TOWARDS SOURCE-FILTER BASED SINGLE SENSOR SPEECH SEPARATION

Michael Stark and Franz Pernkopf

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria
michael.stark@ieee.org, pernkopf@TUGraz.at

ABSTRACT

We present a new source-filter based method to separate two speakers talking simultaneously at equal level mixed into a single sensor. First, the relation between the spectral whitened mixture and the speakers excitation signals is analyzed. Therefore, a factorial HMM capturing also time dependencies is exploited. Then, the estimated excitation signals are combined with best fitting vocal tract information taken from a trained dictionary. We report results on the database of Cooke considering 108 speech mixtures. The average improvement of 2.9 dB in SIR for all data is lower but not significantly lower compared to the Gaussian mixture method which relies on known pitch-tracks. Although the performance is currently moderate we believe in this approach and its significance towards the development of speaker independent single sensor speech separation.

Index Terms— Speech Separation, Source-Filter Representation, Hidden Markov Model, Vector-Quantization

1. INTRODUCTION

The separation of two sound sources mixed into a single channel is in general an ill-posed problem, i.e., recovering the exact waveforms of the underlying signals is impossible without further knowledge about the sources or their interrelationship. For explicit models, the individual source characteristics are stored during a training phase. Afterwards the model is used as prior knowledge about the source without considering the interfering component. The two most established representatives of explicit models are the factorial-max vector quantization in [1] and the factorial-max HMM (FHMM) [2] which also incorporates time dependencies.

In contrast, implicit models try to mimic the ability of the human auditory system. Here, the mixture is a scene to be organized and particular extracted components are merged to form output streams of individual sources. Therefore, features like common on- and off-sets, harmonicity, and amplitude- and frequency modulations are extracted and considered for the signal separation [3]. The computational auditory scene analysis systems in [4] and [5] are the most important representatives. Both systems are heavily based on harmonicity as cue for separation.

Motivated by the factorization of the signal mixture into distinct parts as implicit models does [3], we propose to factorize the spectrum of the mixture in a fine and a coarse spectral structure. For speech the excitation signal produced by the vocal folds mainly represents the fine spectral structure, whereas, the coarse spectral structure can be linked to the shaping of the vocal tract. Separating the coarse from the fine structure of the speech mixture results in a spectral whitening and only the fine structure of the individual speakers remains. Based on this decomposition, the true fine spectral structure of the underlying signals can be estimated for given excitation models. To capture time dependencies and hence avoid permutations among both speakers, HMMs are utilized to include prior knowledge about the respective excitation signals. Using the FHMM approach the excitation signals can be estimated from the speech mixture. Having an estimate related to the excitation signal at hand, the coarse spectral structure can be estimated similarly to an analysis by synthesis problem in speech coding [6], where a trained vector-quantizer codebook models the vocal-tract prior knowledge.

In section 2 the system is presented and models are introduced. Reasons supporting this decomposition are given in sec. 3. Afterwards, the experimental setup is introduced and performance measures are defined. Finally, in sec. 5 we draw some conclusions and point to future aspects.

2. SEPARATION ALGORITHM

We assume a linear instantaneous mixture model of two speakers:

$$x[t] = s_1[t] + s_2[t] + \nu[t], \quad t = [1, \ldots, T], \qquad (1)$$

where $s_i[t]$, with $i \in \{1, 2\}$ is the respective speaker, $\nu[t]$ is a noise signal (e.g., sensor and/or background noise) and T denotes time. Moreover, we consider the component sources to

This research was carried out in the context of COAST-ROBUST, a joint project of Graz University of Technology, Philips Speech Recognition Systems, and Sail Labs Technology. We gratefully acknowledge funding by the Austrian KNet Program, ZID Zentrum fuer Innovation und Technology, Vienna, the Steirische Wirtschaftsfoerderungsgesellschaft mbh, and the Land Steiermark.

be combined at equal level. Before model training, the component signals are divided in their source and filter related parts using the linear prediction technique (LPC) [6] as shown in fig.1. For every speech segment $d = [1, \dots, D]$ source and filter are related as $s_i[t] = -\sum_{n=1}^{N} c_n \cdot s_i[t-n] + e_i[t]$, where N specifies the filter order, c_n the filter coefficients and $e_i[t]$ denotes the residual or excitation signal. Afterwards, the excitation signal of each speaker is firstly transformed to the log-frequency domain $\log|E_i|$ and secondly, used to learn an HMM λ_i^{HMM}. For convenience, time-domain signals will be in lower case and frequency domain signals in upper case letters. Furthermore, signals in the log-frequency domain, i.e., $\log|E_i|$, will be abbreviated by lE_i in future. Additionally, as shown in fig.1 a vector-quantizer (VQ) Codebook, i.e., k-means [6] is trained to capture each speakers vocal tract information. In order to enhance filter stability and increase robustness against quantization errors the LPC coefficients are represented by their line spectral frequencies (LSF) [6]. At

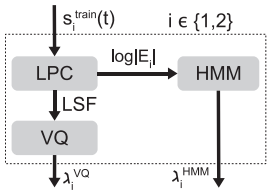


Fig. 1. Block diagram of the training stage.

the beginning of the decoder stage, shown in fig. 2, spectral whitening of the speech mixture $x[t]$ using LPC is performed. Following, the separation is carried out in two steps. First, the remaining spectral fine-structure lE_x in conjunction with the trained models λ_i^{HMM} are utilized as input for the FHMM. The FHMM decodes the excitation mixture lE_x and extracts the individual excitation signals $|\hat{E}_i|$. Given the mixture $|X|$, the models λ_i^{VQ} of each speaker, and the estimated excitation signals, we are able to estimate the best fitting vocal tract information in the *VTE Separation* unit. The best fitting envelope is extracted from λ_i^{VQ} for a particular instant of time in the l_2-norm. The provided output $|\hat{S}_i|$ is an estimate of the underlying signals.

2.1. Source Representation- FHMM

In order to track every speaker over time, i.e., capture also time dependencies, and hence avoid permutations of sources an HMM is used to model vocal-fold related information. The emission distribution contains the vocal-fold specific attributes whereas the transition matrix covers temporal constraints. The posterior probability given the speech mixture and the FHMM in general is given as $p(\text{lE}_1, \text{lE}_2|X) \propto p(\text{lE}_x|\text{lE}_1, \text{lE}_2) \cdot p(\text{lE}_1) \cdot p(\text{lE}_2)$, where $p(\text{lE}_i)$ are the independent priors λ_i^{HMM} and $p(\text{lE}_x|\{\text{lE}_i\})$ is the likelihood function defined as:

$$p(\text{lE}_x|k_1, k_2) = \mathcal{N}(\text{lE}_x|\max(\mu_{k_1}^{\text{lE}_1}, \mu_{k_2}^{\text{lE}_2}), \Sigma), \qquad (2)$$

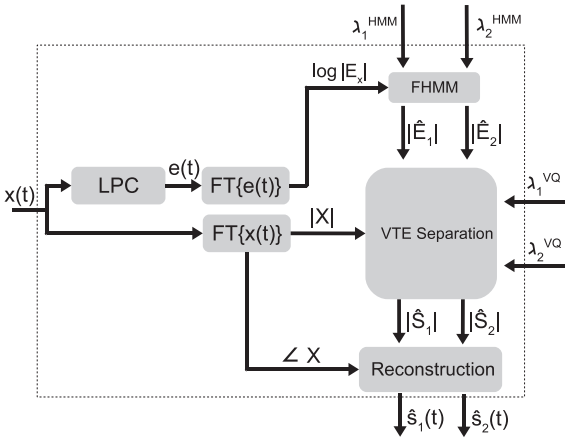


Fig. 2. Block diagram of the separation algorithm.

where $\mathcal{N}$ denotes the normal density, $\max$ is the element-wise maximum operator, k_i are the state indices associated with a particular mean $\mu_{k_i}^{\text{lE}_i}$, and Σ is the covariance matrix shared by all speakers. Introducing time-dependency, the best fitting state for each source is extracted according to:

$$\{k_1^\star, k_2^\star\} = \arg\max_{k_1,k_2} \Big[p\big(\text{lE}_x|k_1(d), k_2(d)\big)$$
$$p\big(k_1(d)|k_1(d-1), \lambda_1^{\text{HMM}}\big)\, p\big(k_2(d)|k_2(d-1), \lambda_2^{\text{HMM}}\big)\Big],$$

where d denotes time segments, $k_i(d)$ the state index for a particular instant of time and $\{k_i^\star\}$ the most probable state of source i given the current observation lE_x and the respective transition probability. Finally, the best sequence can be found using the Viterbi algorithm and the mean $\mu_{k_i^\star(d)}^{\text{lE}_i}$ of the active state is considered to be an estimate for the fine spectral structure of each speaker.

2.2. Envelope Modeling

Given an estimate of each speakers excitation signal $|\hat{E}_i|$ we are now ready to further impose vocal-tract envelope (VTE) related information on $|\hat{E}_i|$ given the statistical models, i.e., λ_i^{VQ}, of the VTE and the observed speech mixture. This unit delivers the unmixed speech signals. We can estimate the posterior probability similar as above assuming uniform priors and a spherical covariance. The best state indices can be found by minimizing the l_2-norm as:

$$\{q_i^\star\} = \arg\min_{\{q_i\}} \Big|\Big|\, |X| - \sum_i E_{k_i^\star} \cdot \lambda_i^{\text{VQ}}(q_i) \,\Big|\Big|_2, \qquad (3)$$

where $q_i^\star$ and q_i are codeword indices. In contrast to the FHMM we do not model any time dependency. The VTE information related to each codeword $q_i^\star$ at every time step can be combined with the estimated excitation signal and the mixed phase to form the estimated speaker:

$$\hat{s}_i = \mathcal{FT}^{-1}\{|\hat{E}_i| \cdot |\hat{H}_i| \cdot \exp j\angle X\}, \qquad (4)$$

where $\mathcal{FT}^{-1}$ denotes the inverse Fourier transform and $|\hat{H}_i| = \lambda_i^{\mathrm{VQ}}(q_i^\star)$ is the estimated VTE envelope. Finally, the speech output sequence is built using the overlap-add method.

3. VALIDATION

In the previous sections all system modules as well as their dependencies have been introduced. However, the relationship between the spectral whitened mixed signal lE_x and the individual LPC residual signals lE_i needs further elaboration. Although there is no closed form solution between these quantities we will explore a reasonable approximation. As defined in Eq. 1 the two component speech signals are additively related in the time domain. This additivity still holds in the Fourier domain assuming that the phase information is included. Depicting the signals with their magnitude and phase values, this relation is given as:

$$X^2 = |S_1|^2 + |S_2|^2 + 2 \cdot |S_1|\,|S_2|\,\cos(\phi), \qquad (5)$$

where ϕ is the phase difference between S_1 and S_2. Recently, [7] have shown that taking the expected value over the logarithm of Eq. 5 results in the max-approximation, i.e., $\log|X| = \max(\log|S_1|, \log|S_2|)$, assuming a uniformly distributed phase between $[0, \ldots, \pi]$. To make it clear, if source one exhibits more energy in a specific time-frequency bin compared to source two, the bin is exclusively assigned to the first source and vice-versa, i.e., the mix-max-approach [7]. The derivation in [7] is independent of any signal characteristics. Hence, it also holds for the excitation signals:

$$\log|E_{s_1,s_2}| = \max[\,\log|E_1|, \log|E_2|\,], \qquad (6)$$

where E_{s_1,s_2} is the Fourier transform of the sum of the respective vocal fold excitations in the time domain, i.e., $e_{s1,s2}[t] = e_1[t] + e_2[t]$. Thus, the only relation to show is if $|E_{s_1 s_2}| \approx |E|$ is valid. Therefore we provide an experiment where two utterances of the same male speaker are mixed at equal level corresponding to an SIR of 0 dB. The mean segmental Source-to-Interference Ratio (SIR_{seq}) over various speech segments is still over 16 dB, where the SIR_{seq} is measured in the log-frequency domain. This is a fairly good value and the approximation can be assumed to be valid. The SIR_{seq} is defined as follows: $\mathrm{SIR}_{seq} = \frac{1}{D}\sum_{d=1}^{D} 10\log_{10}\frac{\sum_f \mathrm{lS}_i(f,d)^2}{\sum_f(\mathrm{lS}_i(f,d)-\hat{\mathrm{lS}}_i(f,d))^2}$, with f the frequency bin index. Figure 3 shows the original excitation mixture $e_{s1,s2}[t]$, the mixture found by spectral whitening $e_x[t]$, and the error signal defined as the difference between the two signals. The SIR_{seq} error of 16 dB, results in an SIR of 8 dB in the time domain shown at the bottom of fig. 3.

4. EXPERIMENTS

The database recently provided by Cooke et al. [8] for the single channel speech separation task has been selected to evaluate the proposed separation algorithm. We compare this algorithm with the methods in [9] where the excitation signal was generated given the respective true pitch-track. The sampling frequency has been resampled to 16 kHz for all files. For calculating the spectrogram the signals have been cut into segments of 32 ms with time shifts of 16 ms. We have used the LPC method of order 24 to separate the filter from the source signal and transformed the filter coefficients to the LSF representation.

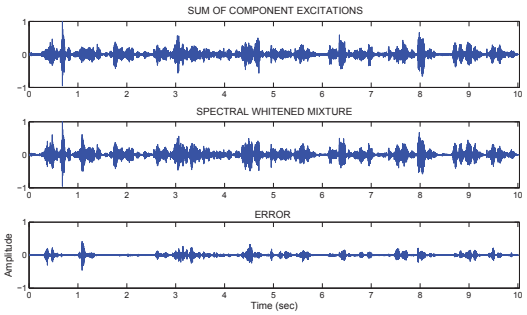


Fig. 3. (a) True excitation mixture $e_{s1,s2}[t]$, (b) Spectral whitened excitation mixture $e_x[t]$, (c) Corresponding error.

For training the speaker models, the remaining files not used for testing are employed, corresponding to approximately 15 min of speech material for each speaker. Two randomly selected male and female speakers, each uttering 3 sentences as shown in table 1 were used for testing. For simplicity we will call these speakers FE1, FE2, MA1 and MA2 in the remainder of this section.

FE1	speaker 18	"lwixzs"	"sbil4a"	"prah4s"
FE2	speaker 20	"lwwy2a"	"sbil2a"	"prbu5p"
MA1	speaker 1	"pbbv6n"	"sbwozn"	"prwkzp"
MA2	speaker 2	"lwwm2a"	"sgai7p"	"priv3n"

Table 1. Labels of speakers and file names used for testing.

For testing all files are mixed at a level of 0 dB SIR and all possible combinations between target speakers and their interfering speakers are evaluated, resulting in altogether 108 mixed signals. Audio examples of the mixtures and the separated files are available at https://www.spsc.tugraz.at/people/michael-stark/SCSS.

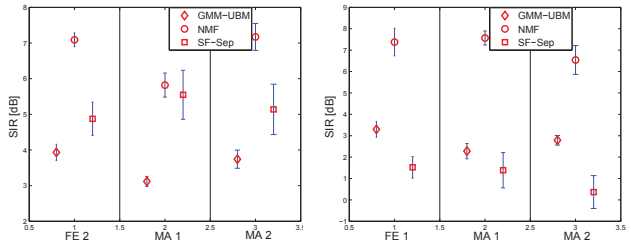
To evaluate the performance the signal-to-interference ratio (SIR) has been used. To avoid synthesis distortions affecting the quality assessment the SIR have been measured by comparing the magnitude spectrograms of the true source and the separated signal as:

$$\mathrm{SIR}_i = \frac{\sum_{f,d}|S_i(f,d)|^2}{\sum_{f,d}(|S_i(f,d)| - |\hat{S}_i(f,d)|)^2},$$

where $f = [1, \ldots, F]$ is the frequency bin index and S_i and

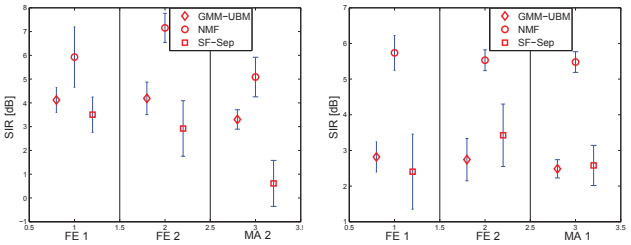
$\hat{S}_i$ are the source and separated signal spectra of speaker i.

For the VQ codebook a dictionary size of 512 has been chosen. The dimension of the model parameters corresponds to the LPC order, i.e., 24. For training we used 200 iterations learning the VQ Codebook and 5 EM-steps for the FHMM. The FHMM method was trained with 1000 states using one Gaussian component per state. The priors are assumed to be uniformly distributed. To reduce complexity and make this method still tractable for estimating the state sequence, a beam search [10] restricting the search to the best 5000 candidates has been used. Figures 4 and 5 report the mean and the standard deviation of the SIR for each target speaker to its interfering speakers. We compare the performance of our method (SF-Sep) to two other approaches where VTE information has been estimated using a Gaussian mixture model (GMM-UBM) and non-negative matrix factorization (NMF) [9]. Both methods are based on a known synthetic excitation signal.



(a) Target Speaker FE1 (b) Target Speaker FE2

Fig. 4. Mean and standard deviation of the SIR from target to interfering speakers. The marker shapes identify methods.



(a) Target Speaker MA1 (b) Target Speaker MA2

Fig. 5. Mean and standard deviation of the SIR from target to interfering speakers. The marker shapes identify methods.

Observing the results, in general the methods with the *a priori* given pitch-tracks used to synthesize a harmonic signal as excitation, slightly outperform the proposed algorithm. This result is not surprising due to the prior knowledge about the true pitch value for every instant of time. Our proposed method works totally unsupervised, i.e., without this knowledge. In general, a good match between the true and estimated vocal-tract envelope can only be found if the excitation estimation works well not only for every time frame but also over time. In any case, the component signal estimates $\hat{S}_i$ can be further used to estimate a binary mask which can be applied on the mixture. The average SIR over all files is shown in tab. 2. An increase of model complexity of the FHMM model from 1000 to 2000 states might be sufficient to outperform the GMM-UBM based method.

GMM-UBM	NMF	SF-Sep
3.2 ± 0.60 dB	6.4 ± 0.70 dB	2.9 ± 1.72 dB

Table 2. Mean and standard deviation of global SIR.

5. CONCLUSION

In this paper, we proposed to tackle the single channel source separation problem by splitting up the signal into its coarse and fine spectral structure. We introduced the model related to each component, namely, a factorial HMM to find a time continuous vocal-fold excitation related signal for each speaker. With these estimates at hand a trained VQ has been employed to further impose matching vocal tract information. We validate this approach by showing that the spectral whitened speech mixture is well approximating the sum of the component excitation signals. We compared the performance to two other methods which search for matching vocal tracked information but rely on the known pitch-track. At the current state of development our proposed method delivers a slightly but not significantly lower performance compared to the Gaussian mixture approach.

6. REFERENCES

[1] Sam T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *EUROSPEECH*, Geneva, Switzerland, Sept. 2003, pp. 1009–1012.

[2] Sam T. Roweis, "One microphone source separation," in *Neural Information Processing Systems, NIPS*, 2000, pp. 793–799.

[3] DeLiang Wang and Guy J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, IEEE Press. John Wiley and Sons Ltd, New Jersey, Oct. 2006.

[4] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[5] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," vol. 10, no. 3, pp. 684–697, 1999.

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

[7] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Monaural speech segregation based on fusion of source-driven with model-driven techniques," *Speech Communication*, vol. 49, no. 6, pp. 464–476, June 2007.

[8] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," in *JASA*, 2006, number 120, pp. 2421–2424.

[9] M. Stark, F. Pernkopf, T. V. Pham, and G. Kubin, "Vocal-tract modeling for speaker independent single channel source separation," in *Cog. Inf. Proc. Workshop*, Greece, June 2008.

[10] Frederick Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, January 1998.