EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

An iterative model-based approach to cochannel speech separation

Ke Hu[1*] and DeLiang Wang[1,2]

Abstract

Cochannel speech separation aims to separate two speech signals from a single mixture. In a supervised scenario, the identities of two speakers are given, and current methods use pre-trained speaker models for separation. One issue in model-based methods is the mismatch between training and test signal levels. We propose an iterative algorithm to adapt speaker models to match the signal levels in testing. Our algorithm first obtains initial estimates of source signals using unadapted speaker models and then detects the input signal-to-noise ratio (SNR) of the mixture. The input SNR is then used to adapt the speaker models for more accurate estimation. The two steps iterate until convergence. Compared to search-based SNR detection methods, our method is not limited to given SNR levels. Evaluations demonstrate that the iterative procedure converges quickly in a considerable range of SNRs and improves separation results significantly. Comparisons show that the proposed system performs significantly better than related model-based systems.

1 Introduction

In daily listening environments, noise corrupts speech and creates substantial difficulty for various applications such as hearing aid design and automatic speech recognition. When noise is a nonspeech signal, existing algorithms often exploit the intrinsic properties of speech/noise for segregation. However, when interference is another voice, the generic properties of speech signals alone are insufficient for separation, and current methods also utilize speaker characteristics. The problem of separating two voices from a single mixture is often referred to as cochannel speech separation. Depending on the information used in cochannel speech separation, we can classify the algorithms into two categories: unsupervised and supervised. In unsupervised methods, speaker identities and pretraining with clean speech are not available, while supervised methods often assume both.

Motivated by human perceptual principles, computational auditory scene analysis (CASA) aims to segregate a voice of interest by exploiting inherent features of speech such as pitch and common onsets [1]. CASA methods are typically unsupervised. For example, pitch and amplitude modulation are utilized to separate voiced portions of cochannel speech, and the estimated pitches in neighboring frames are grouped using pitch continuity [2]. To group temporally disjoint time-frequency (T-F) regions, a system [3] employs speaker models to perform a joint estimation of speaker identities and sequential grouping. Later in [4], the system is extended to handle unvoiced speech based on onset/offset-based segmentation [5] and model-based grouping. Similarly, another CASA system extracts speaker homogeneous T-F regions and employs speaker models and missing data techniques to group them into speech streams [6]. Note that the aforementioned methods use speaker models for sequential grouping, or to group temporally disjoint speech regions, and thus are not completely unsupervised. A recent system [7] applies unsupervised clustering to group speech regions into two speaker groups by maximizing the ratio of between- and within-cluster distances.

Supervised methods often formulate separation as an estimation problem, i.e., given an input mixture, one estimates the two underlying speech sources. To solve this underdetermined equation, a general approach is to represent the speakers by two trained models, and the two patterns (each from one speaker) best approximating the

*Correspondence: huk@cse.ohio-state.edu
[1] Department of Computer Science and Engineering,The Ohio State University, 2015 Neil Ave., Columbus, OH 43210, USA
Full list of author information is available at the end of the article

mixture are used to reconstruct the sources. For example, an early study [8] employs a factorial hidden Markov model (HMM) to model a speaker, and a binary mask is generated by comparing the two estimated sources. In another system [9], Gaussian mixture models (GMM) are used to describe speakers, and speech signals are estimated by a minimum mean-square estimator (MMSE). In MMSE estimation, the posterior probabilities of all Gaussian pairs are computed and used to reconstruct the sources (see [10] for a similar system). The GMM-based methods [9,10] do not model the temporal dynamics of speech. A layered HMM model is proposed to model both temporal and grammar dynamics by transition matrices [11]. A 2-D Viterbi decoding technique is used to detect the most likely Gaussian pair in each frame, and a maximum *a posteriori* (MAP) estimator is used for estimation. In a speaker-independent setting, Stark et al. [12] propose a factorial HMM to model vocal tract characteristics and use detected pitch to reconstruct speech sources. In addition to these methods, other models are applied to capture speakers, including eigenvectors to model and adapt speakers [13], nonnegative matrix factorization-based models [14,15], and sinusoidal models [16].

As pointed out in [9], one problem the model-based methods face is generalization to different input signal-to-noise ratio (SNR) levels (note here that we consider interfering speech as noise). The system [9] does not address this problem and assumes that test mixtures have the same energy level as the training mixtures. Further, the system is designed to only handle 0-dB mixtures. Similarly, a conditional random field-based method in [17] is only applied to separate 0-dB speech mixtures. The factorial HMM system [12] employs a quantile filtering to estimate a gain for each frame and then uses that to adjust the corresponding mean vector in a codebook. Radfar and Dansereau [18] propose a search-based method to detect the input SNR, but one has to specify the search range. In this method, different gains are hypothesized, and the one maximizing likelihood of the whole utterance is taken as the estimate. Radfar et al. [19] use a quadratic function to approximate the likelihood function of a factorial HMM and employ an iterative approach to estimate the gain. The HMM system [11] detects the model gains jointly with the speaker identities given a closed set of speakers and uses an expectation-maximization (EM) algorithm to further adapt the gains. However, the complexity of gain adaptation is quadratic to the number of states, and the convergence speed of the EM algorithm is unknown. Sinusoidal models are also employed to model speakers for joint speaker separation and identification [20], and SNR estimation can be achieved by adapting a universal background model using segregated speech [21].

In this work, we propose an iterative algorithm to generalize to different input SNR conditions given speaker identities. Building on the GMM system [9], we first incorporate temporal dynamics using transition matrices [11]. Then, our algorithm estimates initial T-F masks for two speakers by assuming that the input SNR is 0 dB. The initial masks are used to estimate an utterance-level SNR, which is in turn used to adapt the speaker models. Then, the adapted models are used in a new iteration of separation. The above two steps iterate until both input SNR and the estimated masks become stable. Experiments show that it converges relatively fast and is computationally simple. Compared to the method of [19], our method is simpler and can be applied to factorial HMMs as well as other models (e.g., GMMs). In addition, our method does not require a search range for the estimated input SNR. Comparisons show that the proposed algorithm significantly outperforms related methods.

The rest of the paper is organized as follows. We first present the basic model in Section 2. Section 3 describes iterative estimation. Evaluation and comparison are given in Section 4, and we conclude the paper in Section 5.

2 Model-based separation

We first introduce speaker models and source estimation methods. Throughout the paper, we denote vectors by boldface lowercase and matrices by boldface uppercase letters. Given two speakers a and b, the time-domain cochannel speech signal is a simple addition of two source speech signals. Decomposing the signals into the T-F domain using a linear filterbank and assuming two source signals are uncorrelated at each channel, we have

$$Y(c,m) = X_a(c,m) + X_b(c,m), \tag{1}$$

where $X_a(c,m)$ and $X_b(c,m)$ denote the power spectrum at the T-F unit of channel c and time frame m of speakers a and b, respectively, and $Y(c,m)$ is the spectrum of the mixture. We then take the logarithm of all entities and use log-max approximation to model the relationship between the mixture and sources: in the log-spectral domain, the mixture at each T-F unit is equal to the stronger source. Thus, (1) can be approximated as

$$y(c,m) \approx \max(x_a(c,m), x_b(c,m)), \tag{2}$$

where $x_a(c,m)$, $x_b(c,m)$, and $y(c,m)$ represent the logarithms of $X_a(c,m)$, $X_b(c,m)$, and $Y(c,m)$, respectively. The log-max approximation is originally proposed in [22] to describe the mixing process of speech and noise in robust speech recognition and is later employed in two-speaker separation. A mathematical analysis in [9] shows that the approximation error in (2) is reasonable, but more accurate approximations exist that take both amplitude and phase into consideration [23].

2.1 Speaker models

We use a gammatone filterbank consisting of 128 filters to decompose the input signal into different frequency channels [1]. The center frequencies of the filters spread logarithmically from 50 to 8,000 Hz. Each filtered signal is then divided into 20-ms time frames with 10-ms frame shift, resulting in a cochleagram. The log spectra are computed by taking the element-wise logarithm of the energy in the cochleagram matrix.

Following [9], we build speaker models using GMMs. For each speaker, we build a 128-dimensional GMM from the log spectra of their clean utterances and use a diagonal covariance matrix for each Gaussian for efficiency and tractability. Letting $\mathbf{x}_a$ be the log-spectral vectors of speaker a, the GMM for speaker a can be parameterized as

$$p(\mathbf{x}_a) = \sum_{k=1}^{K} p_a(k) \prod_{c=1}^{128} N(x_a^c; \mu_{a,k}^c, \sigma_{a,k}^c), \tag{3}$$

where K is the number of Gaussians indexed by k, c is the index of frequency channels, and x_a^c is the cth element of $\mathbf{x}_a$. $N(\cdot; \mu_{a,k}^c, \sigma_{a,k}^c)$ denotes a one-dimensional Gaussian distribution with mean $\mu_{a,k}^c$ and variance $\sigma_{a,k}^c$, which correspond to the cth dimension of the kth Gaussian in the GMM. In addition, $p_a(k)$ denotes the prior of kth Gaussian. Similarly, the model of speaker b is

$$p(\mathbf{x}_b) = \sum_{k=1}^{K} p_b(k) \prod_{c=1}^{128} N(x_b^c; \mu_{b,k}^c, \sigma_{b,k}^c). \tag{4}$$

For each speaker, the conditional distribution given a specific Gaussian is a 128-dimensional Gaussian distribution, i.e., $p(\mathbf{x}_a|k_a) = \prod_{c=1}^{128} N(x_a^c; \mu_{a,k_a}^c, \sigma_{a,k_a}^c)$ and $p(\mathbf{x}_b|k_b) = \prod_{c=1}^{128} N(x_b^c; \mu_{b,k_b}^c, \sigma_{b,k_b}^c)$, where k_a and k_b are two Gaussian indices, and $p(x_a^c|k_a)$ and $p(x_b^c|k_b)$ are one-dimensional Gaussians.

Given the above speaker models and the mixing Equation (2), we can derive a per-channel statistical relationship between the mixture and two sources as follows:

$$p(y^c|k_a, k_b) = p_{x_a^c}(y^c|k_a)\Phi_{x_b^c}(y^c|k_b) + p_{x_b^c}(y^c|k_b)\Phi_{x_a^c}(y^c|k_a). \tag{5}$$

Here, we use subscripts x_a^c and x_b^c to differentiate the probability functions for speakers a and b. $\Phi_{x_a^c}(\cdot|k_a)$ and $\Phi_{x_b^c}(\cdot|k_b)$ are their corresponding cumulative distributions. In a probabilistic manner, (5) provides a way of approximating the mixture using two clean speaker models, which in turn can be used to estimate two source signals given the mixture as the observation.

2.2 Source estimation

One method to estimate the sources is the MMSE estimator, which aims to minimize the expectation of the square error between the estimated and underlying true signals given the observations [9]. As a result, for a log-spectral vector $\mathbf{y}$, the cth element of source $\mathbf{x}_a$ can be estimated as

$$\hat{x}_a^c = \int_{-\infty}^{\infty} x_a^c \cdot p(x_a^c|\mathbf{y}). \tag{6}$$

According to the total probability formula, $p(x_a^c|\mathbf{y})$ in (6) can be expanded as follows:

$$p(x_a^c|\mathbf{y}) = \sum_{k_a,k_b} p(k_a, k_b|\mathbf{y}) p(x_a^c|k_a, k_b, y^c). \tag{7}$$

Note that $p(x_a^c|k_a, k_b, y^c)$ here only depends on y^c instead of $\mathbf{y}$ due to the diagonal covariance assumption. The posterior $p(k_a, k_b|\mathbf{y})$ in (7) can be calculated as

$$p(k_a, k_b|\mathbf{y}) = \frac{p_a(k_a)p_b(k_b)p(\mathbf{y}|k_a, k_b)}{\sum_{k'_a,k'_b} p_a(k'_a)p_b(k'_b)p(\mathbf{y}|k'_a, k'_b)}, \tag{8}$$

where $p(\mathbf{y}|k_a, k_b) = \prod_{c=1}^{128} p(y^c|k_a, k_b)$ again because of the diagonal covariance matrix. On the other hand, $p(x_a^c|k_a, k_b, y^c)$ in (7) can be computed by using the Bayes rule:

$$p(x_a^c|k_a, k_b, y^c) = \frac{p(x_a^c, y^c|k_a, k_b)}{p(y^c|k_a, k_b)} \tag{9}$$

$$= \frac{p_{x_a^c}(x_a^c|k_a)p_{x_b^c}(y^c|k_b)}{p(y^c|k_a, k_b)}\delta(x_a^c < y^c) + \frac{p_{x_a^c}(y^c|k_a)\Phi_{x_b^c}(y^c|k_b)}{p(y^c|k_a, k_b)}\delta(x_a^c = y^c). \tag{10}$$

From (9) to (10), the constraint $x_a^c \leq y^c$ and the log-max assumption are used, and a detailed derivation can be found in [22]. We then incorporate (8) and (10) to (7) and combine with (6) to estimate the source speaker a

$$\hat{x}_a^c = \sum_{k_a,k_b} \frac{p(k_a, k_b|\mathbf{y})}{p(y^c|k_a, k_b)} \{p_{x_b^c}(y^c|k_b)[\,\mu_{a,k_a}^c \Phi_{x_a^c}(y^c|k_a) - \sigma_{a,k_a}^c p_{x_a^c}(y^c|k_a)] + \Phi_{x_b^c}(y^c|k_b)p_{x_a^c}(y^c|k_a)y^c\}. \tag{11}$$

The MMSE estimate of speaker b can be computed similarly.

In addition to directly estimating the sources, we estimate a soft mask for speaker a as

$$p(x_a^c > x_b^c|\mathbf{y}) = \sum_{k_a,k_b} p(k_a, k_b|\mathbf{y}) \cdot p(x_a^c > x_b^c|y^c, k_a, k_b) = \sum_{k_a,k_b} p(k_a, k_b|\mathbf{y}) \cdot \frac{p_{x_a^c}(y^c|k_a)\Phi_{x_b^c}(y^c|k_b)}{p(y^c|k_a, k_b)}. \tag{12}$$

Note that the soft mask for speaker b is $p(x_a^c \leq x_b^c|\mathbf{y}) = 1 - p(x_a^c > x_b^c|\mathbf{y})$. In [9], the soft mask is found to perform consistently better than a binarized mask.

An alternative to the MMSE estimator is a MAP estimator. The essence of MAP estimation is similar to MMSE, but instead of using every pair of Gaussians in (7), it only uses the most likely Gaussian pair

$$\{k_a^*, k_b^*\} = \arg\max_{k_a,k_b} p(k_a, k_b|\mathbf{y}), \tag{13}$$

where k_a^* and k_b^* correspond to the pair of Gaussians yielding the highest posterior probability among all possible pairs. The estimate of source signals can be computed similarly to (11) but using only k_a^* and k_b^*. A soft mask can also be derived like (12) using only k_a^* and k_b^*. In experiments, we find that the performance of the MAP estimator is similar to that of MMSE, mainly because at each frame, one pair of Gaussians often approximates the mixture much better than others.

2.3 Incorporating temporal dynamics

The cochannel speech separation system in [9] models speaker characteristics using GMMs and ignores the temporal information of speech signals. A natural extension to the GMMs to incorporate temporal dynamics is using a factorial HMM model. Specifically, for each speaker, we can estimate the most likely Gaussian index for each frame in a clean utterance using a MAP estimator. Each utterance thus generates a sequence of Gaussian indices. The transitions between all neighboring Gaussian indices are then used to build a 2-D histogram, which can then be normalized to produce a transition matrix [11].

In the factorial HMM system, the hidden states of the two HMMs at each frame are the most likely Gaussian indices of two speakers. While the detection of the Gaussian indices is based on only individual frames in a GMM-based model, a 2-D Viterbi search is used in [11] to find the most likely Gaussian index sequences. Specifically, the 2-D Viterbi integrates all frames and the transition information across time to find the most likely two Gaussian sequences, each of which corresponds to one speaker [24].

We use $\delta_t(k_a, k_b)$ to denote the highest probability along a single path (i.e., a sequence of state pairs) accounting for the first t frames and ending at state k_a, k_b

$$\delta_t(k_a, k_b) = \max_{s_a^1, s_b^1, \ldots, s_a^{t-1}, s_b^{t-1}} p(s_a^1, s_b^1, \ldots, s_a^t = k_a, s_b^t = k_b, \mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_t|\lambda), \tag{14}$$

where s_a^t and s_b^t denote the hidden states of speakers a and b at time frame t, respectively, and λ represents the factorial HMM. (14) can be computed iteratively by

$$\delta_t(k_a, k_b) = \max_{k_a', k_b'} \delta_{t-1}(k_a', k_b') \cdot p(k_a|k_a') \cdot p(k_b|k_b') \cdot p(\mathbf{y}_t|k_a, k_b), \tag{15}$$

where $p(k_a|k_a')$ is the transition probability of speaker a from state k_a' to k_a, and $p(k_b|k_b')$ is that of speaker b. $p(\mathbf{y}_t|k_a, k_b)$ can be calculated similarly as in (8). The optimal Gaussian index sequences are detected by a 2-D Viterbi decoding [24], and the MAP estimator is used for estimating sources.

In (15), an exhaustive search for each pair of k_a and k_b across T frames has a complexity of $O(TK^4)$, where K is the number of Gaussians for each speaker and T is the number of frames. It is time consuming if K is relatively large. In our study, we use a beam search to speed up the process (see also [25]). Given a beam width of W, we only search for the W most likely previous state pairs (i.e., k_a' and k_b' in (15)), and the time complexity is reduced to $O(TWK^2)$. The results presented in Section 4 indicate that a beam width of 16 gives a comparable performance to the exhaustive search.

3 Iterative estimation

As mentioned in Section 1, model-based methods such as [9] face the difficulty of generalizing to different mixing conditions. It is partly because the GMMs are trained using log-spectral vectors and hence are sensitive to the overall speech energy. More importantly, if the GMMs of two speakers are trained using clean utterances at certain energy levels, in testing they need to be adjusted according to the input SNR. In [9], mixtures with nonzero input SNR are separated using unadjusted models, but the performance is worse.

We propose to detect the input SNR and use that to adapt the speaker models and re-estimate the sources. To estimate the input SNR from the mixture, one has to first have some source information. Thus, SNR detection and source estimation become a chicken-and-egg problem, i.e., the performance of one task depends on the success of the other. One general approach to deal with this type of problem is to perform an iterative estimation (e.g., [2]). In the initial stage of the iterative procedure, we apply the unadapted speaker models to obtain initial separation. Based on the initial source estimates, we calculate the input SNR and use that to adapt the speaker models. The adapted models are in turn used to re-estimate the sources. The two steps iterate until convergence. As an alternative, we also explore a search-based method which jointly estimates sources and the input SNR.

3.1 Initial mask estimation

For a pair of speakers, we first perform an initial estimate by using their models pre-trained using clean utterances at a per-utterance energy level of 60 dB. Initially, the input SNR is assumed to be 0 dB, and a mixture is scaled to an energy level of 63 dB corresponding to the addition of two 60-dB source signals. We use the

2-D Viterbi decoding based on (15) to detect the most likely Gaussian index sequence and then estimate a soft mask of the target speaker using the MAP estimator in Section 2.2.

3.2 SNR estimation and model adaptation

Denoting the estimated soft masks of speakers a and b as $\mathbf{M}_a$ and $\mathbf{M}_b$, respectively, we use them to filter the mixture cochleagram to obtain the corresponding segregated signals. With the mixture cochleagram $\mathbf{E}_y$, the SNR of the target and interferer in the cochleagram domain can be calculated as

$$R = 10\log_{10}\left(\frac{\sum_{c,m}\mathbf{E}_y(c,m)\cdot\mathbf{M}_a(c,m)}{\sum_{c,m}\mathbf{E}_y(c,m)\cdot\mathbf{M}_b(c,m)}\right), \qquad (16)$$

where $\mathbf{M}_a(c,m)$ denotes the ratio of speaker a at the T-F unit of channel c and frame m, and $\mathbf{M}_b(c,m) = 1 - \mathbf{M}_a(c,m)$. R corresponds to the input SNR of the filtered speech signals. As analyzed in [26], due to gammatone filtering which has a certain passband, one usually should compensate for the loss of energy to calculate the SNR of the original time-domain signals. However, in our work, the frequency range of the gammatone filterbank is between 50 and 8,000 Hz, and both target and interference are speech signals with a sampling frequency of 16 kHz. There is thus little energy loss in the filtering process, and the estimated SNR of filtered signals is close to that of the original time-domain signals. Thus, we directly use the SNR of filtered signals in (16) as our estimate.

We then adapt two speaker models to match the estimated input SNR. In particular, the target speaker model (speaker a) is fixed (i.e., trained by using 60-dB clean utterances), and we adapt the interferer model and the mixture. Given an input SNR of R dB, the interfering signal energy level is thus

$$10\log_{10}(\sum_t x_b^2[\,t]\,/T) = 60 - R, \qquad (17)$$

where $x_b[\,t]$ denotes the time-domain speech of speaker b. That is, instead of using 60-dB utterances, the interferer model should be trained using $60 - R$ dB signals, and the original utterances should be scaled by a multiplicative factor of $10^{-R/10}$. Since the difference lies in a constant factor, we can directly scale the parameters of the GMM models, i.e., the mean and variance. Specifically, the means of the interferer GMM are scaled by an additive factor of $\beta = \log(10^{-R/10})$ since log-spectral vectors are used in training, while the variances will remain unchanged because β is an additive factor.

On the other hand, the mixture energy level can be computed by combining the target and interfering signal levels

$$\begin{aligned}10\log_{10}(y^2[\,t]\,/T) &= 10\log_{10}(\sum_t(x_a^2[\,t] + x_b^2[\,t]\,)/T)\\ &= 60 + 10\log_{10}(1 + 10^{-R/10}), \quad (18)\end{aligned}$$

where $y[\,t]$ is the time-domain cochannel signal, and $x_a[\,t]$ is the source signal of speaker a. In the above calculation, we assume that the time-domain target and interfering signal are uncorrelated at each frame. Given (17) and (18), we have adapted the interfering speaker model and the mixture and created a more matched condition for separation.

3.3 Iterative estimation

Given any input mixture, we first obtain the initial mask estimates $\mathbf{M}_{a,0}$ and $\mathbf{M}_{b,0}$ as described in Section 3.1. Given $\mathbf{M}_{a,0}$ and $\mathbf{M}_{b,0}$, we then estimate the input SNR using (16). The estimated SNR is used to adapt the model of speaker b and mixture by (17) and (18), respectively. They are then used together with the target speaker model to re-estimate the soft masks based on the 2-D Viterbi decoding described in Section 2.3 and the MAP estimator in Section 2.2. To get the maximal performance, the iterative process should continue until neither the estimated input SNR nor speaker masks change. However, empirically, we observe that the separation performance becomes stable when the estimated input SNR change is smaller than 0.5 dB. We thus use this as the stop criterion and terminate the estimation process when the difference of estimated input SNRs between two iterations is less than 0.5 dB.

As an illustration, Figure 1a shows a cochleagram of a cochannel signal at −9 dB consisting of two male utterances, where a brighter unit indicates stronger energy. Figure 1b shows the clean target speech and Figure 1c the clean interfering speech. We show the initially segregated target and interferer in Figure 1d,e, respectively, and the final segregated target and interferer are presented in Figure 1f,g, respectively. As shown in the figure, the iterative estimation improves the quality of segregated speech signals.

3.4 An alternative method

In addition to the iterative method, we have also tried a search-based method to jointly estimate the source state sequences and the input SNR. For example, we use a test corpus described in Section 4 and hypothesize the input SNR in a range from −9 to 6 dB with an increment of 3 dB. At each hypothesized input SNR, we adapt the mixture and interfering speaker model according to (17) and (18) and use them to detect state sequences using the 2-D Viterbi decoding, and then estimate the soft masks based on the MAP estimator. For all hypothesized SNR

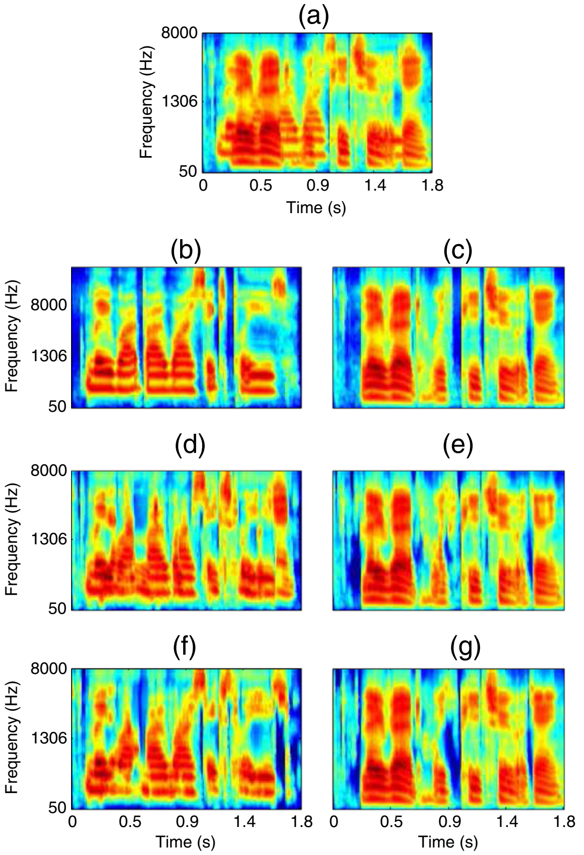

Figure 1 Illustration of separating two male utterances in cochannel conditions. (a) Cochleagram of the cochannel speech with an input SNR of −9 dB. **(b)** Cochleagram of clean target. **(c)** Cochleagram of clean interferer. **(d)** Cochleagram of initially segregated target. **(e)** Cochleagram of initially segregated interferer. **(f)** Cochleagram of segregated target after iterative estimation. **(g)** Cochleagram of segregated interferer after iterative estimation.

conditions, we calculate the joint likelihood of all mixture frames and the Gaussian sequences being generated by the factorial HMM, and the hypothesized input SNR corresponding to the highest likelihood is selected as the detected value. The corresponding state sequence is then used for estimation. We have evaluated the performance of this method using the corpus described in Section 4, and it is about 0.5 dB worse than the iterative method and is computationally more expensive. Note that the discrete SNR range includes the true SNR value in each testing condition to favor the SNR-based search method. How to specify the input SNR levels in search is unclear in practice.

4 Evaluation and comparisons

We use two-talker mixtures in the Speech Separation Challenge (SSC) corpus [27] for evaluation. For each speaker, a 256-component GMM model (i.e., $K = 256$) is trained using all of the speaker's clean utterances in the training set. Here, K is chosen with the consideration of performance and computation complexity. In training, each clean utterance is normalized to a 60-dB energy level, and the log spectra are calculated as described in Section 2.1. An HMM model is then built upon each GMM using the same utterances as described in Section 2.3. We use the test part of the SSC corpus and create two-speaker mixtures at SNRs from −9 to 6 dB (with an increment of 3 dB) for evaluation. We randomly select 100 two-speaker mixtures in each SNR condition for testing. Note that the mixture utterances are the same across different SNRs, and mixtures at opposite SNRs are not symmetric since they are generated by fixing the target and scaling the interfering utterances. The 100 mixtures contain 51 different-gender mixtures, 23 male-male mixtures, and 26 female-female mixtures. All test mixtures are downsampled from 25 to 16 kHz for faster processing.

We evaluate the segregation performance using the SNR gain of the target speaker, which is calculated as the output SNR of segregated target speech subtracted by the corresponding input SNR. For each segregated target, we take its clean speech signal as the ground truth and compute the output SNR as

$$\mathrm{SNR} = 10\log_{10}\left(\sum_{n} x_a^2[t] \,/\, \sum_{n}(x_a[t] - \hat{x}_a[t])^2\right), \tag{19}$$

where $x_a[t]$ and $\hat{x}_a[t]$ are the original clean signals and signals resynthesized from the estimated mask, respectively. Note that a waveform signal can be obtained from a soft mask [1]. In our test conditions, target and interfering speakers are treated symmetrically, e.g., an interferer at 6 dB is considered as a target at −6 dB. Thus, at each input SNR, we calculate the target SNR gain as the average of the target SNR gain at that input SNR and the interferer SNR gain at the negative of that input SNR. For example, the SNR gain at −6 dB is the average of the target SNR gain at the −6 dB SNR and the interferer SNR gain at the 6 dB SNR.

4.1 System configuration

As we mentioned in Section 2.3, an exhaustive 2-D Viterbi search is time consuming, and we use beam search for speedup. The beam width W needs to be chosen to balance the performance and complexity. In Figure 2, we vary W from 1, 4, 16, and 64 to 256, and the corresponding target SNR gains are shown in different curves. For the largest beam width of 256, the beam search already performs comparably to an exhaustive search. On the other hand, a beam width of 1 amounts to a greedy algorithm where we only keep the path with the highest likelihood at each frame. In Figure 2, we observe that when W is

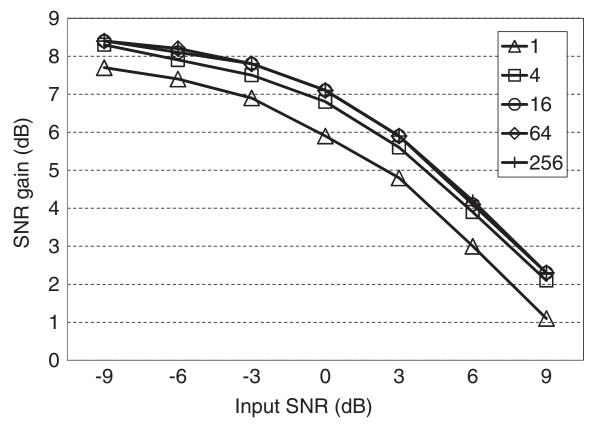

Figure 2 SNR gains of target speaker at different input SNR conditions with beam width varying from 1 to 256.

between 16 and 256, the SNR gains at all conditions are almost the same. However, the gains degrade significantly when W is further reduced. We thus choose W to be 16. Compared to an exhaustive search, the computational complexity is greatly reduced from $O(TK^4)$ to $O(TK^2)$.

Another parameter impacting the system performance is the number of iterations in iterative estimation. In our experiments, we observe that the estimated input SNR and masks become stable quickly. Figures 3 and 4 show the SNR and mask estimation performance, respectively, in terms of the number of iterations. In Figure 3, we measure the SNR estimation performance as the difference of the estimated from the true input SNRs. Each curve in the figure corresponds to the estimation errors at one SNR condition. Before any estimation (i.e., number of iterations $=0$), the input SNR is assumed to be 0 dB and the error is the negative of the underlying true SNR. After the first iteration, the errors decrease significantly for all SNR conditions except for the 0-dB case. This is because at 0 dB, the initial estimate happens to be the same as the true SNR, and any estimation can only deviate away from 0 dB. In this case, we observe that the estimated SNR gets a little worse and then becomes stable. For other SNR conditions, the errors keep decreasing as more iterations are performed, and all of them become stable by the fifth iteration. In Figure 4, we measure the performance of mask estimation by the SNR gain of the segregated target. Initially, the SNR gain is 0 dB, and then, the quality of estimated masks improves substantially after the iteration starts. As shown in the figure, the first iteration brings about 4 to 8-dB improvements for all SNR conditions, and the second iteration mainly improves the performance at -6 and -9 dB (by 1.8 and 3 dB, respectively). The performance at most SNR conditions become stable after three iterations. At -9 dB, the estimated mask gains a small improvement for further iterations. In the experiments, we observe that the estimated masks often become stable when the estimated input SNR changes less than 0.5 dB. Thus, we use this as the stop criterion for iterative estimation. By this criterion, an average of 3 iterations is often enough for convergence.

4.2 Comparisons

We compare the proposed system to related model-based methods, which include the MMSE-based system by [9], a similar system based on a MAP estimator, and an HMM-based system incorporating temporal dynamics. Note that all aforementioned systems are implemented by us in the cochleagram domain for matched comparisons. In training GMMs, we follow [9] and normalize mixtures to have 0 mean and unit variance and use 256 Gaussians in GMMs. We use the soft mask result instead of the direct

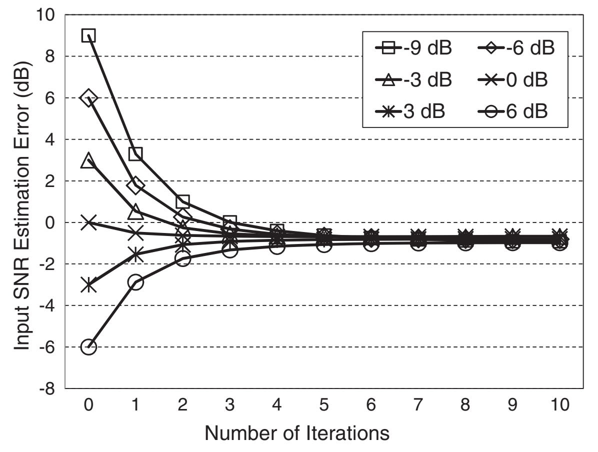

Figure 3 Input SNR estimation error (in dB) with respect to number of iterations used in iterative estimation.

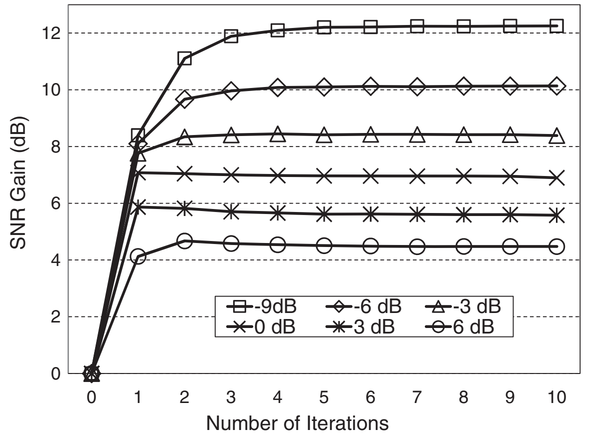

Figure 4 Mask estimation performance in terms of target SNR gain as a function of number of iterations.

estimates in [9] since it gives the best result. The transition probabilities in HMM are calculated according to [11]. The mean SNR gains with 95% confidence intervals of these methods are presented in Figure 5.

-As shown in Figure 5, the proposed system achieves an SNR gain of 11.9 dB at the input SNR of −9 dB, and the gain decreases gradually as the input SNR increases. At 9 dB, the SNR gain is about 3.9 dB. On average, our method achieves an SNR gain of 7.4 dB. Compared to the method of Reddy and Raj, our method performs comparably at 0 dB but significantly better at other input SNRs. For example, the proposed system performs about 2.7 dB better at −9 dB, and the improvement gets smaller as the input SNR gets closer to 0 dB. A similar trend is also observed at positive input SNRs. On average, the proposed system performs 1.2 dB better than the Reddy and Raj method. In the figure, we also show the performance of another MMSE method (black bars), a version of the Reddy and Raj system that does not require the energy levels of training and testing to be the same. In this method, we assume the input SNR to be 0 dB and scale the mixture as described in Section 3.1. As we expect, the performance is a little worse (about 0.3 dB) than the original Reddy and Raj system due to the unmatched signal levels. We also compare to a MAP-based separation method described in Section 2.2. Using only the most likely Gaussian pair for estimation, the MAP method is more efficient than the MMSE method but performs about 0.1 dB worse. Our system performs about 1.6 dB better than the MAP-based method. To isolate the effect of iterative estimation, we have also evaluated the performance of the HMM system alone. As shown in the figure, this method achieves an average SNR gain of about 6.3 dB, about 0.5 dB better than the MAP-based method. This improvement comes from the use of temporal dynamics. Comparing this performance with the proposed system, we get the benefit of iterative estimation, which further increases the SNR gain of the HMM system by about 1.1 dB. In addition, we note that iterative estimation can also be incorporated into other model-based systems. For example, we add iterative estimation to the MMSE method (denoted by as MMSE-iterative in Figure 5) and obtain an improvement of 1.2 dB. Similarly, the MAP-iterative method outperforms the original MAP method by about 1.2 dB. Lastly, to show the upper bound performance of our system, we have utilized the true input SNR and ideal hidden states in estimation. This ideal performance is presented as the HMM ideal in Figure 5. It is about 0.9 dB better than the proposed system, which indicates that our system is close to the ceiling performance.

We have compared to a factorial HMM-based method which is capable of adapting speaker models for separating mixtures with different signal levels [12]. In this method, pitches of two speakers are first estimated by a factorial HMM. Then, vocal tract responses are modeled by vector quantization or nonnegative matrix factorization (NMF) and used with estimated pitches to estimate the source signals. Since the vocal tract responses are normalized in modeling, a gain factor is introduced to scale the source spectra. Specifically, a gain vector is calculated as the difference of the mixture and source spectra, and then quantile filtering is used to select a robust estimate. To compare to this method, we use the criterion of target-to-masker ratio (TMR) as in [12] in the following experiments. In the speaker-dependent case, the method reports about a 6.6-dB gain in terms of TMR at 0-dB input TMR. Specifically, it achieves a TMR of about 7 dB in the same-gender female (SGF) case, 4.5 dB in same-gender male (SGM) case, and 8.3 dB in the different-gender (DG) case. These results correspond to the best performance in a setting where NMF is used for modeling. We evaluate

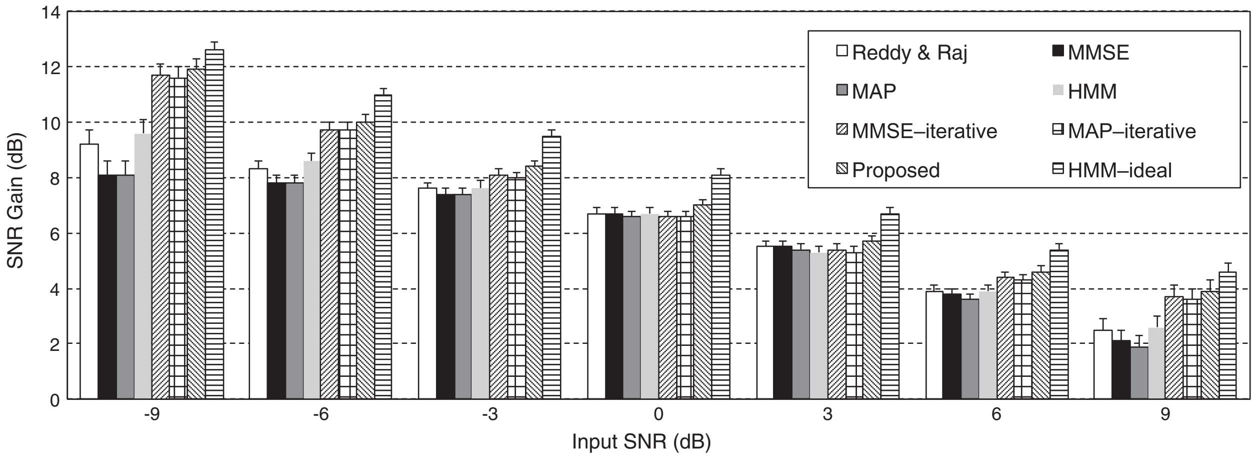

Figure 5 Comparisons to model-based cochannel speech separation algorithms in terms of target SNR gains.

our method using TMR, and the results for 0-dB mixtures are shown in Figure 6. Note that we used the same corpus as in [12], but the exact mixtures may be different. As in [12], we show the TMRs in SGM, SGF, and DG cases separately, and the horizontal lines in the centers of the boxes correspond to means, and the distance between a line and a box boundary depicts standard deviation. The improvements are 9.6, 8.4, and 10.4 dB in the SGF, SGM, and DG cases, respectively, and on average, the improvement is about 9.4 dB. These results show that our system performs substantially better than [12] in all kinds.

In addition to the SNR performance, we also evaluate the system using a hit minus false-alarm (HIT−FA) rate which has been shown to be a good indicator of human speech intelligibility [28]. As in [28], we calculate the hit rate as the percentage of correctly labeled target dominant T-F units and the false alarm (FA) rate as the percentage of incorrectly labeled interferer dominant T-F units. To calculate these rates, we convert the soft masks to binary masks using a threshold of 0.5, i.e., the T-F units with a probability greater than 0.5 are labeled as 1 and 0 otherwise. The HIT−FA rates of our system and the Reddy and Raj system are shown in Figure 7. We observe that the proposed algorithm performs uniformly better than the Reddy and Raj system at all SNR conditions. For our system, the average HIT−FA rate is about 64.4%, and the rates are relatively stable at different input SNR conditions. On average, it is about 7.5% better than the Reddy and Raj system. The performance gap between our system and the Reddy and Raj system are bigger when the input SNR deviates from 0 dB. This again confirms that iterative estimation is effective for generalizing to nonzero SNR mixtures.

Finally, we evaluate our system and compare with the Reddy and Raj system using a short-time objective intelligibility (STOI) [29], which is shown to be highly correlated to human speech intelligibility. As shown in Figure 8, both our method and the Reddy and Raj system perform significantly better than unprocessed mixtures. Our method performs generally better than Reddy and Raj's across a range of SNRs, especially when SNR is far away from 0 dB. As Mowlaee et al. have also evaluated their sinusoidal modeling-based method using STOI [20], it is interesting to draw some comparisons with their performance. Since the exact mixtures in our experiments are different from those in [20], it is more informative to look at the relative STOI improvements over unprocessed mixtures. Roughly speaking, our STOI improvements are comparable to those in [20]. For example, our improvement is about 0.15 at −9 dB and 0.1 at 6 dB, while in [20] (Figure 8), the improvement at −9 dB is about 0.22, but there is no improvement at 6 dB.

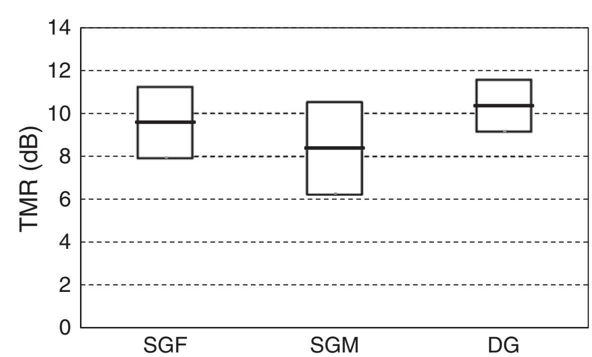

Figure 6 **TMR performance of the proposed algorithm in different kinds of cochannel speech with 0-dB input TMR.**

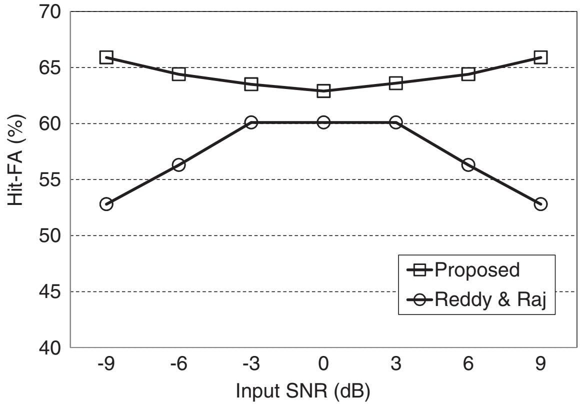

Figure 7 **Comparisons to a model-based speech separation algorithm in terms of HIT−FA rates.**

5 Conclusions

We have proposed an iterative algorithm for model-based cochannel speech separation. First, temporal dynamics is incorporated into speaker models using HMM. We

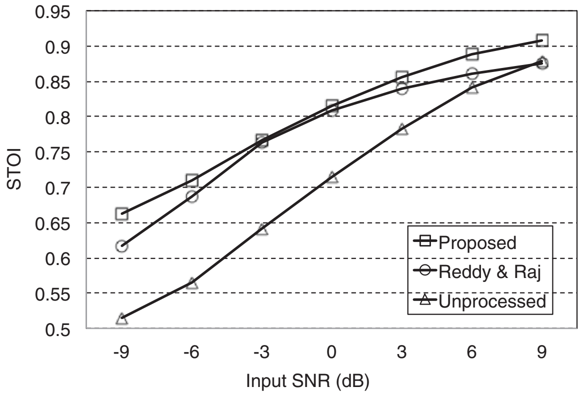

Figure 8 **Comparisons to a model-based speech separation algorithm in terms of STOI scores.**

then present an iterative method to deal with signal level differences between training and test conditions. Specifically, the proposed system first uses unadapted speaker models to segregate two speech signals and detects the input SNR. The detected SNR is then used to adapt the interferer model and the mixture for re-estimation. The two steps iterate until convergence. Systematic evaluations show that our iterative method improves segregation performance significantly and also converges quickly. Comparisons show that it performs significantly better than related model-based methods in terms of SNR gains as well as HIT−FA and STOI scores.

We note that SNR estimation in our system uses the whole mixture, which would not be feasible for real-time applications. However, one can slightly modify it to work in real time. For example, at one frame, one could use only previous frames for Viterbi decoding and SNR detection. The detected SNR could be used to adapt speaker models for separation in later frames and then get updated correspondingly. Such an update may be performed periodically to track the input SNR, and the update frequency would depend on the extent to which the input SNR varies.

In this work, our description is limited to two-talker situations as in related model-based methods. The proposed system could be extended to deal with multi-talker separation problems. For example, the MMSE estimators can be extended to perform three-talker separation according to [9]. As for iterative estimation, one can estimate the energy ratios between multiple speakers instead of the SNR in the two-speaker case and adapt the speaker models accordingly. One issue in multi-talker situations is that the complexity of (13) is exponential to the number of speakers, and a faster decoding method thus needs to be used (e.g., [9,30]).

Competing interests
Both authors declare that they have no competing interests.

Acknowledgements
This research was supported by an AFOSR grant (FA9550-12-1-0130).

Author details
[1]Department of Computer Science and Engineering,The Ohio State University, 2015 Neil Ave., Columbus, OH 43210, USA. [2]Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210, USA.

Received: 24 January 2013 Accepted: 10 June 2013
Published: 26 June 2013

References

1. DL Wang, GJ Brown (eds), *Computational, Auditory Scene Analysis: Principles,Algorithms and Applications*. (Wiley-IEEE Press, Hoboken, 2006)
2. G Hu, DL Wang, A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Trans. Audio, Speech, Lang. Process. **18**, 2067–2079 (2010)
3. Y Shao, DL Wang, Sequential organization of speech in computational auditory scene analysis. Speech Comm. **51**, 657–667 (2009)
4. Y Shao, S Srinivasan, Z Jin, DL Wang, A computational auditory scene analysis system for speech segregation and robust speech recognition. Comput. Speech Lang. **24**, 77–93 (2010)
5. G Hu, DL Wang, Auditory segmentation based on onset and offset analysis. IEEE Trans. Audio, Speech, Lang. Process. **15**, 396–405 (2007)
6. J Barker, N Ma, A Coy, M Cooke, Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. Comput. Speech Lang. **24**, 94–111 (2010)
7. K Hu, DL Wang, An unsupervised approach to cochannel speech separation. IEEE Trans. Audio Speech Lang. Process. **21**, 120–129 (2013)
8. S Roweis, One microphone source separation. Adv. Neural Inf. Process. Syst. **13**, 793–799 (2001)
9. A Reddy, B Raj, Soft mask methods for single-channel speaker separation. IEEE Trans. Audio, Speech, Lang. Process. **15**(6), 1766–1776 (2007)
10. MH Radfar, RM Dansereau, Single-channel speech separation using soft masking filtering. IEEE Trans. Audio, Speech, Lang. Process. **15**(8), 2299–2310 (2007)
11. JR Hershey, SJ Rennie, PA Olsen, TT Kristjansson, Super-human multi-talker speech recognition: a graphical modeling approach. Comput. Speech Lang. **24**, 45–66 (2010)
12. M Stark, M Wohlmayr, F Pernkopf, Source-filter-based single-channel speech separation using pitch information. IEEE Trans. Audio, Speech, Lang. Process. **19**(2), 242–255
13. R Weiss, D Ellis, Speech separation using speaker-adapted eigenvoice speech models. Comput. Speech Lang. **24**, 16–29 (2010)
14. GJ Mysore, P Smaragdis, B Raj, in *Proc. 9th Int. Conf. Latent Variable Analysis and Signal Separation*. Non-negative hidden Markov modeling of audio with application to source separation (Springer Heidelberg, 2010)
15. P Smaragdis, Convolutive speech bases their application to supervised speech separation. IEEE Trans. Audio, Speech, Lang. Process. **15**, 1–12 (2007)
16. P Mowlaee, MG Christensen, SH Jensen, New results on single-channel speech separation using sinusoidal modeling. IEEE Trans. Audio Speech Lang. Process. **19**, 1265–1277 (2011)
17. YT Yeung, T Lee, CC Leung, in *Proc. ICASSP-12 IEEE*. Integrating multiple observations for model-based single-microphone speech separation with conditional random fields (New York, 2012), pp. 257–260
18. MH Radfar, RM Dansereau, in *Proc. WASPAA IEEE*. Long-term gain estimation in model-based single channel speech separation (New York, 2007)
19. MH Radfar, W Wong, RM Dansereau, WY Chan, *Scaled factorial hidden Markov models: a new technique for compensating gain differences in model-based single channel speech separation*, (New York, 2010), pp. 1918–1921
20. P Mowlaee, R Saeidi, MG Christensen, ZH Tan, T Kinnunen, P Franti, SH Jensen, A joint approach for single-channel speaker identification and speech separation. Audio, Speech, and Language Processing, IEEE Transactions on. **20**(9), 2586–2601 (2012)
21. R Saeidi, P Mowlaee, T Kinnunen, ZH Tan, MG Christensen, SH Jensen, P Franti, in *Pattern Recognition (ICPR), 2010 20th International Conference on IEEE,(IEEE*. Signal-to-signal ratio independent speaker identification for co-channel speech signals (New York, 2010), pp. 4565–4568
22. A Nádas, D Nahamoo, MA Picheny, Speech recognition using noise-adaptive prototypes. IEEE Trans. Acoust., Speech, Signal Process. **37**, 1495–1503 (1989)
23. P Mowlaee, R Martin, in *Proceedings of IWAENC 2012; International Workshop on VDE*. On phase importance in parameter estimation for single-channel source separation, in Acoustic Signal Enhancement (IEEE New York, 2012), pp. 1–4
24. AP Varga, RK Moore, *Hidden Markov model decomposition of speech and noise* (IEEE, New York, 1990), pp. 845–848
25. Y Shao, DL Wang, Model-based sequential organization in cochannel speech. IEEE Trans. Audio, Speech, Lang. Process. **14**, 289–298 (2006)
26. A Narayanan, DL Wang, A CASA based system for long-term, SNR estimation. IEEE Trans. Audio Speech Lang. Process. **20**, 2518–2527 (2012)
27. M Cooke, T Lee, Speech, Separation Challenge (21 September 2006). [http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm]

28. G Kim, Y Lu, Y Hu, PC Loizou, An, algorithm that improves speech intelligibility in noise for normal-hearing listeners. **126**(3), 1486–1494 (2009)
29. CH Taal, RC Hendriks, R Heusdens, J Jensen, A short-time objective intelligibility measure for time-frequency weighted noisy speech, in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on IEEE, 4214–4217 (2010)
30. S Rennie, J Hershey, P Olsen, Single channel multi-talker speech recognition: graphical modeling approaches. IEEE Signal Process. Mag. **27**(6), 66–80 (2010)

doi:10.1186/1687-4722-2013-14
Cite this article as: Hu and Wang: **An iterative model-based approach to cochannel speech separation.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:14.