A distance measure to combine monaural and binaural auditory cues for sound source segregation

Sarinah Sutojo, Steven van de Par, Joachim Thiemann

CvO University Oldenburg, Acoustics Group, Cluster of Excellence 'Hearing4all', sarinah.sutojo@uni-oldenburg.de

Introduction

The blind segregation of audio sources from a sound mixture is one of the main problems in computer-based analysis of audio signals. If several sources overlap in time and frequency it becomes particularly challenging to distinguish between them. One method of segregating such mixtures is the estimation of an ideal binary mask (IBM). Considering one of the present signals as the target signal, the IBM labels time-frequency units (t-f units) in which the target signal is dominant with a 1 while all other t-f units are set to 0. Assuming that the relevant information about the target is preserved in the parts where the target is dominant, missing data classifiers can be used to recognize speech or identify speakers.

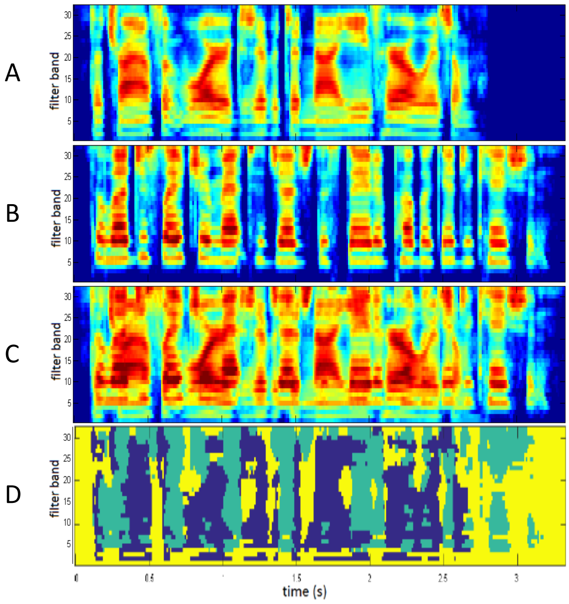

Figure 1: A typical mixture of speech signals. A: time-frequency representation of sentence 1 (”The triumphant warrior exhibited naive heroism.”), B: t-f representation of sentence 2 (”The legislature met to judge the state of public education.”), C: mixture of both sentences. D: Each t-f unit is labeled according to the dominant source (green and blue label the two speakers, yellow represents background noise).

In computational auditory scene analysis (CASA) the IBM is usually estimated by considering the prominent auditory features of the mixture and assigning t-f units to different sources, based on the observed features. The aim of the presented CASA system is to effectively combine periodicity and binaural cues for the estimation of the IBM and build a framework that allows us to integrate further feature types.

Different algorithms have been suggested that also rely on joint pitch and spatial cues. One example can be found in [5] where the pitch and azimuth of concurrent sources are tracked across time. Other systems can be found that have a similar structure but take different approaches on how the available features are analyzed and eventually combined to achieve the segregation [6].

System Overview

Since the human auditory system shows the ability to perform blind audio source segregation, it often serves as a model for the design of CASA systems. Especially principles from *auditory scene analysis* (ASA) [4] are considered in this context. There are two main processing steps in ASA which motivate the structure of the presented system, namely the segmentation and the grouping stage. Segmentation describes the decomposition of the incoming mixture into units that likely originate from the same source. The resulting segments are thereafter linked with each other if they belong to the same source (grouping).
Fig. 2 shows an overview of the proposed system. After a pre-processing stage which transfers the incoming time-domain signal into a time-frequency representation, the acoustic features are extracted for each of the resulting t-f units. Within the local clustering module, the similarities between the extracted periodicity features are analysed on a local level to form clusters of relatively homogenous acoustic attributes (comparable to the segmentation stage in ASA). The obtained clusters are passed on to the global grouping unit in which the prominent spatial cues of each cluster are used to link the segments that originate from similar directions. Like the above described grouping stage, the previously disconnected clusters are then assigned to a common source. In this way, an estimate of the IBM is made. The estimated masks are evaluated by comparing them with the masks that are derived from the sound mixture components prior to mixing.

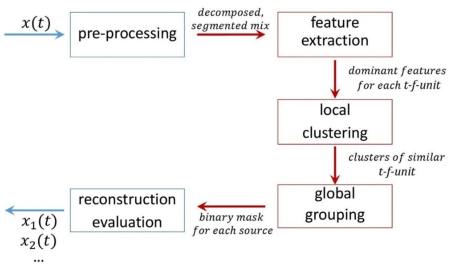
$x(t)$
pre-processing
decomposed, segmented mix
feature extraction
dominant features for each t-f-unit
local clustering
clusters of similar t-f-unit
global grouping
binary mask for each source
reconstruction evaluation
$x_1(t)$
$x_2(t)$
...

Figure 2: System diagram

Pre-Processing

The left and right ear signals are first passed through a gammatone filterbank with 32 filters. The center frequencies are spaced on the equivalent rectangular bandwidth (ERB) scale and range from 80 to 5000 Hz. Each filter output is then divided into time frames of 20 ms length with a 10 ms time shift. Thus, a cochleagram for each ear is created that represents the signal in an array of t-f units and which is the basis of the following feature extraction.

Feature Extraction

In every t-f unit a set of features is extracted which includes location, periodicity and energy. For the latter two, the sum of the t-f units in the right and left ear channel is used.

Periodicity Features

The periodicity is analysed with the *periodicity degree* (PD) as proposed in [2]. For every time frame j and filter band k, the *normalized autocorrelation* (NAC) and *comb filtering ratio* (CFR) are calculated for a range of period candidates p. For filter bands with center frequencies above 1.5 kHz the NAC and CFR are calculated from the envelopes of the filter outputs instead of using the outputs directly as it is done in the lower subbands. Both values are then combined using

$$PD\left(j,k,p\right) = max\left[0.01, NAC\left(j,k,p\right)\cdot CFR\left(j,k,p\right)\right]. \tag{1}$$

For every t-f bin a vector of PD values per period candidate is extracted. The originally described algorithm for pitch extraction averages the periodicity information across frequency bands within a time frame, this is omitted here because we want to make frequency specific estimates of the grouping of t-f units.

Spatial Cues

For the extraction of spatial cues, a probabilistic localization algorithm is used [1]. From each pair of t-f units for the left and right ears the *interaural time* and *level differences* (ITDs, ILDs) are calculated. Based on Gaussian mixture models that are trained with ITD and ILD data for each subband and a range of azimuths a, the likelihood $\mathcal{L}\left(j,k,a\right)$ of each t-f (j,k) unit originating from a certain location a is computed with

$$\mathcal{L}\left(j,k,a\right) = \log p\left(\overrightarrow{x_{j,k}}|\lambda_{k,\phi_a}\right), \tag{2}$$

where $\overrightarrow{x_{j,k}}$ represents the observed binaural feature vector consisting of ITDs and ILDs, and λ_{k,ϕ_a} the frequency- and azimuth dependent Gaussian mixture model. For every t-f bin a vector of likelihoods corresponding to a range of azimuth (-90°, - 85°, ... 90°) is available.

Local Clustering

The local clustering is based on each t-f units' information on periodicity and energy. Assuming that the PD between two neighbouring units varies little if they are both dominated by the same source, adjacent units are assigned to the same cluster if their PD vectors are highly correlated. However, if they are dominated by different sources the correlation should be significantly lower and a contour can be drawn between the units to mark the onset of a different source. An illustration of the method is given in Fig. 3.

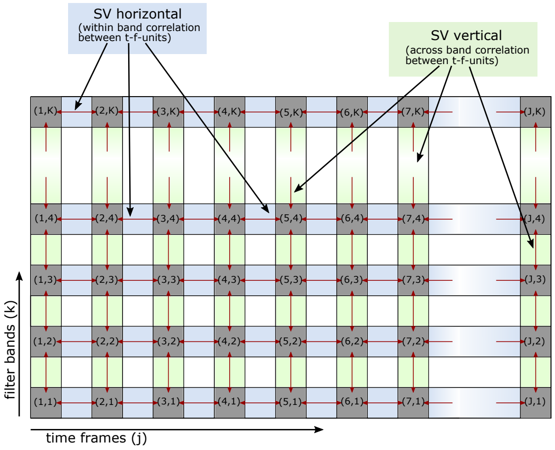

Figure 3: Similarity Values (SV) between t-f units are mapped to an expanded matrix.Thresholds applied to the across and within frequency band correlations determine if two adjacent units are grouped.

The vector of PD values in unit (j, k) is correlated to its 4 direct neighbours (across frequency band and across time) by using Pearson's correlation coefficient between the neighbouring PD vectors. The correlation coefficient gives a measure of the similarity between the neighbours. If it does not reach a given threshold, a 1 is noted between the neighbouring units to mark a contour line. After doing so for every transition between t-f units, the empty elements of the matrix (white squares in Fig 2.) are also set to 1 if two or more of the 4 adjacent values are set to 1. Thus, a map of contours is generated which enclose regions of similar or slowly changing PD values. T-f units with low energy produce rather unreliable PD values which results in low correlations between neighbours. Preferably, these units should be summarized in the same cluster since they most likely belong to background noise. To avoid this problem the low energy units are excluded from the periodicity based clustering and are summarized in regions that are enclosed by contour lines. To eventually create clusters, the obtained contour map is processed with a region growing algorithm [3] that fills an area within closed contour lines with the same integer.

Global Grouping

The previously formed clusters are grouped together based on their spatial properties. To achieve this, the vectors of azimuth dependent log-likelihoods for all t-f units within a contiguous cluster are summed. Purpose of this summation is to attain a more reliable spatial estimate as compared to the available information within a single t-f unit. Every cluster is then assigned the azimuth location which reveals the highest log-likelihood after the summation. To determine where the sources are located throughout the whole signal, the azimuth log-likelihoods are also summed over the entire time-frequency plane. The dominant azimuths are detected and clusters at these position (±15°) are grouped together. Fig 3 displays the results of the local clustering and global grouping for an exemplary mixture of two simultaneous speakers (male + female) originating from static positions.

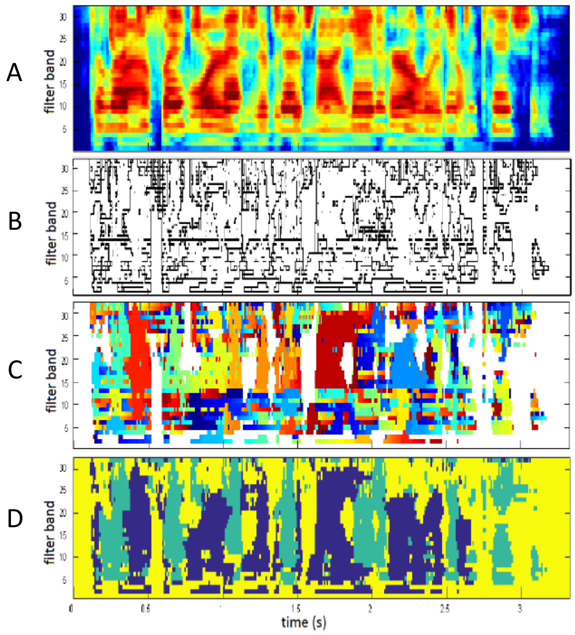

Figure 4: Estimation of the ideal mask from a two speaker mixture. A: Cochleagram of speech mixture, B:estimated contours, C:local clusters, D:localized clusters.

Evaluation

The presented algorithm was tested in 3 noise conditions and for 10 different speaker mixtures. Speech material was taken from the TIMIT Corpus and each stimulus consisted of a male and a female talker that were assigned to static positions at -70° and 60° on the horizontal plane. We render the sources at these positions using the HRTF database described in [7]. The two speakers were always assigned different sentences and presented simultaneously to create spectral and temporal overlap. Speech and transfer functions were adjusted to a sampling frequency of 16 kHz. Diffuse spatial noise was generated and mixed with the speaker mixtures at SNRs of 20,10 and 5 dB SNR. Each t-f unit was classified to one of three groups, either background noise, or one of the two dominant azimuth positions as detected over the entire time-frequency plane (assuming that the speakers were actually located close to these azimuths). The map of classified t-f units (bottom panel in Fig. 4) was compared with the ideal classification map, derived from the signal components prior to mixing (Fig.1). The estimated segregation was evaluated based on the percentage of correctly classified t-f units. The mean percentages of correctly identified t-f elements and standard deviations for the three noise conditions are displayed in Tab. 1.

Table 1: Percent correctly identified t-f units.

condition/dB SNR	mean ± std
20	(66,4 ± 8,5)%
10	(63,0 ± 8,8)%
5	(54,2 ± 7,5)%

For the evaluated two talker mixtures these preliminary results are promising. In the conditions with 20 and 10 dB SNR the algorithm was able to correctly classify more than 60% of the t-f units. One can see that the performance is significantly lower though when the SNR is decreased to 5 dB.

Conclusion

A clustering of neighbouring t-f units based on periodicity and energy properties is useful for estimating an ideal mask. Even though the feature extraction in a small t-f unit is likely to cause error-prone information it is still possible to successfully group units with pronounced periodicity based on next-neighbour similarity. By applying a localizer to these clusters, the disconnected segments could be grouped when originating from the same direction. This allows to effectively combine benefits of monaural and binaural cues. Through integration of more acoustic features and the use of optimization methods, the performance is supposed to be further improved.

References

[1] T. May, S. van de Par, A. Kohlrausch, (2012), “A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation,“ IEEE Transactions on Audio, Speech, and Language Proc., Vol. 20 (7), pp. 2016-2030)

[2] Z. Chen and V. Hohmann, (2015), “Online Monaural Speech Enhancement Based on Periodicity Analysis and A Priori SNR Estimation,“ IEEE Transactions on Audio, Speech, and Language Proc., Vol. 23 (11), pp. 1904-1916

[3] Gonzales, R., Woods, R., Eddins, S., “Digital image processing using MATLAB”, Pearsons Education, New Delhi (2009).

[4] Bregman, A.,“Auditory scene analysis: the perceptual organization of sound”, MIT Press, Cambridge Mass. (1994)

[5] J. Woodruff and D. L. Wang, (2013), “Binaural Detection, Localization, and Segregation in Reverberant Environments Based on Joint Pitch and Azimuth Cues“ IEEE Transactions on Audio, Speech, and Language Proc., Vol. 21 (4), pp. 806-815.

[6] A. Josupeit, N. Kopco, V. Hohmann, (2016), “Modeling of speech localization in a multi-talker mixture using periodicity and energy-based auditory features“ J.Acoust.Soc.Amer., Vol. 139 (5), pp. 2911-2923.

[7] Thiemann, J., Escher, A. and van de Par, S., “Multiple Model High-Spatial Resolution HRTF Measurements”, DAGA 2015.