Self-Adaption in Single-Channel Source Separation

Michael Wohlmayr[1,a,b], *Ludwig Mohr*[2], *Franz Pernkopf*[2,b]

[1] Commend International, Salzburg, Austria
[2] Signal Processing and Speech Communication Laboratory
Graz University of Technology, Graz, Austria

m.wohlmayr@commend.com, mohr@tugraz.at, pernkopf@tugraz.at

Abstract

Single-channel source separation (SCSS) usually uses pre-trained source-specific models to separate the sources. These models capture the characteristics of each source and they perform well when matching the test conditions.

In this paper, we extend the applicability of SCSS. We develop an EM-like iterative adaption algorithm which is capable to adapt the pre-trained models to the changed characteristics of the specific situation, such as a different acoustic channel introduced by variation in the room acoustics or changed speaker position. The adaption framework requires signal mixtures only, i.e. specific single source signals are not necessary. We consider speech/noise mixtures and we restrict the adaption to the speech model only. Model adaption is empirically evaluated using mixture utterances from the CHiME 2 challenge. We perform experiments using speaker dependent (SD) and speaker independent (SI) models trained on clean or reverberated single speaker utterances. We successfully adapt SI source models trained on clean utterances and achieve almost the same performance level as SD models trained on reverberated utterances.

Index Terms: single-channel source separation, self-adaptation, MLLR

1. Introduction

The aim of single-channel source separation (SCSS) is to divide a mixture of two signals into its underlying source signals. This is in general an ill-posed problem. Implicit models such as computational auditory scene analysis (CASA) and explicit models known as under-determined blind source separation methods are applied [1]. Implicit models try to imitate the remarkable ability of the human auditory system to recover individual sound components in adverse environments. CASA systems are based on harmonicity as cue for separation. In contrast, explicit models incorporate prior knowledge, i.e. the individual source characteristics are learned during a training phase using source specific data. The two most prominent explicit models are the factorial-max vector quantization (VQ) [2] and the factorial-max hidden Markov model [3] which also integrates time dependencies. Another method for identifying components with temporal structure in a time-frequency representation is non-negative matrix factorization (NMF) [4, 5]. All these explicit models usually require sufficient speaker/source specific data for learning which restricts their applicability to scenarios with matching training/test conditions and known sources. In case of model mismatch adaption techniques can be applied. Some of the most successful approaches for model adaption in the context of speech recognition are the maximum likelihood linear regression (MLLR) framework [6, 7], maximum a posteriori (MAP) estimation [8], and rapid adaption in eigenvoice space [9]. While these approaches assume that adaption data consists of clean speech, methods for adaption of *undistorted* source models from contaminated speech have been also developed, e.g. in [10]. Rose et al. [11] extended this in terms of a more general interaction and background noise model based on Gaussian mixture models (GMMs). In [12], the eigenvoice approach is generalized to adapt individual speaker models given a superposition of two speech signals.

In this paper, we develop an algorithm for model adaption to overcome any mismatch between training and testing conditions in SCSS. The aim is to adapt models to a novel acoustic environment using only mixture data. In particular, we consider speech/noise mixtures and we restrict the adaption to the speech model only. Furthermore, we perform self-adaption, i.e. adaption is performed on the same test mixtures used for separation. We propose an EM-based iterative algorithm using MLLR for adaption of the speaker model from the noise/speech mixtures in spectral domain. The model for the speakers and the noise is based on GMMs. Model adaption is empirically evaluated using mixture utterances from the CHiME 2 challenge using speaker dependent (SD) models, speaker independent (SI) models, and models trained on clean data and reverberated data. We use the perceptual evaluation of speech quality (PESQ) measure [13] and metrics of the blind source separation evaluation (BSS EVAL) toolkit [14] for evaluation. Generally, the GMMs trained on reverberated data perform better than the GMMs from clean data. Self-adaption using clean SI models leads to almost the same PESQ performance as obtained for SD models trained on reverberated utterances. Furthermore, self-adaption is more beneficial for mixtures with larger SNRs. While our model is similar to the model proposed in [11], we differ in the following aspects: (i) We use MLLR to adapt from noise signal mixtures. This has the advantage of requiring only little adaption data. In particular, we tie all Gaussian components and use only one global transformation. In [11], a full re-estimate of speaker model parameters is performed. (ii) We apply the model for SCSS of speech/noise mixtures what can be also viewed as model-based speech enhancement. In [11], the model is applied to speaker identification in noise.

The paper is organized as follows: In Section 2 we intro-

[a] This work was performed while M. Wohlmayr was at Graz University of Technology.

[b] This work was supported by the Austrian Science Fund (FWF) under the project number P25244-N15 and the K-Project ASD. The K-Project ASD is funded in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), the Province of Styria - Government of Styria and the Technology Agency of the City of Vienna (ZIT). The programme COMET is conducted by Austrian Research Promotion Agency (FFG).

duce the notation, SCSS, the interaction model, and the GMMs. The model adaption framework is presented in Section 3. In Section 4 empirical results are reported. Section 5 concludes with a perspective on future work.

2. Single-Channel Source Separation

Model-based SCSS for separating the observed noisy speech determines the unknown speech and noise components. Our approach is based on the mixture-maximization (MIXMAX) interaction model [10], i.e. the observed log-magnitude short-time Fourier transform (STFT) of the signal mixture $\boldsymbol{y}^{(t)}$ is approximated by the element-wise maximum of their respective single-source log-magnitude STFTs of the two sources $\boldsymbol{s}_1^{(t)}$ and $\boldsymbol{s}_2^{(t)}$, i.e. $\boldsymbol{y}^{(t)} \approx \max(\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)})$, where vector $\boldsymbol{s}_k^{(t)} \in \mathbb{R}^D$ corresponds to D bins of the log-magnitude STFT of source k at time frame t and similar $\boldsymbol{y}^{(t)} \in \mathbb{R}^D$. This approximation is based on the sparse nature of speech in time-frequency representations where each bin of a mixture spectrogram is dominated by a single source. The log-magnitude STFT $\boldsymbol{s}_k^{(t)}$ is modeled as GMM according to[1]

$$p(\boldsymbol{s}_k|\boldsymbol{\Theta}_k) = \sum_{m=1}^{M_k} \alpha_k^m \mathcal{N}\left(\boldsymbol{s}_k|\boldsymbol{\theta}_k^m\right) = \sum_{m=1}^{M_k} \alpha_k^m \mathcal{N}\left(\boldsymbol{s}_k|m, \boldsymbol{\theta}_k^m\right), \tag{1}$$

where in the last equality we emphasize that each component m is represented as Gaussian, $M_k \geq 1$ is the number of mixture components of source k, and α_k^m denotes the weight of component m; $\alpha_k^m \geq 0$ and $\sum_{m=1}^{M_k} \alpha_k^m = 1$. The GMM for source k is specified by the parameter set $\boldsymbol{\Theta}_k = \{\alpha_k^m, \boldsymbol{\theta}_k^m\}_{m=1}^{M_k}$, where $\boldsymbol{\theta}_k^m = \{\boldsymbol{\mu}_k^m, \boldsymbol{\Sigma}_k^m\}$ is the mean and diagonal covariance matrix of component m. The parameters of the GMMs are obtained by the EM-algorithm [15].

At each time frame, the observation $\boldsymbol{y}^{(t)}$ is considered to be produced jointly by the two single-source emissions $\boldsymbol{s}_1^{(t)}$ and $\boldsymbol{s}_2^{(t)}$ using the MIXMAX model, i.e. $p(\boldsymbol{y}^{(t)}|\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)}) = \delta\left(\boldsymbol{y}^{(t)} - \max(\boldsymbol{s}_1^{(t)}, \boldsymbol{s}_2^{(t)})\right)$. We obtain the component-conditional observation probability $p(\boldsymbol{y}|m_1, m_2)$ by marginalization over $\boldsymbol{s}_k$, i.e.

$$p(\boldsymbol{y}|m_1, m_2) = \int\int p(\boldsymbol{y}|\boldsymbol{s}_1, \boldsymbol{s}_2) p(\boldsymbol{s}_1|m_1, \boldsymbol{\theta}_1^{m_1}) p(\boldsymbol{s}_2|m_2, \boldsymbol{\theta}_2^{m_2}) \mathrm{d}\boldsymbol{s}_1 \mathrm{d}\boldsymbol{s}_2. \tag{2}$$

This can be solved in closed form using single-component GMMs $p(\boldsymbol{s}_k|m_k, \boldsymbol{\theta}_k^{m_k})$ and the MIXMAX interaction model, i.e.

$$p(\boldsymbol{y}|m_1, m_2) = \prod_{d=1}^{D} \left\{\mathcal{N}(y_d|\theta_1^{m_1,d})\Phi(y_d|\theta_2^{m_2,d}) + \Phi(y_d|\theta_1^{m_1,d})\mathcal{N}(y_d|\theta_2^{m_2,d})\right\}, \tag{3}$$

where y_d denotes the d^{th} element of $\boldsymbol{y}$, $\theta_k^{m_k,d}$ is the d^{th} element of the corresponding mean and variance of the single-speaker model of speaker k, and $\Phi(y|\theta) := \int_{-\infty}^{y} \mathcal{N}(x|\theta)\mathrm{d}x$ represents the univariate cumulative normal distribution (details are in [16, 10]).

Given the observation sequence $\mathcal{Y} = \bigcup_{t=1}^{T} \boldsymbol{y}^{(t)}$, the aim is to infer the best combination of components m_1^* and m_2^* maximizing the conditional distribution, i.e.

$$\left\{m_1^{*,(t)}, m_2^{*,(t)}\right\} = \arg\max_{m_1, m_2} p(\boldsymbol{y}^{(t)}|m_1, m_2). \tag{4}$$

Hence, the search space for separation of two sources is $\mathcal{O}(M_1 \cdot M_2)$ at each t. In [17], we derive bounds for efficient determination of m_1^* and m_2^*.

Once we have found the optimal indices $\left\{m_1^{*,(t)}, m_2^{*,(t)}\right\}$ for all t, we use the corresponding Gaussian components as approximation of the speaker and noise log-magnitude STFTs denoted by $\boldsymbol{\mu}_1^{m_1^*}$ and $\boldsymbol{\mu}_2^{m_2^*}$, respectively. These approximations enable to derive a softmask $G(t, d)$ in frequency domain, i.e.

$$G(t,d) = \frac{\mu_{1,d}^{m_1^{*,(t)}}}{\mu_{1,d}^{m_1^{*,(t)}} + \mu_{2,d}^{m_2^{*,(t)}}}, \tag{5}$$

where $\mu_{k,d}^{m_k}$ denotes the d^{th} frequency bin in $\boldsymbol{\mu}_k^{m_k}$. To re-synthesize time signals, the softmask is multiplied with the original noisy spectrogram $\mathcal{Y}$ and the inverse STFT followed by an overlap-and-add procedure is applied. The phase of the noisy signal is used for reconstruction.

3. Model Adaptation

We might encounter different channel conditions during separation, i.e. the spectral characteristics of each source signal might have changed due to multi-path propagation in a room or a different microphone transfer function. Any mismatch between the source models and the actual condition in a mixture results in a degraded separation accuracy. Model self-adaption tunes the available source models to the specific source characteristics and channel conditions present in a previously unseen recording using only the observed mixture signal. We limit our adaption framework to the adaption of the speaker model only, i.e. w.l.o.g. the first GMM $\boldsymbol{\Theta}_1$ is assumed to be the speaker model. We use MLLR to adapt the means of $\boldsymbol{\Theta}_1$. Extensions to the covariances are considered in future work. We assume the same affine transform for each component in $\boldsymbol{\Theta}_1$, i.e. $\hat{\boldsymbol{\mu}}_1^{m_1} = \boldsymbol{T}\boldsymbol{\xi}_1^{m_1}$, where $\boldsymbol{T}$ is a $D \times (D+1)$ transformation matrix and $\boldsymbol{\xi}_1^{m_1} = (1, \boldsymbol{\mu}_1^{m_1})^T$ is the mean vector and a bias. Since different GMM components model different characteristics of speech, it may be reasonable to assume that an improved adaption performance can be achieved if those components modeling the same characteristic are tied together and updated with a separate transformation matrix [6]. We use a global transformation matrix since the amount of adaption mixture data is limited to a few seconds.

The transformation matrix is obtained by maximizing the log-likelihood $LL(\boldsymbol{T}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ of the transformed speaker GMM, given a signal mixture $\mathcal{Y}$, i.e.

$$LL(\boldsymbol{T}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) = \sum_{t=1}^{T} \ln p(\boldsymbol{y}^{(t)}|\boldsymbol{T}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2). \tag{6}$$

This models the joint distribution of $\mathcal{Y}$ depending on the transformation matrix $\boldsymbol{T}$, i.e. $\ln p(\mathcal{Y}|\boldsymbol{T}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$.

The distribution of the observation at one time instance is

[1] We omit the explicit dependence of random variables on t, where appropriate throughout the manuscript.

(cf. (2))

$$p(\boldsymbol{y}^{(t)}|\boldsymbol{T},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) = \tag{7}$$
$$\int\int p(\boldsymbol{y}^{(t)}|\boldsymbol{s}_1^{(t)},\boldsymbol{s}_2^{(t)})p(\boldsymbol{s}_1^{(t)}|\boldsymbol{T},\boldsymbol{\Theta}_1)p(\boldsymbol{s}_2^{(t)}|\boldsymbol{\Theta}_2)\mathrm{d}\boldsymbol{s}_1^{(t)}\mathrm{d}\boldsymbol{s}_2^{(t)}$$
$$= \sum_{m_1=1}^{M_1}\sum_{m_2=1}^{M_2}\alpha_1^{m_1}\alpha_2^{m_2}\int\int p(\boldsymbol{y}^{(t)}|\boldsymbol{s}_1^{(t)},\boldsymbol{s}_2^{(t)})$$
$$\times\, p(\boldsymbol{s}_1^{(t)}|m_1,\boldsymbol{T},\boldsymbol{\theta}_1^m)p(\boldsymbol{s}_2^{(t)}|m_2,\boldsymbol{\theta}_2^m)\mathrm{d}\boldsymbol{s}_1^{(t)}\mathrm{d}\boldsymbol{s}_2^{(t)}. \tag{8}$$

We apply Jensen's inequality in (6) to construct a lower bound, which is in general easier to optimize [18]. For any distribution $q(\cdot)$, and any joint probability $p(a,b)$, it follows from Jensen's inequality that

$$\ln\sum_a p(a,b) = \ln\sum_a q(a)\frac{p(a,b)}{q(a)} \geq \sum_a q(a)\ln\frac{p(a,b)}{q(a)},$$

and equality holds if and only if $q(a) = p(a|b)$. We systematically apply Jensen's inequality to construct the following sequence of variational lower bounds on $LL(\boldsymbol{T},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)$:

$$LL(\boldsymbol{T},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2) \geq \text{const} + \sum_t\sum_{m_1}\sum_{m_2} q(m_1^{(t)},m_2^{(t)})$$
$$\times \ln\int\int p(\boldsymbol{y}^{(t)}|\boldsymbol{s}_1^{(t)},\boldsymbol{s}_2^{(t)})$$
$$\times\, p(\boldsymbol{s}_1^{(t)}|m_1,\boldsymbol{T},\boldsymbol{\theta}_1^m)p(\boldsymbol{s}_2^{(t)}|m_2,\boldsymbol{\theta}_2^m)\mathrm{d}\boldsymbol{s}_1^{(t)}\mathrm{d}\boldsymbol{s}_2^{(t)}, \tag{9}$$
$$\geq \text{const} + \sum_t\sum_{m_1}\sum_{m_2} q(m_1^{(t)},m_2^{(t)})$$
$$\times \int\int q(\boldsymbol{s}_1^{(t)},\boldsymbol{s}_2^{(t)})\ln p(\boldsymbol{s}_1^{(t)}|m_1,\boldsymbol{T},\boldsymbol{\theta}_1^m)\mathrm{d}\boldsymbol{s}_1^{(t)}\mathrm{d}\boldsymbol{s}_2^{(t)}, \tag{10}$$

where 'const' refers to all terms independent of $\boldsymbol{T}$. Starting with an initial guess for $\boldsymbol{T}$, a local maximum of (6) can be found using the EM algorithm consisting of the following two steps:

E-Step: The variational distributions are set such that the lower bound is tight[2] at the current parameter estimate, i.e.

$$q(m_1^{(t)},m_2^{(t)}) = p(m_1^{(t)},m_2^{(t)}|\boldsymbol{y}^{(t)},\boldsymbol{T}^{(old)},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2), \text{ and} \tag{11}$$
$$q(\boldsymbol{s}_1^{(t)},\boldsymbol{s}_2^{(t)}) = p(\boldsymbol{s}_1^{(t)},\boldsymbol{s}_2^{(t)}|\boldsymbol{y}^{(t)},m_1,m_2,\boldsymbol{T}^{(old)},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2), \tag{12}$$

where the posterior of the components in (11) can be obtained by using Bayes rule in (3) and assuming the MIXMAX interaction model.

M-Step: The lower bound of the $LL(\cdot)$ in (10) also known as auxiliary function is maximized with respect to the parameter $\boldsymbol{T}$, i.e.

$$\boldsymbol{T}^* = \arg\max_{\boldsymbol{T}} Q(\boldsymbol{T}) = \arg\max_{\boldsymbol{T}}\sum_t\sum_{m_1}\sum_{m_2} q(m_1^{(t)},m_2^{(t)})$$
$$\times\, \mathrm{E}_{(\boldsymbol{s}_1^{(t)},\boldsymbol{s}_2^{(t)})}\left\{\ln p(\boldsymbol{s}_1^{(t)}|m_1,\boldsymbol{T},\boldsymbol{\theta}_1^m)\right\}, \tag{13}$$

where the unknown single-speaker spectrum $\boldsymbol{s}_1^{(t)}$ has been replaced by its conditional expected value where the expectation $\mathrm{E}_{(\boldsymbol{s}_1^{(t)},\boldsymbol{s}_2^{(t)})}\{\cdot\}$ is with respect to the distribution in (12).

[2]Up to a term that does not depend on adaption parameters.

In the following, we introduce parameter $\boldsymbol{T}$ in GMM $\boldsymbol{\Theta}_1$ and derive the update equations for the M-Step of the EM algorithm. Each mixture component for $\boldsymbol{s}_1^{(t)}$ is of the form

$$p(\boldsymbol{s}_1^{(t)}|m_1,\boldsymbol{T},\boldsymbol{\theta}_1^m) = \mathcal{N}(\boldsymbol{s}_1^{(t)}|\boldsymbol{T}\boldsymbol{\xi}_1^{m_1},\boldsymbol{\Sigma}_1^{m_1}). \tag{14}$$

Inserting (14) in the auxiliary function $Q(\boldsymbol{T})$ for the speaker model in (13) results in

$$Q(\boldsymbol{T}) = \sum_t\sum_{m_1}\sum_{m_2} q(m_1^{(t)},m_2^{(t)})$$
$$\times\, \mathrm{E}_{(\boldsymbol{s}_1^{(t)},\boldsymbol{s}_2^{(t)})}\left\{\ln\mathcal{N}(\boldsymbol{s}_1^{(t)}|\boldsymbol{T}\boldsymbol{\xi}_1^{m_1},\boldsymbol{\Sigma}_1^{m_1})\right\}. \tag{15}$$

Since $Q(\cdot)$ is concave in $\boldsymbol{T}$, a global optimum can be obtained by setting the derivative to zero [19], i.e. $\frac{\partial Q(\boldsymbol{T})}{\partial \boldsymbol{T}} = 0$, and we obtain

$$\sum_{m_1}\boldsymbol{A}_{m_1}\boldsymbol{T}\boldsymbol{B}_{m_1} = \boldsymbol{C}, \tag{16}$$

where

$$\boldsymbol{A}_{m_1} = \sum_{m_2}\sum_t q(m_1^{(t)},m_2^{(t)})(\boldsymbol{\Sigma}_1^{m_1})^{-1} \tag{17}$$
$$\boldsymbol{B}_{m_1} = \boldsymbol{\xi}_1^{m_1}(\boldsymbol{\xi}_1^{m_1})^T \tag{18}$$
$$\boldsymbol{C} = \sum_{m_1}\sum_{m_2}\sum_t q(m_1^{(t)},m_2^{(t)})$$
$$\times\, (\boldsymbol{\Sigma}_1^{m_1})^{-1}\mathrm{E}\left\{\boldsymbol{s}_1^{(t)}|\boldsymbol{y}^{(t)},m_1,m_2,\boldsymbol{T},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2\right\}(\boldsymbol{\xi}_1^{m_1})^T. \tag{19}$$

In (19), $\mathrm{E}\left\{\boldsymbol{s}_1^{(t)}|\boldsymbol{y}^{(t)},m_1,m_2,\boldsymbol{T},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2\right\}$ is the expected single-speaker spectrum conditioned on the observation at time t and the components m_1 and m_2. The d^{th} dimension of this expectation is calculated as

$$\mathrm{E}\left\{s_{1,d}^{(t)}|y_d^{(t)},m_1,m_2,\boldsymbol{T},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2\right\} =$$
$$\frac{y_d^{(t)}\Psi_1^{m_1,d} + \left(\mu_1^{m_1,d} - (\sigma_1^{m_1,d})^2\Psi_1^{m_1,d}\right)\Psi_2^{m_2,d}}{\Psi_1^{m_1,d}+\Psi_2^{m_2,d}}, \tag{20}$$

where

$$\Psi_k^{m_k,d} = \frac{\mathcal{N}(y_d|\theta_k^{m_k,d})}{\Phi(y_d|\theta_k^{m_k,d})} \tag{21}$$

is the ratio of the normal density and the cumulative normal distribution of observation y_d. For a derivation of (20), we refer the reader to [20].

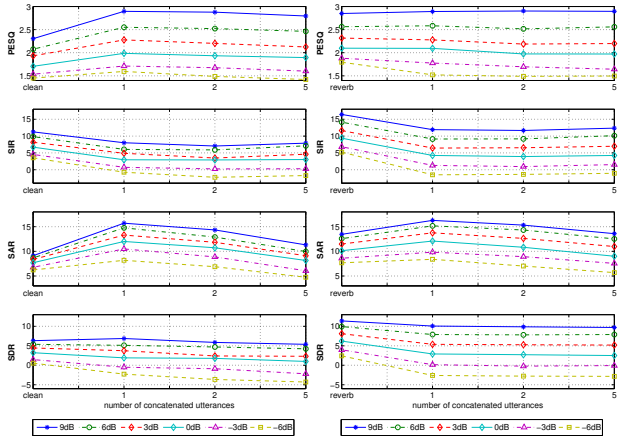
The matrix equation in (16) can be solved in closed form [21], i.e.

$$\text{vec}(\boldsymbol{T}) = \left(\sum_{m_1}(\boldsymbol{B}_{m_1})^T\otimes\boldsymbol{A}_{m_1}\right)^{-1}\text{vec}(\boldsymbol{C}), \tag{22}$$

where $\otimes$ denotes the Kronecker product and $\text{vec}(\boldsymbol{T})$ is a vector obtained by sequentially stacking the columns of $\boldsymbol{T}$.[3]

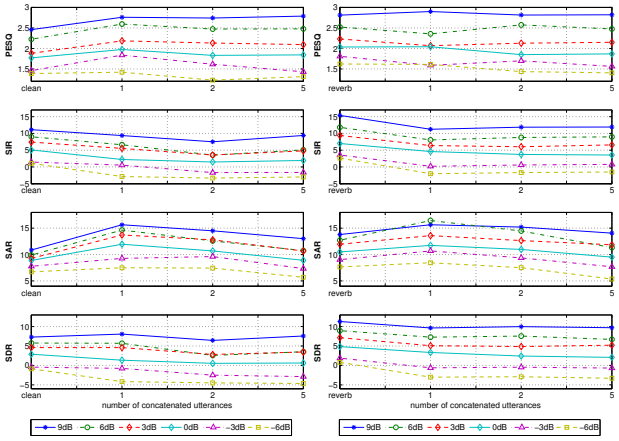
The relevant steps of the adaption algorithm can be summarized as follows: During the E-Step, the expectation of the single source spectrum and the component posterior are estimated from the signal mixture based on the current models. During the M-Step, the expected single source spectrum used as surrogate for the unknown single source spectrum and the component posteriors are employed for determining the adaption parameters $\boldsymbol{T}$ for the speaker.

[3]Using the Kronecker product, a product of three matrices $\boldsymbol{ATB}$ can be re-expressed as $\text{vec}(\boldsymbol{ATB}) = (\boldsymbol{B}^T\otimes\boldsymbol{A})\text{vec}(\boldsymbol{T})$ [21].



(a) (b)

Figure 1: Mean performance scores for self-adaption performed for 1, 2 and 5 concatenated noisy mixtures: (a) clean SD models, (b) reverb SD models.



(a) (b)

Figure 2: Mean performance scores for self-adaption performed for 1, 2 and 5 concatenated noisy mixtures: (a) clean SI models, (b) reverb SI models.

4. Experiments

We use the proposed framework to perform self-adaption on signal mixtures from the small vocabulary track of the CHiME 2 challenge [22]. We use the PESQ measure [13] and the BSS EVAL toolkit [14]. The PESQ measure returns a mean opinion score (MOS) between 0.5 and 4.5. The BSS Eval measures are the signal-to-distortion (SDR) ratio, the signal-to-interference (SIR) ratio, and the signal-to-artifact (SAR) ratio in [dB]. Both the PESQ measure and the BSS EVAL toolkit require the true source signals for score calculation. Hence, only the training set is applicable for evaluation. The development/test set do not provide the clean speech and reverberated speech utterances as reference. Therefore, we divided the training set consisting of 500 utterances for each of the 34 speakers at random into an actual training set of 400 utterances and a test set consisting of the remaining 100 utterances.

We learn GMM $\boldsymbol{\Theta}_2$ modeling the background noise from a subset of 30 utterances of the isolated noise data. The speaker GMMs $\boldsymbol{\Theta}_1$ are either learned using all 400 clean (clean) or reverberated (reverb) single speaker utterances from their respective training set. Furthermore, the speaker models are either trained in an SD way for each of the 34 speakers or in an SI way on 400 utterances chosen randomly from the training sets of all but speaker IDs 2, 5, 22, and 23. These speakers are used for evaluating the SI models. Each GMM consists of 128 components. The features $\boldsymbol{s}^{(t)}$ and $\boldsymbol{y}^{(t)}$ are based on the log-magnitude of the spectrogram of the signals computed via the 1024 point STFT, using a Hamming window of length 32ms and step size of 10ms, i.e. $D = 513$. The sampling frequency is $f_s = 16$kHz. The noisy mixtures are available with signal to noise ratios (SNR) ranging from -6dB to 9dB in 3dB steps. The performance is evaluated for each SNR level separately. Self-adaption is performed on 1, 2 or 5 concatenated time-frequency representations of noisy mixtures picked at random from the test set of the respective speaker. In particular, for each SNR level and number of concatenated mixtures we generate three test utterances. Adaption is applied up to a maximum of 120 iterations.

Figure 1 show the mean score averaged over the test mixtures of all 34 speakers for the self-adaption in the SD case starting from clean and reverb models. Mixtures with low SNR (i.e., those with an SNR below 0dB) do not benefit from self-adaption. The reverb models show generally a better performance. This is due to the superior initial model. For the reverb models no mismatch between training and testing is present, i.e. the PESQ and SDR measures do not significantly enhance with self-adaption. The improvement in terms of SAR is canceled out by the deterioration of the SIR. For the clean models, a slight improvement in terms of SAR is visible which is reflected by the PESQ score. The SIR shows a slight degradation, the SDR is almost constant. The PESQ is not influenced by the amount of adaption data. Altogether, it seems that on average one noisy utterance is sufficient for self-adaption in order to achieve good performance.

Figure 2 show the corresponding results for the SI case, again for clean and reverb models. The SI models show an overall slightly inferior performance, nevertheless the same behavior as for the SD models can be observed. Again, we reach results for the adapted speaker independent models comparable to those of the speaker dependent ones. In summary, SCSS benefits from self-adaption, especially for large SNRs. Furthermore, it can be observed that a small amount of adaption data is sufficient for the proposed algorithm.

5. Conclusions

We developed an MLLR adaption framework capable of adapting pre-trained speaker models onto previously unseen conditions using mixture data only. All developed methods are empirically compared using data from the CHiME 2 challenge. We were able to show that self-adaption improves the PESQ measure for GMMs trained on clean single speaker utterances. Using this adaption framework, we are able to achieve with speaker independent models almost the same performance as with speaker dependent models. The proposed model adaption algorithm is able to achieve this with a minimum amount of adaption data. Furthermore, self-adaption is more useful for mixtures with larger SNRs. We plan to extend our MLLR-based adaption framework to additionally adapt the covariances of the speaker models and the second source model.

6. References

[1] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, ser. IEEE Press. J. Wiley and Sons Ltd, 2006.

[2] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH*, Switzerland, Sep. 2003, pp. 1009–1012.

[3] ——, "One microphone source separation," in *Neural Information Proc. Sys.*, 2000, pp. 793–799.

[4] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[5] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoustics*, 2003.

[6] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

[7] M. Gales and P. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249 – 264, 1996.

[8] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

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

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

[11] R. Rose, E. Hofstetter, and D. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, 1994.

[12] R. Weiss and D. Ellis, "A variational EM algorithm for learning eigenvoice parameters in mixed signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 113 –116.

[13] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *ITU-T Recommendation P.862*, 2000.

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

[15] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. B30, pp. 1–38, 1977.

[16] M. Wohlmayr, M. Stark, and F. Pernkopf, "A probabilistic interaction model for multipitch tracking with factorial hidden Markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 799 –810, 2011.

[17] M. Wohlmayr and F. Pernkopf, "Model-based multiple pitch tracking using factorial HMMs: Model adaptation and inference," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 8, pp. 1742–1754, 2013.

[18] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[20] M. Wohlmayr, "Probabilistic model-based multiple pitch tracking of speech," Ph.D. dissertation, Graz University of Technology, 2012.

[21] T. Moon and W. Stirling, *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall, 2000.

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