VOCAL-TRACT MODELING FOR SPEAKER INDEPENDENT SINGLE CHANNEL SOURCE SEPARATION

Michael Stark, Franz Pernkopf, Tuan Van Pham, Gernot Kubin

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria
michael.stark@ieee.org, pernkopf@TUGraz.at, v.t.pham@tugraz.at, g.kubin@ieee.org

ABSTRACT

In this paper, we investigate two statistical models for the source-filter based single channel speech separation task. We incorporate source-driven aspects by pitch estimation in the model-driven method which models the vocal-tract part as *a priori* knowledge. This approach results in a speaker independent (SI) source separation method. For modeling the vocal tract filters Gaussian mixture models (GMM) and non-negative matrix factorization are considered. For both methods, the final fusion of the source and filter parameters results in a reformulation of the models that finally are used for separation. Furthermore, for the GMM method we propose a new gain compensation and pitch adjustment method. Performance is evaluated and compared to the speaker dependent (SD) factorial Hidden Markov Model [1]. Although the SD method delivers the best quality our SI methods show promising results and possess a lower complexity in terms of used parameters.

Index Terms— single channel speech separation, source-filter modeling, Gaussian mixture model, non-negative matrix factorization

1. INTRODUCTION

Recently, Single Channel Source Separation (SCSS) [2] is enjoying great popularity in the field of signal processing. The task of SCSS is to separate the linear instantaneous mixture $y = s_1 + s_2$ of two source signals s_1 and s_2.

This paper is restricted to the most challenging case of two concurrent speech signals, overlapping in time and frequency. As the number of observations y is less than the unknowns s_1 and s_2, we are talking of an under-determined problem. In principle, there exist two main approaches to solve the problem: One, motivated by the remarkable ability of the human auditory system to recover individual sound components in adverse environments, called computational auditory scene analysis [2], also often known as source-driven approach. Based on auditory motivated features, these systems try to mimic the segregation performance of the human brain. The other method relies on a statistical approach, incorporating *a priori* knowledge of the sources to solve the under-determined problem and is called model-based method [2] emerging from the field of Blind Source Separation (BSS) [3]. This study is based on the idea first proposed by [4] to combine both, the source-driven with the model-driven approach. Radfar et al. [4] suggest to also consider the speech signal characteristics and use them as an additional cue. Using this as basis the signal can be decomposed into an excitation signal and a filter representing the vocal tract. The source-driven part extracts the fundamental frequency (F0) of each speaker used to create an artificial excitation. The vocal-tract filters (VTFs) are estimated based on a probabilistic model-driven approach. This decomposition results in a speaker independent (SI) system in contrast to most other methods. Speaker independency is achieved by training a SI VTF model. This paper in particular investigates the model driven part and its impact on separation performance by introducing two statistical methods, one based on the Gaussian mixture model (GMM) approach and the other on non-negative matrix factorization (NMF). We propose a new gain compensation and pitch adjustment method for the GMM model resulting in a reduction of the model complexity. Additionally, we apply NMF the first time for this source-filter based approach. The performances of the introduced methods are compared to the speaker dependent factorial Hidden Markov Model (F-HMM) based separation method [1].

The remaining paper is structured as follows: In sec. 2 we introduce the fusion of the model-driven and the source-driven approach. The proposed VTF models are introduced in sec. 3. The experimental setup and results are presented in sec. 4. Finally, we conclude and give future perspectives in sec. 5.

2. COMBINING THE SOURCE-DRIVEN AND THE MODEL-DRIVEN APPROACH

In the source-filter model the speech signal is composed of an excitation signal that is shaped by the vocal tract acting as a filter process. The final speech output is the convolution ($\star$) of the excitation with the VTF in the time domain resulting in a multiplicative ($\times$) relation in the frequency domain as:

$$s_i = e_i \star h_i \rightarrow S_i = E_i \times H_i,$$

where s is a speech segment, with e its excitation and h its VTF. The speaker index is given as $i \in \{1, 2\}$. Signals in time domain are denoted by lower case and signals in the magnitude spectrum by uppercase characters.

The overall system is shown in fig. 1 and consists of the following building blocks: A single pitch extraction unit followed by the excitation generation unit is representing the source-driven part. VTFs, know as spectral envelopes, are extracted from SI training data x_{train} and used to train a model λ^{GMM} or λ^{NMF}. The excitation signals e_i and the model are used to formulate the speaker dependent model λ_{SD}^{GMM} or λ_{SD}^{NMF} and perform separation. We recon-

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.

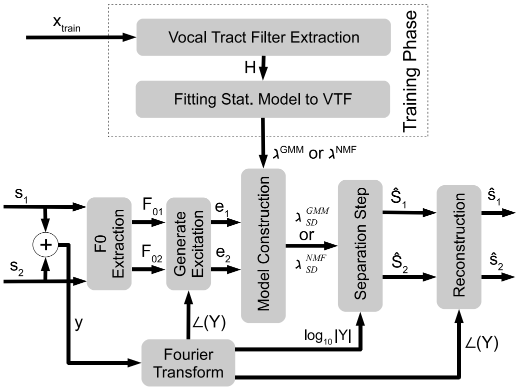


Fig. 1. Blockdiagram of the separation algorithm.

struct the underlying signals s_i by finding the most probable signals $\hat{S}_1$ and $\hat{S}_2$ having produced the mixture Y. This paper is restricted to the model-driven part, assuming that the speakers pitch tracks are known. In a speech mixture multi-pitch detection methods are used to extract the respective F0 of the underlying signals. In decoupling the problem into two subtasks, namely, the multi-pitch detection and the estimation of the VTFs we firstly reduce the complexity and secondly allow to tune parameters for each system separately. For the given F0 of the respective source and the phase of the mixed signal $\angle(Y)$ the harmonic part of the excitation signal is modeled as:

$$\hat{e}_i(t, \omega_{0_i}, \angle(Y(u))) = \sum_{u=1}^{U(\omega_{0_i}, f_{max})} sin(u\,\omega_{0_i} t + \angle(Y(u))), \quad (1)$$

where U denotes the number of harmonics up to a specified highest frequency f_{max} set to 4kHz, ω_0 is F0 in radians and $t = [1, \dots, T]$ is the time index. For unvoiced signals, a Gaussian random signal is used as an excitation model and for voiced signal, a high-pass filtered Gaussian random signal is added to equation 1 above the frequency f_{max}. Equation 1 is similar to the harmonic plus noise modeling [5] but with missing amplitude weightings for the harmonics which will be provided in our case by the VTF estimation algorithm.

In the training phase, the VTF or spectrum envelope, is extracted using Linear Prediction (LPC) [6] and a smoothing as post-processing. The envelopes are extracted from speech segments from speakers of a large training database not used for separation. Hence, we deal with a SI SCSS system. These envelopes are used to train the model described in the next section.

3. VOCAL TRACT FILTER MODELS

In this section, two different statistical separation algorithms are proposed. The first is based on the Maximum Likelihood (ML) search performing the separation using the mixture maximization (MIXMAX) approach [7] and the second uses NMF to separate the mixed signal into its constituents.

3.1. Maximum Likelihood Based Source Separation

A GMM is trained to model the density $p_h(\log_{10}(H))$ of the extracted VTFs in the log-frequency domain:

$$p_h(\log_{10}(H)) = \lambda^{GMM}(c, \mu, \Sigma) = \sum_{k=1}^{K} c^k\, \mathcal{N}(\log_{10}(H), \mu^k, \Sigma^k),$$

where K is the number of Gaussian components $\mathcal{N}$ with c^k, μ^k and Σ^k denote the prior probabilities, mean values and the diagonally assumed covariance matrices, respectively. Because the GMM models densities with the same shape at different gain levels with separate components, the VTFs are mean normalized with the advantage of reducing model complexity and increasing robustness in the model learning. By this operation the gain information is lost and has to be recovered by a gain estimation proposed in the next section.

For the separation of the speech signals, we rely on the sparse nature of speech in its high-resolution time-frequency representation. This directly results in the MIXMAX formulation to separate sparsely distributed signals. The log-spectrum of a mixed signal $\log(Y)$ can be approximated by the element-wise maximum of the log-spectra of the constituent signals S_1 and S_2: $\log(Y) \approx \max[\log(S_1), \log(S_2)]$. This leads to the notion of the binary mask (BM). Per definition the binary mask of the speakers are complementary eg.: $\text{BM}_1 = \overline{\text{BM}_2}$. In the following, we will first construct speaker dependent models. Afterwards we use the MIXMAX approach to combine two models and construct a model which is approximating the mixture $\log_{10}(Y)$. Using an ML search the most likely Gaussian component representing $\log_{10}(Y)$ is selected. The BMs are applied to the mixed signal to finally obtain the estimates of the underlying speech signals, $\hat{s}_1$ and $\hat{s}_2$ respectively.

3.1.1. Gain Normalization and Speaker Dependent Model Adaptation

Each components mean value of the trained GMM can be thought of representing a prototype VTF. Using this knowledge we can formulate a speaker dependent (SD) GMM by incorporating E_i and g_i^k, the logarithmic gain factor, to the means as $\mu_{SD_i}^k = (\mu^k + \log(E_i^k) + g_i^k)$ resulting in SD models $\lambda_{SD_i}^{GMM}(c, \mu_{SD_i}, \Sigma)$ each consisting of $K_1 = K_2 = K$ components.

As the ML approach is prone to gain mismatches, the gain has been removed from the VTFs, hence the SD models still have to be gain adjusted. For this purpose we use the MIXMAX approach. The bin along frequency, comprising the maximum value of the underlying signal S_i can not exceed the value of the mixture Y at this frequency bin which is exactly the property we are exploiting for gain compensation. The left hand side of fig. 2 shows the mixture Y and the first components mean $\mu_{SD_1}^{k_1}$ of the first speaker without gain compensation and the right hand side shows the same with gain compensation. The spectrum is zoomed in to a frequency range of $0 - 4$ kHz.

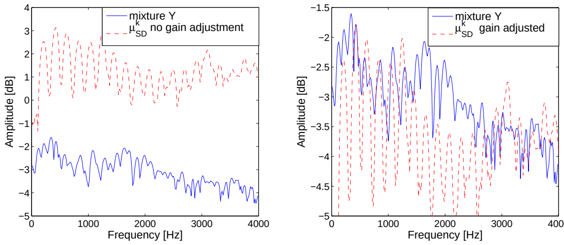


Fig. 2. Illustration of the gain adjustment method. The left hand side shows one segment of the mixture's spectrum in dB (blue solid) and the first component μ_{SD}^k of the adapted SD GMM without gain compensation (red dashed). The right hand side shows the same but with gain compensation.

The gain g_i^k is calculated for each component k and each time instant as follows: Calculate the difference of the maximum value in each components mean vector $(\mu_k + \log(E_i))$ and the value of the mixture Y at the same frequency bin.

For pitch adjustment the approximation of the second derivative over frequency of the mixture is investigated. The curvature at this frequency bin must exhibit a local maximum, hence the value must be negative. If the curvature at this frequency bin is not a local maximum the neighbors (i.e., one frequency bin before and afterwards) are considered and if a local maximum is found there, the spectrum of the mean μ_k is shifted by one bin to the local maximum subjectively resulting in better results. The gain g_i^k is recalculated and finally used to adjust the mean value of the considered component. The benefit of this method is the decoupled shape and gain of the SD GMM means and a reduction of complexity.

3.1.2. Model Construction and Source Separation

Having established the SD models the goal of separating the mixture Y can be achieved by applying the MIXMAX approach to all possible combinations of the respective model components. Consequently, we found the SD binary masks BM_1 and BM_2 for all $K_1 \cdot K_2$ combinations. The model $\lambda(\hat{c}^{k_1,k_2}, \hat{\mu}^{k_1,k_2}, \hat{\Sigma}^{k_1,k_2})$ with $K_1 \cdot K_2$ components is made up of the following parameters:

$$\begin{aligned} \hat{c}^{k_1,k_2} &= c^{k_1} \cdot c^{k_2} \\ \hat{\mu}^{k_1,k_2} &= \mu_{SD_1}^{k_1} \cdot \text{BM}_1 + \mu_{SD_2}^{k_2} \cdot \text{BM}_2 \\ \hat{\Sigma}^{k_1,k_2} &= \Sigma^{k_1} \cdot \text{BM}_1 + \Sigma^{k_2} \cdot \text{BM}_2 \end{aligned}$$

The objective of the maximum likelihood estimator is to find the component maximizing $p(\log_{10}(Y)|\lambda^{GMM})$, defined as:

$$\{k_1^\star, k_2^\star\}_{\text{ML}} = \\ \arg \max_{\{k_1,k_2\}} \Big[p\big(\log_{10}(Y)|\lambda^{GMM}(\hat{c}^{k_1,k_2}, \hat{\mu}^{k_1,k_2}, \hat{\Sigma}^{k_1,k_2})\big)\Big].$$

The estimated component indices $\{k_1^\star, k_2^\star\}_{\text{ML}}$ can be directly used to select the associated binary masks BM_i and apply them on the mixed signal to get the respective speakers spectrum.

3.2. Separation Using Non-Negative Matrix Factorization

For modeling the VTF distribution we have furthermore investigated NMF [8, 9] instead of GMMs. NMF approximates a non-negative matrix $V^{M\times N}$ by the product of two also non-negative matrices $W^{M\times R}$ and $A^{R\times N}$. In our case the VTFs correspond to V and the bases W are the quantity we are interested in, summarized as model λ^{NMF}. For factorization of V, in W and R, the method described in [8], minimizing a distance function akin to the Kullback-Leibler distance, is used. While in the training phase the bases W are estimated, in the separation phase the weights A are of interest. The weights are specifying the contribution of each bases for the approximation of the target signal V. Typically, in the separation step a union of all source dependent bases is constructed by combining them as $W = W_1 \cup W_2 \cup \ldots \cup W_N$. This results in an N-times larger bases set. Similarly to the ML-based case, we have to reintroduce E_i into the bases to be applicable for the SI source separation task. This is achieved by multiplying the spectrum of the excitation signal of the first and second speaker with the bases:

$$W_{\text{SD}} = W_1 \cup W_2 = W \cdot E_1 \cup W \cdot E_2.$$

In the last step, the spectrogram of the mixed signal Y is given as V the non-negative matrix to be approximated. Fixing the introduced bases W_{SD} and estimating the weights A best approximating Y, finally yields in the separation of the mixed signal. The reconstruction is done by first splitting up the bases matrix W_{SD} and the trained weight matrix A into the parts belonging to the corresponding sources and finally reconstructing the source signals as:

$$\hat{S}_i = W_i \cdot A_i \quad \text{with} \quad i \in \{1, 2\},$$

where $\hat{S}_i$ is the respective spectrum of a speakers segment.

4. EXPERIMENTS

To evaluate the proposed separation algorithms, the database recently provided by Cooke et al. [10] for the single channel speech separation task has been selected. As at this stage of our implementations no multi-pitch detection algorithm has been developed only data from the training corpus, for both the training and testing are used. The sampling frequency was resampled to 16 kHz for all files. For calculating the spectrogram the signal was cut into segments of 32 ms with time shifts of 10 ms. For pitch extraction *PRAAT* [11] has been used. To assess the performance of our algorithm we compared them to the F-HMM algorithm [1]. We used the LPC method of order 24 to separate the filter from the source signal and transformed these parameters to the log-frequency representation. For training the SI models for both algorithms 10 male (MA) and 10 female (FE) speakers each producing 30 sec. of speech are used. The label of the speakers are shown in table 1. For training the SD models of the F-HMM method, 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 2 were used for testing. For simplicity we will call these speakers FE1, FE2, MA1 and MA2 in the future.

	speaker									
FE	4	7	8	11	15	16	21	22	23	24
MA	3	5	6	9	10	12	13	14	17	19

Table 1. Label of female and male speakers used for training speaker independent models.

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 2. Labels of speakers and file names used for testing.

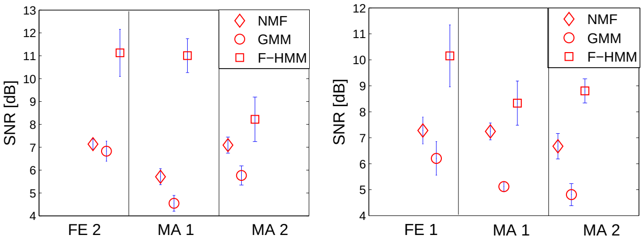
For testing all files are mixed at a level of 0 dB SNR and all possible combinations between target speakers and their interfering speakers are evaluated, resulting in altogether 108 mixed signals. Audio examples of the mixtures and the separated files are available at https://www.spsc.tugraz.at/people/michaelstark/SCSS.

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

$$\text{SNR}_j = \frac{\sum_{d,t} S_j^2(d,t)}{\sum_{d,t}(S_j(d,t) - \hat{S}_j(d,t))^2},$$

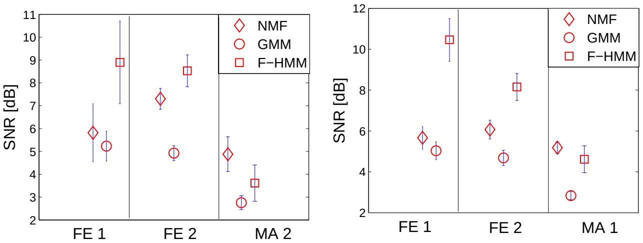
where $d = [1, \ldots, D]$ is the index of the frequency bin and S_j and $\hat{S}_j$ are the source and separated signal spectra of the considered speaker j.

For the NMF method we trained 200 bases for each gender, yielding an SI model of 400 bases. The GMMs were also trained in a gender dependent way with 64 components, which are combined with equal weighting yielding a model with 128 components altogether. The dimension of the model parameters corresponds to the number of frequency bins used in the spectrogram, which was 512. For training we used 200 iterations in the NMF case and 10 EM-steps in the GMM case. The F-HMM 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 BM, a beam search [12] restricting the search to the best 5000 candidates has been used. Figures 3 and 4 show the mean and the standard deviation of the SNR for each target speaker to its interfering speakers.



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

Fig. 3. Mean and standard deviation of the SNR from the target speakers to the interfering speakers. The shape of the markers identifies Methods.



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

Fig. 4. Mean and standard deviation of the SNR from the target speakers to the interfering speakers. The shape of the markers identifies Methods.

Investigating the results we can clearly observe the dominance of the F-HMM method. This is not surprisingly as this method on the one hand is speaker dependent and on the other hand has the highest model complexity. From a complexity point of view the SI NMF method uses a larger number of parameters than the GMM and one could argue that for this reason it achieved the second best performance in terms of SNR improvement. But on the other hand the NMF method uses the synthesized signals only and does not derive a binary mask applied on the mixed signal. We assume that the reason for the inferior performance of the ML-based method is its sensitivity to prototype envelope mismatches alleviated by the weighted sum of the NMF bases. Furthermore, the SNR of the NMF method is more constant over different interferences compared to the F-HMM hence the NMF method might be better suited for applications in adverse situations. Conducting the same experiments with SD trained models does not increase the performance significantly. This supports our assumptions to employ SI models. Finally, it should be noted that the phonetic content of some utterances was almost similar with only one word difference and that we investigated also mixtures Y with same genders in contrast to most other literature.

5. CONCLUSION AND OUTLOOK

We have presented two statistical approaches for modeling the VTFs, a maximum likelihood (ML) based one and one using non-negative matrix factorization, and we have used them for the source-filter based single channel speech separation task. The decomposition of the speech signals into a source and a filter component was motivated by the work of [4] yielding a speaker independent (SI) separation method. For the ML-based method we trained SI Gaussian mixture models (GMM) used as *a priori* knowledge. For this method we proposed a new algorithm for gain compensation based on the mixture-maximisation approach. NMF bases have been trained using the same data as the GMM method. The synthesis has been carried out using only the artificially generated signals and it shows superior performance compared to the GMM and lower variance across different mixed signals compared to the method of Roweis. Moreover, the proposed methods have the advantage of being less complex in terms of the used number of parameters and they are speaker independent. As next step, a multi-pitch tracking algorithm is going to be implemented and further refinements in the generation of the excitation signal and envelope modeling are going to be considered. Finally, listening tests will be carried out.

6. REFERENCES

[1] S. T. Roweis, “One microphone source separation,” in *Neural Information Proc. Sys., NIPS*, 2000, pp. 793–799.

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

[3] A. Hyvärinen, J. Karhunen, and W. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[4] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, “A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation,” *EURASIP Journal on Audio, Speech, and Music Proc.*, vol. 1, pp. 15, 2007.

[5] J. Laroche, Y. Stylianou, and E. Moulines, “Hns: Speech modification based on a harmonic+noise model,” in *IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., ICASSP*, 27-30 April 1993, vol. 2, pp. 550–553.

[6] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[7] A. Nadas, D. Nahamoo, and M. A. Picheny, “Speech recognition using noise-adaptive prototypes,” *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, vol. 37, no. 10, pp. 1495–1503, Oct. 1989.

[8] D. D. Lee and H. S. Seung, “Learning the parts of objects by nonnegative matrix factorization,” *Nature*, vol. (401):788, 1999.

[9] P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 15, no. 1, pp. 1–12, 2007.

[10] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” in *Journal of the Acoustical Society of America*, 2006, number 120, pp. 2421–2424.

[11] P. Boersma and D. Weenink, *Praat: doing phonetics by computer (Version 4.5.14) [Computer program]*, 2007.

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