Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2008) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Monaural speech separation based on MAXVQ and CASA for robust speech recognition

Peng Li [a,*], Yong Guan [b], Shijin Wang [a], Bo Xu [a,b], Wenju Liu [b]

[a] *Digital Content Technology Research Centre, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[b] *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

Received 6 July 2007; received in revised form 25 March 2008; accepted 29 May 2008

Abstract

Robustness is one of the most important topics for automatic speech recognition (ASR) in practical applications. Monaural speech separation based on computational auditory scene analysis (CASA) offers a solution to this problem. In this paper, a novel system is presented to separate the monaural speech of two talkers. Gaussian mixture models (GMMs) and vector quantizers (VQs) are used to learn the grouping cues on isolated clean data for each speaker. Given an utterance, speaker identification is firstly performed to identify the two speakers presented in the utterance, then the factorial-max vector quantization model (MAXVQ) is used to infer the mask signals and finally the utterance of the target speaker is resynthesized in the CASA framework. Recognition results on the 2006 speech separation challenge corpus prove that this proposed system can improve the robustness of ASR significantly.

Keywords: Monaural speech separation; Computational auditory scene analysis (CASA); Factorial-max vector quantization (MAXVQ); Automatic speech recognition (ASR)

1. Introduction

Robustness is one of the main problems for automatic speech recognition (ASR) applications. Speech recognition performance degrades greatly when the systems are used in noisy environments or when there exist mismatches between the training and the testing conditions. Mismatches between training and testing may result from many factors, especially the nonstationary background noise. Hence, noise robust ASR is one of the main topics for many researchers in this field. (Acero, 1992; Gong, 1995; Junqua and Haton, 1996).

The approaches to improve speech recognition performance in noisy environments can be classified into four major categories, e.g. signal enhancement, feature compensation, model adaptation, and noise contamination training (Gong, 1995; Furui, 1997; Lee, 1998).

* Corresponding author. Tel.: +86 10 82872970 211; fax: +86 10 62570224.
E-mail addresses: pengli@hitic.ia.ac.cn (P. Li), yguan@nlpr.ia.ac.cn (Y. Guan), sjwang@hitic.ia.ac.cn (S. Wang), xubo@hitic.ia.ac.cn (B. Xu), lwj@nlpr.ia.ac.cn (W. Liu).

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.05.005

Signal enhancement is used to increase the quality of the signals sent to the recognizers. The representative approaches include spectral subtraction (Boll, 1979), Wiener filter (Mcaylay and Malpass, 1980), and advanced front-end of distributed speech recognition released by ETSI, 2002 etc. However, the performance improvement is not obvious, because there is no direct relationship between the SNRs (or perceptual qualities) of the enhanced signals and the recognition performance, even when the acoustic models are trained via the same processing.

Feature compensation aims at constructing a compact and robust feature set to compensate for the mismatches of the acoustic spaces between training and testing. It can be operated either in the feature space or in the model space. The main approaches include signal bias removal (Rahim and Juang, 1996; Lawrence and Rahim, 1999), stochastic matching (Sankar and Lee, 1996; Lee, 1998), noise modeling and masking (Nadas et al., 1989; Sanches, 2000; Zhao, 2000), and parallel model combination (Gales, 1995; Gales and Yound, 1996).

Model adaptation techniques attempt to adapt system parameters to the speakers or the testing environments by using representative testing data samples. The popular adaptation schemes are maximum a posteriori (MAP; Gauvain and Lee, 1994) or maximum likelihood linear regression (MLLR; Leggetter and Woodland, 1995; Gales and Woodland, 1996; Gales, 1998).

Besides the three categories of techniques described above, in the literature there exists an approach known as training data contamination (Das et al., 1993; Daytrich et al., 1983). It provides a way to train acoustic models more robustly and more representatively by using the real noisy data for training instead of using clean speeches. In practical implementations, training data are produced by injecting noise into the clean speech utterances. Although obviously this approach is time-consuming and may be impractical when the training data are huge, it does offer some advantages. For instance, it is free from the negative spectrum problem in noise subtraction (Furui, 1992).

Although the above approaches do improve the robustness of ASR systems more or less, these methods still have not solved the robustness problem completely, especially in nonstationary noisy conditions. As a result, researchers try to borrow some ideas from other related fields to deal with noise robustness. Since human auditory system shows a remarkable ability on catching target source in a wide range of listening conditions, more and more researchers start to introduce the achievements in psychophysics and psychoacoustics into speech processing to let computers imitate the same capability.

Great achievements have been made on the speech perception in the past several decades. In 1990, Bregman first proposed the concept of auditory scene analysis (ASA; Bregman, 1990). According to Bregman, ASA takes place in the brain in two stages: The first stage decomposes an auditory scene into segments (or sensory elements) and the second stage groups segments into streams. Generally speaking, grouping processes may be primitive, or schema-based. The former is based on general heuristics and performed in a bottom-up manner, while the latter is based on specific knowledge and performed in a top-down manner.

With the development of the researches on speech perception, many auditory cues have been used for perceptual organization of the decomposed signals (Darwin and Carlyon, 1995; Cooke and Ellis, 2001; Darwin and Hukin, 2000; Brown and Wang, 2005). These cues can be classified into two types. One is the primitive cues, including event boundaries such as onset and offset, temporal modulations such as common FM, periodicity such as harmonicity and fine-structure periodicity, spatial location such as interaural time difference (ITD) and interaural intensity difference (IID), and event sequence such as good continuation and similarity. The other is the source-specific cues, including speaker-specific cues such as vocal tract size, shape, and accent, and linguistic cues such as phonotactics, intonation and so on. Different cues have different strengths, and they compete to control the grouping. For example, the perceptual segregation of sounds in a sequence depends upon differences in their frequencies, pitches, timbres (spectral envelopes), central frequencies (of noise bands), amplitudes, and locations, and upon sudden changes of these variables. The perceptual fusion of simultaneous components depends on their onset and offset synchrony, frequency separation, regularity of spectral spacing, binaural frequency matches, harmonic relations, parallel amplitude modulation, and parallel gliding of components.

The study on ASA greatly promoted the development of computational auditory scene analysis (CASA). According to the known principles of ASA, many CASA systems have been proposed (Weintraub, 1985; Cooke, 1991; Brown and Cooke, 1994; Wang and Brown, 1999; Ellis, 1999; Hu and Wang, 2004; Li et al.,

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

2006). These systems offer a new way to perform speech separation and enhance the qualities of the input data for ASR. A typical CASA system consists of two main stages: segmentation and grouping. In the segmentation stage, the acoustic input is decomposed into sensory segments, each of which should be originated from a single source. In the grouping stage, those segments probably originated from the same source are grouped together. The CASA-based separation systems are not only able to realize the objective of speech segregation, but also need no strong assumption on the acoustic properties of interferences. However, most of the proposed CASA systems nowadays are mainly based on pitch (Wang and Brown, 1999; Hu and Wang, 2004; Li et al., 2006). As a result, they could only be used to segregate voiced speech. Therefore, there exists the great challenge to separate unvoiced segments in monaural speech. To solve this problem, some researchers attempted to combine other valuable grouping cues, such as onset, offset, and timbre (Hu and Wang, 2007; Godsmark and Brown, 1999). Other researchers considered the acoustic and the phonetic characteristics of individual unvoiced consonants and introduced them in CASA (Hu and Wang, 2005).

Meanwhile, as many proposed speech and audio signal processing methods, the technique of machine learning has potential to be applied to the monaural source separation. Although natural speech exhibits a lot of regularities in the way that energies are distributed across the time-frequency plane, the machine learning community does not pay much attention to monaural speech separation. Therefore, it is valuable to explore the statistical learning methods to discover the regularities from a large amount of speeches to perform speech separation in the CASA framework.

Considering the facts described above, in this paper we attempt to separate the voiced and unvoiced target speech simultaneously from non-stationary interfering sources by using factorial-max vector quantization (MAXVQ; Roweis, 2003) and some related techniques of CASA. For convenience, we focus our objective on the separation of speech signals from two talkers, which is also one of the most difficult problems in robust ASR.

The organization of this paper is as follows: Section 2 first gives an overview of our proposed system, and then the main components of the model are explained in detail in Sections 3–5, including model training, speaker identification and speech separation etc. In Section 6, the proposed method is systematically evaluated with a standard corpus. Finally, a further discussion is given in Section 7.

2. System overview

The main objective of the proposed system is to separate the speech of two talkers by combining CASA with statistical learning technique and provide a robust front-end to ASR. Fig. 1 shows the detailed structure of the proposed system. Obviously the system can be divided into two stages, e.g. training and testing.

In the training stage, lots of prepared isolated clean speech signals of different speakers are used. The speech signals first pass through a decomposition and feature extraction module. Different from the general feature extraction method, in this system an auditory filterbank is used to analyze the input signal in consecutive time frames. Via this processing the input signal is decomposed into a two-dimensional time-frequency map. Each unit of this map is termed as a T-F unit corresponding to a certain filter at a certain time frame. Then the

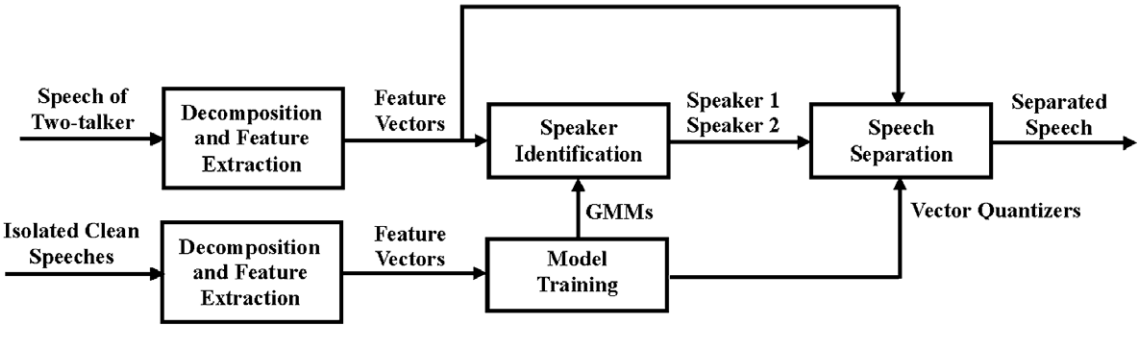

Fig. 1. Schematic diagram of the proposed system.

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

features (the natural logarithmic energies of all the units) are extracted and sent to the model training module. For each speaker a Gaussian mixture models (GMM) and a vector quantizer (VQ) are trained using a modified K-means clustering algorithm to find K good local optima. The pre-trained GMMs and VQs will be utilized in the testing stage to guide speaker identification and speech segregation.

The testing stage is composed of three steps. When a degraded speech (not included in the training data) is sent to the system, it is first decomposed into T-F units and the feature vectors are extracted in the same way as in the training stage. Then, speaker identification is performed on the extracted feature vectors. The speaker identification is implemented based on the IBM 2006 speech separation challenge system (Kristjansson et al., 2006) with a few modifications. After this processing, a MAXVQ model consisting of two VQs corresponding to the two recognized speakers is introduced to infer the binary mask signals for the resynthesis in the CASA framework. Finally the separated utterance is obtained, which more or less comes from the same speaker and as a result its quality is better than the original degraded speech.

Detailed explanations on the main components of the proposed system will be presented in the following sections.

3. Model training

The aim of model training is to obtain the *a priori* knowledge for speaker identification and speech separation. For each speaker, two models are trained. One is the GMM with K_1 mixtures used for speaker identification; the other is the VQ with K_2 codewords used for speech segregation. Both of the two models can be trained using a modified K-means clustering algorithm, which is performed by finding the K local optima from the training samples of the corresponding speaker. To reduce the redundant training and get the two models simultaneously, we set $K_1 = K_2 = K$. That is, the mixture number of the GMM is same as the codeword number of the VQ.

More specifically, given some isolated clean recordings, all the feature vectors from one speaker are clustered into K classes. For each class, the mean vector, the covariance matrix and the prior probability of the corresponding class are estimated. Then, the estimated mean vectors, covariance matrices and prior probabilities of the K clusters form the K mixtures of the GMM and the K codewords of the VQ for the speaker. These pre-trained GMMs and VQs will be used in the testing stage to guide speaker identification and speech separation.

4. Speaker identification

In order to identify the speakers presented in the mixed speech, a model-based speaker identification method (Kristjansson et al., 2006) is utilized. Given feature vectors extracted from a mixture, the method could identify which frames are uttered by a single speaker, which are not and in the single/mixed frames who are present.

More specifically, assume that there are M_{T} speakers included in the training data, each speaker is described by one GMM:

$$p(\mathbf{x}_t|m) = \sum_k \pi_m^k p\left(\mathbf{x}_t|\mathbf{v}_m^k, \sum_m^k\right), \quad m \in \{1,2,\ldots,M_{\mathrm{T}}\}, \quad k \in \{1,2,\ldots,K\} \tag{1}$$

where m is the index of the speaker, K is the number of mixture components, and $\mathbf{x}_t$ is the extracted feature vector at frame t. π_m^k, $\mathbf{v}_m^k$ and $\sum_m^k$ denote the *a priori* probability, mean vector and covariance matrix of the kth mixture component for speaker m, respectively.

Since the testing features may not match the training features in some sense, a gain information is introduced the same as Kristjansson et al., 2006. Thus Eq. (1) can be modified as:

$$p(\mathbf{x}_t|m) = \sum_g \sum_k \pi_g \pi_m^k p(\mathbf{x}_t|\mathbf{v}_m^k + g, \sum_m^k), g \in \{g_1, g_2, \ldots, g_G\} \tag{2}$$

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

where the gain parameter g is modeled as a discrete variable corresponding to G kinds of different target-to-masker ratio (TMR) with the *a priori* π_g as uniform.

Assume each dimension of the feature vector is independent from other dimensions and $\sum_m^k$ can be constrained as a diagonal matrix (Note that although this assumption is physically unrealistic, it is extremely valuable to simplify the computation and in practice this approximation could provide acceptable results). Sequentially, Eq. (2) could be rewritten as:

$$p(\mathbf{x}_t|m) = \sum_g \sum_k \pi_g \pi_m^k \frac{1}{(2\pi)^{D/2} \prod_d \sigma_{md}^k} \exp\left(-\frac{1}{2} \sum_d \left(\frac{x_{td} - v_{md}^k - g}{\sigma_{md}^k}\right)^2\right) \tag{3}$$

where D is the dimension number of feature vectors, x_{td} is the dth dimension of $\mathbf{x}_t$, v_{md}^k is the dth dimension of $\mathbf{v}_m^k$, and σ_{md}^k is the dth diagonal element of $\sum_m^k$.

It is found that if one dimension of a feature vector departs greatly from v_{md}^k, the probability $p(\mathbf{x}_t|m)$ will be mainly dominated by this dimension and as a result it will be much less than the true one. To reduce this effect, we modify Eq. (3) as:

$$p(\mathbf{x}_t|m) = \sum_g \sum_k \pi_g \pi_m^k \frac{1}{(2\pi)^{D/2} \prod_d \sigma_{md}^k} \exp\left(-\frac{1}{2} \sum_d \phi(x_{td}, v_{md}^k + g, (\sigma_{md}^k)^2)\right) \tag{4}$$

where $\phi(x_{td}, v_{md}^k + g, (\sigma_{md}^k)^2)$ is defined as:

$$\phi(x_{td}, v_{md}^k + g, (\sigma_{md}^k)^2) = \begin{cases} (x_{td} - v_{md}^k - g)^2/(\sigma_{md}^k)^2, & if\,(x_{td} - v_{md}^k - g)^2/(\sigma_{md}^k)^2 < \lambda \\ \lambda, & \text{otherwise} \end{cases}. \tag{5}$$

where the threshold λ is determined by the experiments on the development set.

To gain the reliable estimation of $p(m|\mathbf{x})$, the *a posteriori* probability of $\mathbf{x}$ belonging to speaker m, only the frames with high entropy and fitting the model very well are selected to estimate which speakers are present in the mixture. The practical estimation is performed as below:

(1) Given $\mathbf{x}_t$, compute the normalized likelihood of speaker m for each frame

$$b_{\mathbf{x}_t}(m) = p(\mathbf{x}_t|m) \Big/ \sum_{m'} p(\mathbf{x}_t|m') \tag{6}$$

(2) Approximate the component class likelihood

$$p(\mathbf{x}|m) = \sum_t \psi(b_{\mathbf{x}_t}(m)) \cdot b_{\mathbf{x}_t}(m) \tag{7}$$

where $\psi(b_{\mathbf{x}_t}(m))$ is a confidence weight related to $b_{\mathbf{x}_t}(m)$ and defined as

$$\psi(b_{\mathbf{x}_t}(m)) = \begin{cases} 1, & \max b_{\mathbf{x}_t}(m) > \gamma \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where γ is a threshold determined by the experiments on the development set.

(3) Compute the component class posteriori via:

$$p(m|\mathbf{x}) \propto p(\mathbf{x}|m)p(m) \tag{9}$$

Therefore, the frames, which have higher $b_{\mathbf{x}_t}(m)$ but can not fit any of the speaker models, are discarded; and only the frames, which fit the model very well, are utilized to estimate which speakers are present in the mixture speech. Given a short-list of finalists chosen according to $p(m|\mathbf{x})$ as computed above, the present components are identified by applying a max-based approximate EM algorithm to identify the most probable speaker combination.

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

5. Speech separation

In this section, the method of speech separation, the key component of the proposed system, is presented. The separation is performed in the CASA framework, and the details, including the binary mask applied in CASA, the MAXVQ model, the inference of the mask signals and the processing of resynthesis will be described sequentially.

5.1. Binary mask in CASA

The aim of CASA is to organize the energies in a complex sound mixture that corresponds to different sources. Since the two-dimensional time-frequency representation, where the time dimension consists of a sequence of time frames and the frequency dimension consists of a bank of auditory filters (e.g., gammatone filters), is widely used in the CASA literature, speech separation based on CASA could be further treated as a process with the objective to classify the time-frequency regions (T-F units) into two streams, corresponding to the target and intrusion, respectively. Based on this, binary masks have been widely used in the CASA literature as an output representation to label the origins of the mixed speech (Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004; Li et al., 2006).

The idea of binary mask is originated from the auditory masking phenomenon: within a critical band a weaker signal is masked by a stronger one (Moore, 1997). It is effective for speech separation based on CASA and is well-defined no matter how many intrusions exist and how many targets need to be segregated (Wang, 2005). Within this representation, the key consideration behind the notion of the binary mask is to retain the T-F units of a target sound that are stronger than the interference and discard the units that are weaker than the interference. More specifically, 1 is used to indicate that the target energy is stronger than the interference energy within the corresponding T-F unit and 0 is used otherwise.

When a gammatone filterbank is used to generate the time-frequency representation, a technique proposed by Weintraub, 1985 can be introduced to resynthesize a waveform signal using the binary mask (see also Brown and Cooke, 1994; Wang and Brown, 1999). This provides a convenient way to combine the CASA system with other techniques.

According to the above description, the problem remaining is how to compute the mask signals automatically. Fortunately, natural speech exhibits a lot of regularity in the way that energy is distributed across the time-frequency plane. Grouping cues based on these regularities have been studied by psycho-physicists and are utilized in many CASA systems (Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004). Different from these systems, in this paper we propose a statistical approach to discover these regularities from a large amount of speech data and then use the MAXVQ model to compute the binary mask signals to perform separation.

5.2. MAXVQ

MAXVQ is a useful approach to model complicated sensory observations using a number of separate but interacting causes. It is motivated by the observation that the log energy in a narrow band of a mixture is almost exactly the maximum of the individual logarithmic energy in that band. The MAXVQ model usually has a fixed number of VQs, each of which stochastically selects a prototype to model the observation vector. The final output vector is a noisy composite of the set of proposed prototypes, obtained by taking the element-wise maximum of the set (Roweis, 2000; Roweis, 2003).

MAXVQ is a latent variable probabilistic model for D-dimensional data vectors $\mathbf{x}$. It consists of N_T vector quantizers, each of which has K codewords of mean vectors $\mathbf{v}_n^l$ and covariance matrices $\sum_n^l$ (where, $n \in \{1,2,\ldots,N_T\}, l \in \{1,\ldots,K\}$). Latent variables $z_n \in \{1,\ldots,K\}$ control which codeword vector each VQ selects.

Assume that each VQ chooses its codebook entries independently with fixed rates π_n^l, that is,

$$p(z_n = l|\pi) = \pi_n^l, \tag{10}$$

The probability of selecting a setting of the latent $\mathbf{z}$ can be computed as:

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

$$p(\mathbf{z}) = \prod_n p(z_n), \quad \mathbf{z} = (z_1, \ldots, z_{N_T}) \tag{11}$$

Given the N_T selections, the final output is generated as a noisy version of the elementwise maximum of the selected codewords. More specifically, we first determine the index a_d of the quantizer selected in the dth dimension of the output as:

$$a_d = \arg\max_n (v_{nd}^{z_n}), \quad d \in \{1, \ldots, D\} \tag{12}$$

Then, all the selected indexes in the D dimensions form the vector $\mathbf{a} = (a_1, a_2, \ldots, a_D)^T$. Therefore, the output's mean vector $\mathbf{v_a}$ and covariance matrix $\sum_{\mathbf{a}}$ can be written as:

$$\mathbf{v_a} = (v_{a_1 1}^{z_{a_1}}, v_{a_2 2}^{z_{a_2}}, \ldots, v_{a_D D}^{z_{a_D}})^T \tag{13}$$

$$\sum\nolimits_{\mathbf{a}} = diag\left(\sum_{a_1 1}^{z_{a_1}}, \sum_{a_2 2}^{z_{a_2}}, \ldots, \sum_{a_D D}^{z_{a_D}}\right) \tag{14}$$

Since every dimension of the feature is assumed to be independent, given a selected setting of $\mathbf{z}$, we can compute the observation probability of x_d as:

$$p\left(x_d | a_d, \mathbf{v_a}, \sum_{\mathbf{a}}\right) = N\left(x_d | v_{a_d d}^{z_{ad}}, \sum_{a_d d}^{z_{ad}}\right) \tag{15}$$

where, $N(*)$ means the Gaussian distribution. As a result, the observation probability of the feature vector $\mathbf{x}$ at the condition of $\mathbf{v_a}$,$\sum_{\mathbf{a}}$ and π can be computed as:

$$p\left(\mathbf{x} | \mathbf{v_a}, \sum_{\mathbf{a}}, \pi\right) = \sum_{\mathbf{z}} p(\mathbf{z}|\pi) p\left(\mathbf{x} | \mathbf{z}, \mathbf{v_a}, \sum_{\mathbf{a}}\right) \tag{16}$$

Through comparing the observation probabilities of $\mathbf{x}$ in the conditions of all the possible settings of $\mathbf{z}$,we can find the most likely setting of $\mathbf{z}$ by selecting the one which makes the observation probability of the feature vector $\mathbf{x}$ have the maximum value.

5.3. Inference of binary mask signals

It has been emphasized that the objective of introducing MAXVQ is to infer the mask signals needed in the resynthesis of CASA. For two-talker speech, there are two "sources" (speakers). Each speaker can be treated as an independent cause contributing to the observed signal. Therefore, a MAXVQ can be obtained with the two VQs which are determined according to the results of speaker identification.

Given an observation vector $\mathbf{x}$, to fulfill the inference process, we have to find the single most likely setting of $\mathbf{z}$ in all the possible joint settings of the hidden selection variable $\mathbf{z}$. Once the MAP setting of $\mathbf{z}$ is computed, the mask signals can be estimated. Considering that the short-time logarithmic energy in cochleagram's frequency bands are used as feature vectors in analysis, the inference could be performed as follows: for each frame of the input cochleagram, if the output of a band (channel) proposed by the model corresponding to the target speaker is the maximum over the two models (corresponding to the target speaker and the intrusion speaker, respectively), then the mask signal of this band is set to 1. For other bands, the masks are 0.

5.4. Resynthesis

Having inferred the mask signals, a speech waveform could be resynthesized from the foreground stream using the method proposed by Weintraub, 1985. In practical implementations, it should be noted that:

(1) Since the speaker identification module does not offer which of the two recognized speakers is the target source, resynthesis should be implemented twice to produce waveforms corresponding to both of the sources by switching the mask signals from 1 to 0 and 0 to 1.

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

(2) Considering that the objective of the proposed system is to offer a good front-end for ASR, if we retain the energies of the units masked with 1 and reject all the energies of the units masked with 0 as usual, it may lead to information loss. Therefore, a slight modification is made and real value masks are employed to replace the binary masks. More specifically, the mask of a T-F unit is not binary any longer but a real value, which is proportional to the amplitude ratio of the target to the mixture in that unit. Since the feature vector is composed of the natural logarithmic energies of the T-F units and the weighting operation is operated on the outputs of gammatone filters, the real value mask can be derived by:

$$\theta_d^t = \sqrt{e^{v_{1d}^t}} \Big/ \left(\sqrt{e^{v_{1d}^t}} + \sqrt{e^{v_{2d}^t}} \right) \tag{17}$$

where θ_d^t is the mask value of the T-F unit of channel d in frame t; and v_{1d}^t, v_{2d}^t are the dth dimension of the mean vectors $\mathbf{v}_1^t$ and $\mathbf{v}_2^t$ of the two optimal codewords from the two VQs of MAXVQ in frame t, corresponding to the target speaker and interference speaker, respectively.

After the above processing, the resynthesized speech will more or less come from the same source and its quality should be better than the original mixture. More details about the resynthesis can be referred to (Weintraub, 1985; Brown and Cooke, 1994; Wang and Brown, 1999).

6. Evaluation

In this section, the proposed system will be evaluated on a standard corpus with an HMM based recognizer.

6.1. Evaluation corpus

The evaluation corpus is the 2006 speech separation challenge corpus (Cooke and Lee, 2008). It was used as a standard corpus to test the performance of some separation systems in ICSLP'2006 (Kristjansson et al., 2006; Barker et al., 2006; Srinivasan et al., 2006; Ming et al., 2006; etc.).

The corpus contains three sets: training, testing, and development. The training set consists of 17,000 sentences (500 from each of the 34 talkers). The testing and the development sets are both composed of 2 types of data: sentences in speech-shaped noise and sentences of two-talker. The two-talker data contain pairs of sentences mixed at 6 different target-to-masker ratio (TMR): −9, −6, −3, 0, 3, and 6 dB. Each TMR condition contains 600 sentences in testing set and 300 in development set. In the 600 testing sentences about one third consists of same talker mixtures (ST); another one third mixtures of different talkers with the same gender (SG); and the remaining one third different gender mixtures (DG).

Since isolated utterances are needed to learn the models (GMMs and VQs) and the corpus does not have isolated noise utterances in the speech-shaped noise set, only the two-talker data is used to evaluate the performance of the proposed system.

6.2. Practical implementations

In practical implementations, we decompose the input signals with a 128-channel "gammatone" filterbank (Patterson et al., 1988) whose center frequencies are quasi-logarithmically spaced from 80 to 10,000 Hz. The gammatone filterbank is a standard model of cochlear filtering (de Boer and de Jongh, 1978). The impulse response of the gammatone filter is:

$$g(t) = \begin{cases} t^{l-1} \exp(-2\pi bt) \cos(2\pi ft), & t \geqslant 0 \\ 0, & \text{else} \end{cases} \tag{18}$$

where $l = 4$ is the order of the filter, b is the equivalent rectangular bandwidth, and f is the center frequency of the filter. In each filter channel, the output of decomposition is then divided into 20-ms time frames with 10-ms overlapping between consecutive frames. As a result of band pass filtering and short-time windowing, the input is decomposed into a two-dimensional T-F representation, or a collection of T-F units. Then we obtain the

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

feature vectors by a natural logarithm operation on the energy of each unit. That is, given a certain frame index, the log-energies of the 128 channels compose the feature vector.

In the training stage, since the number of the speakers presented in the corpus are 34, we set $M_T = 34$ and train 34 GMMs and 34 VQs on the training data by using a modified K-means algorithm. Each GMM consists of $K = 256$ mixtures and each VQ has a codebook with $K = 256$ codewords. Every model corresponds to a speaker presented in the corpus.

In the testing stage, the 34 GMMs pre-trained in the training stage are used together to identify the two speakers presented in a test mixture at first. 6 discrete gain values of g, corresponding to the 6 TMR conditions of −9, −6, −3, 0, 3, and 6 dB, are selected. According to the experimental results on the development set, the optimal values of the thresholds λ and γ are set to 7.5 and 0.9, respectively. Although the values of g are selected according to the known TMRs here, it does not mean that it is necessary to know the TMRs of the mixtures beforehand. In fact, we also attempt to use 13 discrete values of g, corresponding to −10.5, −9, −7.5, −6, −4.5, −3, −1.5, 0, 1.5, 3, 4.5, 6, and 7.5 dB, which provides almost the same speaker identification results. Besides, the thresholds λ and γ are stable. After the above processing, a MAXVQ model consisting of 2 VQs corresponding to the two speakers presented in the mixture is used to infer the mask signals for the test utterance. Finally, resynthesis is performed to recover the waveform of the separated speech as described in 5.4.

To evaluate the recognition performance of the proposed system on ASR, the separated signals are fed into an HMM based recognizer. The recognizer is based on the HTK package version 3.1 (Cooke and Lee, 2008). The input speech waveforms are parameterized into standard 39-dimensional Mel frequency cepstral coefficients (MFCCs), i.e., 12 Mel-cepstral coefficients and the logarithmic frame energy plus the corresponding delta and acceleration coefficients (MFCC_E_D_A). The words are modeled as whole-word HMMs with a left-to-right model topology with no skips over states and 32 Gaussian mixtures per state with diagonal covariance matrices. The number of states for each word is based on 2 states per phoneme. Finally, the recognition results are scored automatically as follows: for every TMR condition, each utterance receives a score of 0, 1 or 2 (based on the letter and digit keywords). This is turned into an average and the results are further scored based on whether the talker and masker are the same, or have the same gender, or have different gender.

6.3. Evaluation results

6.3.1. Results of speaker identification

The proposed speaker identification algorithm described in Section 4 is firstly evaluated. Table 1 shows the results on the two-talker testing set. The percentages of files where both speakers are identified as one of the two most probable source classes are reported. It is shown that on average over all the conditions the accuracy of the speaker identification is 99%. Compared with the results of Kristjansson's (Kristjansson et al., 2006) listed in Table 2, it can be found that they are in the same level. Obviously the modified speaker identification algorithm could solve the problem satisfactorily.

Table 1
Speaker identification results of the proposed system (Percentage of utterances with both speakers in the 2-best list output by the described algorithm. ST, same talker; SG, same gender and DG, different gender)

TMR	ST	SG	DG	Ave
6 dB	100	98	98	99
3 dB	100	99	99	99
0 dB	100	99	99	99
−3 dB	100	99	99	99
−6 dB	100	99	98	99
−9 dB	100	94	97	97
Ave	100	98	98	99

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

Table 2
Speaker identification results of Kristjansson's (Percentage of utterances with both speakers in the 2-best list output by the described algorithm)

TMR	ST	SG	DG	Ave
6 dB	100	97	99	99
3 dB	100	98	99	99
0 dB	100	98	98	99
−3 dB	100	97	98	98
−6 dB	100	97	97	98
−9 dB	99	96	96	97
Ave	100	97	98	98

6.3.2. Results of speech recognition

Tables 3 and 4 list the recognition results in different TMR conditions of the baseline system and the proposed system on the 2006 speech separation challenge corpus, respectively. Separated speech is synthesized with the real weights mentioned in 5.4. The two tables include the results of all the three subsets: ST, SG and DG. The average recognition rates over all subsets for one TMR condition are shown in the last column of the tables, while the average recognition rates over all TMR conditions on one subset are listed in the last row of the tables. Note that the recognition results of ST reported here are obtained by selecting the better result of the two separated speeches resynthesized with different mask signals from a mixture utterance, as mentioned in Section 5.4.

As shown in Tables 3 and 4, the absolute improvements over the 6 TMR conditions are 1.17%, 16.58%, 22.41%, 23.25%, 18.08% and 12.58%, respectively. The proposed system improves the performances significantly for almost all TMRs except 6 dB and provides an average improvement of 15.68% over the original results.

Table 5 lists the relative performance improvements of the proposed system compared with the baseline system. The average relative improvement is 94.71%. The lower the TMR condition is, the larger the relative improvement is obtained. All the results show that the proposed system is effective to deal with the robustness problem of ASR, especially for the speeches of two talkers.

Table 3
Baseline system's recognition results

TMR	ST (%)	SG (%)	DG (%)	Ave (%)
6 dB	62.44	64.25	64.25	63.58
3 dB	46.15	44.13	46.75	45.75
0 dB	29.64	32.96	33.50	31.92
−3 dB	18.10	20.95	19.50	19.42
−6 dB	9.73	14.53	11.50	11.75
−9 dB	5.66	7.26	7.50	6.75
Ave	28.62	30.68	30.50	29.86

Table 4
Proposed system's recognition results

TMR	ST (%)	SG (%)	DG (%)	Ave (%)
6 dB	42.08	70.67	84.50	64.75
3 dB	37.10	70.39	83.00	62.33
0 dB	27.60	63.69	75.50	54.33
−3 dB	19.68	46.93	64.25	42.67
−6 dB	14.48	32.40	44.50	29.83
−9 dB	13.80	17.60	27.00	19.33
Ave	25.79	50.28	63.13	45.54

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

Table 5
Relative performance improvements between the proposed system and the baseline system

TMR	6 dB	3 dB	0 dB	−3 dB	−6 dB	−9 dB	Ave
Imp.	1.84%	36.24%	70.21%	119.72%	153.87%	186.37%	94.71%

However, the performances of the proposed method on the subsets of ST, SG and DG are not consistent: the performances of DG and SG are good while the performance of ST is not. The reason is that in the ST condition the two VQs used to infer the mask signals are the same. Thus, it is difficult to determine which of the two selected codewords corresponds to the target speech. As a result, the proposed system could not give accurate mask signal sequence to resynthesize the separated speech. Obviously we should spend more efforts to solve this problem in the future.

6.3.3. Comparison with other systems

The evaluation of the proposed model has been elaborated above. It is clear that the recognition performance of the separated speeches is distinctly improved. In this section, we will compare our system with some other systems.

Firstly, as we know, the errors in the speech identification may bring about segregation errors. To further evaluate the effect of speaker identification, we employ the true identities of the present speakers for speech segregation. True identities are obtained from the file name of the mixture. The results of the system using true speaker identities instead of the speaker identification module are given in Table 6. Compared with Table 4, the results of SG and DG are slightly better (about 0.14% and 0.16%, respectively), however the results of ST are worse (about 0.87%). The main reason that the results of true speaker identities in ST subset are not better as expected is that the proposed speaker identification method may identify two different talkers (one is the target talker and the other is someone else who is similar to the target talker) instead of one talker in ST subset. As a result, the confusion in organizing the T-F units would be decreased and better sequences of mask signals might be obtained. The above results also imply that the speaker identification used in our system is quite reliable.

Secondly, we compare the performance of our proposed system with a typical system designed to deal with the 2006 speech separation challenge. Although there are several systems which could achieve better performance on the 2006 speech separation challenge (Kristjansson et al., 2006; Barker et al., 2006; Srinivasan et al., 2006; Ming et al., 2006; etc.), we choose Srinivasan's system (Srinivasan et al., 2006) for comparison because it employs the concept of binary mask and is also implemented in a CASA framework. Table 7 gives the recognition results of Srinivasan's. It can be found that our system outperforms Srinivasan's in about half of the conditions especially in the low TMRs of ST mixtures and the high TMRs of DG mixtures. Although the average performance of our proposed system is about 0.52% lower than Srinivasan's, the difference between the two systems is not distinct. There are two reasons, which result in the performance difference. The first is that the recognizers of the two systems are not the same. In our experiments, a conventional MFCC feature based recognizer is used, while in Srinivasan's system a missing data based recognizer like Cooke et al., 2001 is employed. Therefore, Srinivasan's system embodies a tighter link of CASA and ASR systems, which would greatly facilitate the performance of speech recognition. The second is that in the proposed system, to

Table 6
Recognition results of the proposed system under the condition that the speakers are known

TMR	ST (%)	SG (%)	DG (%)	Ave (%)
6 dB	40.50	70.95	84.50	64.25
3 dB	35.97	70.39	82.75	61.83
0 dB	26.92	63.69	75.75	54.17
−3 dB	19.00	47.21	64.75	42.67
−6 dB	15.38	32.68	44.75	30.33
−9 dB	11.76	17.60	27.25	18.67
Ave	24.92	50.42	63.29	45.32

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

Table 7
Recognition results of the Srinivasan's system

TMR	ST (%)	SG (%)	DG (%)	Ave (%)
6 dB	57.69	77.37	82.00	71.67
3 dB	40.95	68.72	79.25	62.00
0 dB	27.15	63.97	70.00	52.42
−3 dB	17.65	49.16	58.50	40.67
−6 dB	15.61	33.80	45.00	30.83
−9 dB	11.09	16.76	29.00	18.75
Ave	28.36	51.63	60.63	46.06

reduce the computing time, only simultaneous organization is performed, while in Srinivasan's system, both simultaneous and sequential organization are performed, which may correct some errors in grouping and improve in some sense the final recognition performance, especially the performance of ST. The above reasons also give some suggestions on our research, which would be studied further in future.

Thirdly, since the mixture and the clean target signal before mixing are available, the *a priori* masks can be easily obtained. Considering that the separation method of our proposed system is mainly based on the concept of real value mask, we construct a separation system which employs the *a priori* masks to resynthesize the separated speeches and compare the recognition performance with that provided by the proposed system. Note that in the resynthesis stage, the *a priori* mask system also adopts the real value masks similar to those mentioned in 5.4. The difference is that the real value mask estimation is performed as:

$$\theta_d^t = \sqrt{E_{Td}^t} / \left(\sqrt{E_{Td}^t} + \sqrt{E_{Nd}^t} \right) \tag{19}$$

where E_{Td}^t and E_{Nd}^t are the energies of the clean speech and the interference speech in channel d and frame t. Table 8 lists the results of the *a priori* mask system. Since it utilizes the *a priori* information about the isolated target speech and interfering speech, the results in Table 8 are uniformly better than all the systems mentioned before. Compared with our proposed system and Srinivasan's, the average performance improvement for the entire corpus is about 47.84% and 47.32% (although the results of the *a priori* mask system could not be compared with the results of Srinivasan's system directly because the two systems used different recognizers, the results still give us a qualitative description of the performance gap between them), which indicates how much the performance of our proposed system could be improved. The reason that there is a substantial gap is because the qualities of the real value masks estimated in our system are not as accurate as those in the *a priori* mask system. In fact, the separation system based on masks could be treated as a highly nonstationary Wiener filter, and the masks have a strong impact on the final performance. Since in our system, the mean vectors of the two selected codewords of MAXVQ, which are trained with clean speeches, are used to estimate the real value masks, it will inevitably result in estimation errors when there exists interference. This is because the practical energies of the observed feature vectors in the mixture do not equal to the mean energies of the two selected codewords corresponding to the target speaker and interference speaker. Therefore, the key point to improve the performance is to further increase the estimation accuracy of the masks.

Finally, we also compare the performance of our system with that of listeners. The results of listeners are listed in Table 9, which are reported in Cooke et al., 2008. From Table 9, we can find that the performance of

Table 8
Recognition results of the *a priori* mask system

TMR	ST (%)	SG (%)	DG (%)	Ave (%)
6 dB	93.67	94.97	95.25	94.58
3 dB	93.21	95.25	95.00	94.42
0 dB	92.76	94.41	95.00	94.00
−3 dB	91.86	94.13	93.75	93.17
−6 dB	90.50	93.58	93.50	92.42
−9 dB	89.82	93.58	92.00	91.67
Ave	91.97	94.32	94.08	93.38

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

Table 9
Recognition results of Listeners

TMR	ST (%)	SG (%)	DG (%)	Ave (%)
6 dB	89.80	93.00	94.10	92.30
3 dB	71.80	85.10	91.00	82.60
0 dB	54.40	76.10	86.40	72.30
−3 dB	51.90	72.40	87.60	70.70
−6 dB	60.20	76.80	86.70	74.60
−9 dB	67.90	79.70	83.00	76.90
Ave	66.00	80.52	88.13	78.23

our proposed system is 32.69% lower than that of listeners. Although the performance gap is still substantial, it could also be found that the results of the *a priori* mask system obviously outperform the listener's results in all the conditions (the average improvement is about 15.15%). This indicates that the separation based on real value masks is promising to solve the robustness problem of speech recognition. If more accurate masks could be acquired, it will be possible to make computers have the comparable or even better ability to recognize two-talker mixtures than human being.

7. Conclusions

In this paper, a monaural speech separation method based on MAXVQ and CASA is proposed. By using an inference step in a factorial model to provide the mask signals for resynthesis, separation of the two-talker mixtures is successfully realized. Systematically evaluation on the 2006 speech separation challenge database shows that the proposed method can not only separate the mixture speech of two talkers simultaneously but also improve the recognition rate significantly.

Although the proposed method is effective to solve the separation problem, there are still some drawbacks, which should be paid more attention to in our future work.

First, in the proposed system we only perform the simultaneous organization and do not consider the sequential organization in the separation module, which leads to the confusion on judging the T-F units corresponding to the target source in the ST condition. Therefore, the inferred mask signals for the speech mixed by the same talker are not accurate, and the improvements on recognition performance are not gained as expected. To solve the problem, a straightforward method is to introduce the sequential organization to the separation process. This may be achieved by using pitch and some other cues to guide the organization as many CASA systems used. Of course, extending the MAXVQ model to the Factorial-Max Hidden Markov Model is another way which could be considered.

Second, as we all know, binary mask is originated from the auditory masking phenomenon and is widely used in the CASA literature. Since binary mask signals offer an appropriate classification of the reliable and the unreliable features (Cooke et al., 2001; Green et al., 2001), it provides an excellent front-end for automatic speech recognizer based on missing data technique. In our system, binary mask signals are only used to resynthesize the separated speech. If we combine the binary mask signals with a missing data recognizer, it would further improve the recognition performance.

Third, in our system to infer the mask signals, we have to select a codeword from 256 codewords belonging to each of the two selected vector quantizers' codebooks and compute 256×256 pairs in every frame to find the optimal pair. This leads to huge processing time (roughly 3 min for a 2-s long mixture on a 2.0-GHz Pentium PC). As a result, it is difficult to let the approach work real-time. In fact, it is still a big challenge to find a fast implementation in the CASA field. If this problem could be solved, it would greatly facilitate the real application of CASA undoubtedly.

Acknowledgements

This work is supported by the National Grand Fundamental Research 973 Program of China under Grant No. 2004CB318105 and the National High-Tech Research and Development Plan of China under Grant Nos.

2006AA010103 and 2006AA01Z194. The authors would like to thank the two anonymous reviewers for their great help in improving the structure of this paper.

References

Acero, A., 1992. Acoustical and Environmental Robustness in Automatic Speech Recognition. Kluwer.

Barker, J., Coy, A., Ma, N., Cooke, M., 2006. Recent advances in speech fragment decoding techniques. In: ICSLP'2006.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustic Speech Signal Processing 27 (2), 113–120.

Bregman, A.S., 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. MIT Press, Cambridge.

Brown, G.J., Cooke, M.P., 1994. Computational auditory scene analysis. Computer Speech and Language 8, 297–336.

Brown, G.J., Wang, D.L., 2005. Separation of speech by computational auditory scene analysis. In: Benesty, J., Makino, S., Chen, J. (Eds.), Speech Enhancement. Springer, New York, pp. 371–402.

Cooke, M.P., 1991. Modeling auditory processing and organization, Ph.D. Thesis, University of Sheffield.

Cooke, M.P., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication 34, 267–285.

Cooke, M.P., Ellis, D.P.W., 2001. The auditory organization of speech and other sources in listeners and computational models. Speech Communication 31, 141–177.

Cooke, M.P., Lee, T.-W., 2008. The 2006 Speech separation challenge. Computer Speech and Language.

Cooke, M.P., Garcia Lecumberri, M.L., Barker, J.P., 2008. The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. Journal of the Acoustical Society of America.

Darwin, C.J., Carlyon, R.P., 1995. Auditory grouping. In: Moore, B.C.J. (Ed.), The handbook of perception and cognition, Hearing. Academic, London, pp. 387–424.

Darwin, C.J., Hukin, R.W., 2000. Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. Journal of the Acoustical Society of America 107 (2), 977–979.

Das, S., Bakis, R., Nadas, A., Nahamoo, D., Picheny, M., 1993. Influence of background noise and microphone on the performance of the ibm tangora speech recognition system. In: Proceedings of the ICASSP'93, pp. 95–98.

Daytrich, B.A., Rabiner, L.R., Martin, T.B., 1983. On the effects of varying filter bank parameters on isolated word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 31 (4), 793–897.

de Boer, E., de Jongh, H.R., 1978. On cochlear encoding: potentialities and limitations of the reverse-correlation techniques. Journal of the Acoustical Society of America 63, 115–135.

Ellis, D.P.W., 1999. Using knowledge to organize sound: the prediction-driven approach to computational auditory scene analysis and its application to speech nonspeech mixtures. Speech Communication 27, 281–298.

ETSI, 2002. ETSI draft standard doc speech processing, transmission and quality aspects; distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm. ETSI ES 202 050 V0.1.0.

Furui, S., 1992. Robust speech recognition under adverse conditions. In: Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions, pp. 31–42.

Furui, S., 1997. Recent advances in robust speech recognition. In: Proceedings of ESCANATO Workshop on Robust Speech Recognition for Unknown Communication Channels, pp. 11–20.

Gales, M.J.F., 1995. Model-based techniques for noise robust speech recognition. Ph.D. Thesis, Cambridge University, Cambridge, England.

Gales, M.J.F., Yound, S.J., 1996. Robust speech recognition using parallel model combination. IEEE Transactions on Speech and Audio Processing 4 (5), 352–359.

Gales, M.J.F., Woodland, P.C., 1996. Mean and variance adaptation within the MLLR framework. Computer Speech and Language 10, 249–264.

Gales, M.J.F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language 12, 75–98.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing 2 (2), 291–298.

Godsmark, D., Brown, G.J., 1999. A blackboard architecture for computational auditory scene analysis. Speech Communication 27 (3-4), 351–366.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Communication 16, 191–261.

Green, P., Barker, J., Cooke, M.P., Josifovski, L., 2001. Handling missing and unreliable information in speech recognition. In: AISTATS'2001.

Hu, G.N., Wang, D.L., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Transactions on Neural Network 15 (5), 1135–1150.

Hu, G.N., Wang, D.L., 2005. Separation of fricatives and affricates. In: ICASSP'2005.

Hu, G.N., Wang, D.L., 2007. Auditory segmentation based on onset and offset analysis. IEEE Transactions on Audio, Speech, and Language Processing 15 (2), 396–405.

Junqua, J.C., Haton, J.P., 1996. Robustness in Automatic Speech Recognition. Kluwer.

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005

Kristjansson, T., Hershey, J., Olsen, P., Rennie, S., Gopinath, R., 2006. Super-human multi-talker speech recognition: the IBM 2006 speech separation challenge system. In: ICSLP'2006.

Lawrence, C., Rahim, M., 1999. Integrated bias removal techniques for robust speech recognition. Computer Speech and Language 13, 283–298.

Lee, C.H., 1998. On stochastic feature and model compensation approaches to robust speech recognition. Speech Communication 25, 29–47.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language 9, 171–185.

Li, P., Guan, Y., Xu, B., Liu, W.J., 2006. Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech. IEEE Transactions on Audio, Speech, and Language Processing 14 (6), 2014–2023.

Mcaylay, R.J., Malpass, M.L., 1980. Speech enhancement using a soft-decision noise suppression filter. IEEE Transactions on Acoustic Speech Signal Processing 28, 137–145.

Ming, J., Hazen, T.J., Glass, J.R., 2006. Combining missing-feature theory, speech enhancement and speaker-dependent/-independent modeling for speech separation. In: ICSLP'2006.

Moore, B.C.J., 1997. An Introduction to the Psychology of Hearing, fourth ed. Academic, San Diego, CA.

Nadas, A., Nahamoo, D., Picheny, M., 1989. Speech recognition using noise-adaptive prototypes. IEEE Transactions on Acoustics, Speech and Signal Processing 37 (10), 1495–1503.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P., 1988. An efficient auditory filterbank based on the gammatone function, Applied Psychological Unit, Cambridge University, Cambridge, UK, APU Report 2341.

Rahim, M., Juang, B.H., 1996. Signal bias removal by maximum lielihood estimation for robust telephone speech recognition. IEEE Transactions on Speech and Audio Processing 4, 19–30.

Roweis, S., 2000. One microphone source separation. In: NIPS' 2000.

Roweis, S., 2003. Factorial models and refiltering for speech separation and denoising. In: Eurospeech' 2003.

Sanches, I., 2000. Noise-compensated hidden Markov models. IEEE Transactions on Speech and Audio Processing 8 (5), 533–540.

Sankar, A., Lee, C.H., 1996. A maximum likelihood approach to stochastic matching for robust speech recognition. IEEE Transactions on Speech and Audio Processing 4, 190–202.

Srinivasan, S., Shao, Y., Jin, Z.Z., Wang, D.L., 2006. A computational auditory scene analysis system for robust speech recognition. In: ICSLP'2006.

Wang, D.L., 2005. On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P. (Ed.), Speech Separation by Humans and Machines. Kluwer Academic, Norwell MA, pp. 181–197.

Wang, D.L., Brown, G.J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Transactions on Neural Networks 10 (3), 684–697.

Weintraub, M., 1985. A theory and computational model of auditory monaural sound separation. Ph.D. Dissertation, Department of Electrical Engineering, Stanford University, Stanford, CA.

Zhao, Y., 2000. Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises. IEEE Transactions on Speech and Audio Processing 8 (3), 255–266.

Please cite this article in press as: Li, P. et al., Monaural speech separation based on MAXVQ and CASA ..., Computer Speech and Language (2008), doi:10.1016/j.csl.2008.05.005