \mathbf{v}_{k,L} \quad \mathbf{v}^*_{k,L}]^T$, where $\mathbf{v}^*_{k,i}$ denotes the complex conjugate of $\mathbf{v}_{k,i}$. An N-length frame of the mixture, $z(n)$, comprised of K speakers can be expressed as

$$z(n) = \sum_{k=1}^{K} s_k(n), \quad n = 0, \cdots, N-1 \quad (2)$$

For the current study, only the case of $K = 2$ speakers was considered.

2.2. Sinusoidal features based source separation

Speaker-specific split-VQ codebooks were generated for each speaker according to the scheme described by Mowlaee et al. [6]. The task of separating the sources present in a binary mixture translates to finding the set of unknowns $\{\hat{A}_{1,i}, \hat{A}_{2,i}, \hat{\mathbf{v}}_{1,i}, \hat{\mathbf{v}}_{2,i}\}_{i=1}^{L}$ from the codebook $\mathbb{C}_1$ for speaker one and $\mathbb{C}_2$ for speaker two, so that a cost function, taking the code vector distances into account, is minimized. This amounts to solving the following minimization problem

$$\underset{\hat{A}^r_{1,i}, \hat{A}^q_{2,i}, \hat{\mathbf{v}}^r_{1,i}, \hat{\mathbf{v}}^q_{2,i}}{\arg\min} \sum_{i=1}^{L} \|(A_{z,i}\mathbf{v}_{z,i} - \hat{A}^r_{1,i}\hat{\mathbf{v}}^r_{1,i} - \hat{A}^q_{2,i}\hat{\mathbf{v}}^q_{2,i})\|_2^2 \quad (3)$$

where $\hat{A}^r_{1,i}$ and $\hat{\mathbf{v}}^r_{1,i}$ are the rth codevector from the codebook $\mathbb{C}_1$ and $\hat{A}^q_{2,i}$ and $\hat{\mathbf{v}}^q_{2,i}$ are the qth codevector from the codebook $\mathbb{C}_2$. The utterances are reconstructed using the overlap add method (OLA), once the codebook entries are selected.

3. Proposed Approach

3.1. Initial model using Eigenvoice adaptation

We extend the idea of Eigenvoice adaptation (EA), originally proposed by Kuhn et al. [11] and used for SCSS by Weiss and Ellis [2], to incorporate the use of sinusoidal features. Codevectors of each speaker-specific codebook $\mathbb{C}_i$ are arranged in a single row to form a supervector $\mathbf{m}_i$. Supervectors from M different speakers (where $M \neq K$) are then concatenated to form a supervector Matrix $\mathbf{U}$ of M rows which can be expressed as

$$\mathbf{U} = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_M]. \quad (4)$$

We then perform Principal Component analysis (PCA) on $\mathbf{U}$ to yield $M-1$ Eigenvectors, $\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_{M-1}$ Any speaker-specific codebook, $\mathbb{C}_i$, can be represented as a linear combination of the Eigenvectors and the mean supervector, $\bar{\mathbf{m}}$, as

$$\mathbf{m}_i = \sum_{j=1}^{M-1} w_{ij}\hat{\mu}_j + \bar{\mathbf{m}}, \quad (5)$$

where w_{ij} is the corresponding Eigenweight of $\mathbf{m}_i$ with respect to the Eigenvector $\hat{\mu}_j$.

To construct initial estimates of the unknown sources constituting a mixture, Weiss and Ellis [2] quantized the Eigenweights of each of M speakers to three levels, and used the Iroquois speaker identification algorithm [12] to develop initial source models following a bottom up construction. In the current work, using quantized Eigenweights to construct initial amplitude codebooks degraded performance relative to unquantized Eigenweights so they were not quantized. To simplify the current approach, only amplitude codebooks of the novel speakers were created. Reconstruction was performed by evaluating the amplitude codevectors from the codebooks that give minimum distortion against the text mixtures sinusoidal amplitudes, and using the test mixtures frequency and phase vectors. The following algorithm is proposed to develop initial codebooks for the novel speakers present in the mixed signal:

1. For $i \in [1, M]$, construct M speaker codebooks using $\mathbf{m}_{1i} = w_{i1}\hat{\mu}_1 + \bar{\mathbf{m}}$.
2. Select the winner speaker model $\mathbf{m}_{w1}$ from these M speaker models along the lines of the speaker recognition scheme of Weiss and Ellis [2]:
   (a) For each frame of the mixed signal, calculate the minimum distortion with respect to each of the M codebooks $\mathbf{m}_{K1}, \mathbf{m}_{K2}, , \mathbf{m}_{KM}$. ($K$ is initialized to 1).
   (b) Calculate the mean distortion from all the minimum distortions, and set a suitable *threshold* as θ times the mean distortion. The *threshold* is chosen as a small fraction of the mean distortion to highlight the contributions from the dominant speaker.
   (c) If the minimum distortion is less than the *threshold*, the cumulative score for the corresponding codebook $\mathbf{m}_{Ki}$ is increased by the minimum distortion. Otherwise it is increased by unity.
   (d) Repeat (a)-(c) for all frames of the mixed signal.The speaker codebook corresponding to the minimum score, denoted as $\mathbf{m}_{Wk}$, is declared the winner of the kth stage.
3. Construct $\mathbf{m}_{2i} = w_{i2}\hat{\mu}_2 + \mathbf{m}_{W1}$ for $i = 1, 2, \cdots, M$ and use the speaker identification system outlined in step 2. to get $\mathbf{m}_{W2}$.
4. Define $\mathbf{m}_{3i} = w_{i3}\hat{\mu}_3 + \mathbf{m}_{W2}$, and identify the speaker model $\mathbf{m}_{W3}$ using step 2.
5. Keep repeating this strategy to ultimately get $\mathbf{m}_{W(M-1)}$ using the $(M-1)$th Eigenvector and the corresponding Eigenweights $W_{i(M-1)}$.

This algorithm uses a modified version of the VQ-based speaker recognition algorithm outlined in Jialong et al. [13]. A penalty term has been included in step 2-(c) to limit contributions to only those from the most dominant source as outlined by Rennie et al. [12].

The EA approach develops an initial codebook for the most dominant speaker present in the mixture. For all the speech mixtures in which the target speaker is mixed at target Signal to mask Signal Ratios (SSR) greater than 0 dB, the algorithm essentially returns an initial model for the target speaker. On the other hand for all negative target to mask SSRs, the returned initial model is that of the mask speaker. In its present form, the approach is unable to develop initial models for the less dominant speaker. To reconstruct an estimate of the dominant speakers utterance, an approximate version of equation (3) is used in which contributions from the less dominant of the two speakers are ignored.

3.2. Iterative Codebook Update Algorithm

To improve the estimate of the reconstructed source, the preliminary codebook was modified using the codebook update algorithm proposed by Gersho and Yano [14]. First, the sinusoidal features of the reconstructed source are extracted.

Then, the minimum distortions (squared distances) for each frame with respect to the codevectors and the centroid (mean) of all the frames mapping to a particular codevector $\mathbb{C}_i$ are computed. If the average distortion using $\mathbb{C}_i$ is greater than λ times the mean distortion, for some threshold λ, $\mathbb{C}_i$ is replaced by the centroid of the frames that were originally mapped to $\mathbb{C}_i$. The source then is reconstructed using the modified codebook. This algorithm is repeated 25 times. The source is reconstructed using the mixed utterance and the updated codebook. Assuming prior knowledge of SSR, an estimate of the less dominant speaker is also reconstructed. This is achieved by using the prior knowledge of SSR to transform the mixture into a 0 dB mixture, and subtracting the estimate of the dominant speaker from the mixture.

3.3. Successive Mixtures Codebook Replacement

An actual single-channel recording may include more than one mixed utterance of the two talkers. To exploit the possible availability of multiple mixed signals, a second enhancement technique is proposed. The novel Successive Mixtures Codebook Replacement (SMCR) strategy relies on the fact that, since a codebook for the dominant talker has already been created using a single utterance, successive utterances involving the same talker can be separated using the previously-developed codebook as the initial codebook. Further updates can be carried out on this initial codebook to get an updated version with increased separation performance. This new approach is approximately 5 times faster than using EA algorithm each time a new mixed utterance is encountered.

4. Experimental Results

The GRID corpus, provided in the speech separation and recognition challenge [1], was used as the speech database. Codebooks of 20 speakers from the corpus were created. As suggested by Mowlaee et al. [6], 500 utterances of each speaker were down-sampled from 25 KHz to 8 KHz and used to create amplitude codebooks with 2048 codevectors. For framing, a 32 ms von Hann window was used with a frame-shift of 8 ms. Throughout the simulations presented here, the frequency range of interest was 60,3850 Hz uniformly distributed along the Mel Scale. The order of each codevector was kept at 50 as Mowlaee reported good performance with codevectors of this order. Each of the 2048×50 codevectors was then arranged as a single supervector from which a supervector matrix was formed. Nineteen Eigenvectors and the corresponding Eigenweights for each speaker were then evaluated.

To test the validity of the proposed approach, 20 utterances from four talkers not among the 20 used in developing codebooks (7, 8, 17 and 21 of the GRID Corpus) were mixed to form 120 binary mixtures at seven SSRs (-9, -6, -3, 0, 3, 6 and 9 dB).The dominant source was reconstructed according to the two proposed schemes: Eigenvoice Adaptation with Iterative Codebook Update (EA with iteration) and SMCR.

Perceptual Evaluation of Speech Quality (PESQ) scores [15] of the reconstructed sources with respect to the original sources were used to assess the performance of the proposed SCSS system. PESQ scores have been reported in prior SCSS performance evaluations [6, 10], and have been shown to be highly correlated with overall speech quality [16]. Further, Mowlaee et al. have demonstrated that PESQ scores correspond well to subjective opinions of separation quality [10, 17].

Figure 1, top panel compares, as a function of SSR, the perceived quality of EA iterated model results and the SMCR approach against the ideal scenario where prior target and mask speaker-specific codebooks are available and the optimal indices in (3) are available (from [6]). Two ideal speaker-dependent scenarios are depicted. One shows the separation achieved using mixture frequency and phase vectors while the other depicts an upper bound on performance if the phase is known. For negative SSRs, the figure depicts equal PESQ scores for both the EA with iteration and SMCR approaches because neither approach specifically deals with the less dominant speaker. These scores are evaluated after reconstructing the target speaker by the approximate method mentioned previously for the less dominant speaker, and reported as such for both approaches.

Figure 1, bottom panel depicts the same algorithms PESQs for the mask speaker. It can be seen that the SMCR strategy improves PESQ scores of the dominant speaker over the Eigenvoice adaptation approach with iteration. As should be expected, neither performs as well as methods based on the target and mask speakers having speaker-specific models already in the database. For SSRs greater than -3dB, the iterative model results outperform the initial model, evident by higher scores for both the EA with iteration and SMCR approaches. From Figure 1 it is observed that the PESQ results shown for the target and masker are nearly mirror images of each other. This happens as we include symmetric range for SSR to consider both target and masker roles at each utterance. If we average the results over whole speakers and SSRs in the corpus, the curves are exact mirrors of each other (similar observations reported in [6, 10]).

5. Discussions and Future Work

The proposed approach extends the Eigenvoice adaptation technique to sinusoidal features, leading to speaker-independent SCSS. The proposed Successive Mixtures Codebook Replacement (SMCR) approach is successful in separating the dominant source with good accuracy, as shown by the PESQ scores. Additionally, it also benefits from being computationally inexpensive. Most model-based approaches in the SSC were speaker dependent although one speaker independent approach was demonstrated by Weiss and Ellis [2], and two other groups

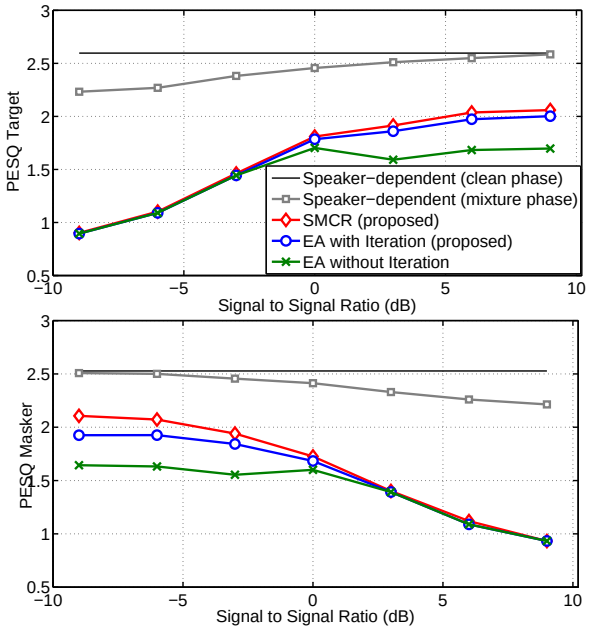

Figure 1: Comparison of PESQ scores for reconstruction of Speaker 1 (Target) a) and Speaker 2 (Mask) b) at different SSRs using two proposed techniques where speaker is unknown against two known/ speaker-dependent models (from [6]).25 iterations are used. The PESQ scores obtained using only initial Eigenvoice adapted models without the iterations are also shown.

used partially speaker-independent approaches which required training data for one of the speakers [18, 19]. This paper presents an approach that relaxes the speaker dependence constraint completely. Further, no assumptions about talker gender have been made. The test mixtures included mixed and same gender mixtures in equal proportions.

According to our PESQ evaluations, the perceived signal quality results are relatively low, in line with the results reported in [10] for the state-of-the-art single-channel separation methods that participated in the challenge [1]. Therefore, one remaining issue, is to improve the quantization performance of the sinusoidal coders used. One drawback of the proposed approach is that it does not reconstruct the less dominant source with good accuracy. While no a priori knowledge of the SSR is required for the dominant speaker, reconstruction of the less dominant speaker requires this information. Future work will attempt to address both these issues. Improvements in reconstruction using iterative approaches will also be explored. Possible other applications of the approach in areas such as music-voice separation will also be investigated by using a generic sinusoidal features model for voice.

Finally, in this work, we focused on the scenario where no additive background noise was present in the single-channel mixed signals. In real applications, this is not the case. We aim at adapting the proposed method to real scenarios, e.g., the realistic and natural reverberant environments using many simultaneous sound sources as recently presented in [20].

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] R. J. Weiss and D. P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 16–29, 2010.

[3] 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.

[4] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[5] P. Mowlaee and A. Sayadiyan, "Model-based monaural sound separation by split-VQ of sinusoidal parameters," in *Proc. European Signal Processing Conf.*, Aug. 2008.

[6] P. Mowlaee, M. Christensen, and S. Jensen, "New results on single-channel speech separation using sinusoidal modeling," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 5, pp. 1265 – 1277, 2011.

[7] P. Mowlaee, M. G. Christensen, Z. H. Tan, and S. H. Jensen, "A MAP criterion for detecting the number of speakers at frame level in model-based single-channel speech separation," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2010, pp. 538–541.

[8] P. Mowlaee, M. G. Christensen, and S. H. Jensen, "Improved single-channel speech separation using sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2010, pp. 21–24.

[9] P. Mowlaee, R. Saeidi, Z. H. Tan, M. G. Christensen, P. Fränti, and S. H. Jensen, "Joint single-channel speech separation and speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2010, pp. 4430–4433.

[10] P. Mowlaee, R. Saeidi, Z. H. Tan, M. G. Christensen, T. Kinnunen, P. Fränti, and S. H. Jensen, "A joint approach for single-channel speaker identification and speech separation," *IEEE Trans. Audio, Speech, and Language Process. to appear*, Jun. 2012.

[11] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 8, no. 6, pp. 695–707, 2000.

[12] S. Rennie, P. Olsen, J. Hershey, and T. Kristjansson, "The Iroquois model: Using temporal dynamics to separate speakers," in *Workshop on Statistical and Perceptual Audio Processing (SAPA)*, Sept. 2006.

[13] Jialong He, Li Liu, and G. Palm, "A new codebook training algorithm for vq-based speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1997, pp. 1091 –1094.

[14] A. Gersho and M. Yano, "Adaptive vector quantization by progressive codevector replacement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1985, pp. 133 – 136.

[15] 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.

[16] Yi Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 1, pp. 229 –238, Jan. 2008.

[17] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin, "Subjective and objective quality assessment of single-channel speech separation algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Mar. 2012, pp. 69–72.

[18] J. Barker, M. Ning, A. Coy, and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 94 – 111, 2010.

[19] J. Ming, T. J. Hazen, and J. R. Glass, "Combining missing-feature theory, speech enhancement, and speaker-dependent/-independent modeling for speech separation," *Elsevier Computer Speech and Language*, vol. 24, no. 1, pp. 67 – 76, 2010.

[20] 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, pp. 1918–1921.