Segmentation of Multitalker Mixtures Based on Local Feature Contrasts and Auditory Glimpses

Sutojo, Sarinah; May, Tobias; Van de Par, Steven

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

Link to article, DOI:
10.1109/TASLP.2022.3155285

Publication date:
2022

Document Version
Peer reviewed version

Citation (APA):
Sutojo, S., May, T., & Van de Par, S. (2022). Segmentation of Multitalker Mixtures Based on Local Feature Contrasts and Auditory Glimpses. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1249-1262. https://doi.org/10.1109/TASLP.2022.3155285
Segmentation of Multitalker Mixtures Based on Local Feature Contrasts and Auditory Glimpses

Sarinah Sutojo, Tobias May, and Steven van de Par

Abstract—The separation of unknown sources from an acoustic scene remains one of the main challenges in the analysis of audio signals. In this paper we address the problem of source separation in multitalker scenarios, attempting a speaker-independent segmentation of the scene, based on the analysis of relative changes in the grouping features, followed by an ideal binary mask (IBM) estimation. In the presented algorithm, periodicity, power, and spatial features are combined in a general framework to estimate the on- and offsets of acoustic sources. A deep neural network (DNN) is then applied to convert relative feature changes into spectro-temporal contrast maps which are used to extract auditory glimpses, i.e. spectro-temporal segments which are dominated by the same source. By exploiting relative feature changes instead of absolute source properties, the presented glimpse formation is well applicable to unseen acoustic conditions or speakers. Using retrospective labeling, the formed glimpses are then assigned to fixed azimuth locations, providing the final IBM estimate. A systematic evaluation with up to five competing speakers and interfering diffuse noise shows that the segmentation procedure generalizes well to unseen speakers as well as to untrained noise conditions and source numbers. A comparison between the final IBM estimates and the output of a baseline system, which performs the source segregation based on spatial cues only, shows that the contrast based segmentation utilizing multiple features and spectro-temporal context can improve the quality of the estimated IBMs, in particular in conditions with multiple interfering talkers.

Index Terms—computational auditory scene analysis, binaural speech segregation, multi-talker, deep learning

I. INTRODUCTION

EVERYDAY listening scenarios, such as noisy environments or settings with multiple talkers, pose the challenge to the listener to separate the different acoustic sources from the noisy mixture. While human listeners are remarkably good at listening to a particular target source even in highly complex acoustic scenes, machine-based solutions to accomplish the same task cannot yet achieve the performance and robustness of the human auditory system. Although the performance of technical systems has substantially increased, variations in the acoustic scene and specifically the generalization to a larger number of competing talkers, remains a significant challenge. However, a robust machine-based solution to the problem of automatic source separation would be of fundamental importance for a variety of applications, such as speech enhancement, automatic speech recognition (ASR) or speaker identification in complex acoustic settings ([1], [2], [3], [4], [5]).

One method to segregate single sources from a mixture is to estimate an ideal binary mask (IBM) which labels time-frequency (T-F) units that are dominated by the target source [6]. The reliable T-F units according to the IBM can be further used, for example, as input to a missing data classifier ([7], [8]) to perform speaker identification, or by directly applying the IBM to the mixture and removing all T-F units which are dominated by interfering sources [9]. IBMs have been shown to improve ASR performance under noisy conditions [10] and are applicable to a wide range of scenarios, independent of type or number of interfering sources, which makes them a suitable goal for source segregation. The application of binary masks to select T-F regions belonging to the target speaker creates artifacts which can be reduced with ideal ratio masks (IRMs) ([11], [12], [13], [5]). IRMs apply a similar principle as IBMs but are not limited to binary values. Instead, the IRM provides the ratio of target energy to interfering sources within each T-F unit. Nevertheless, when using IBMs or IRMs to disentangle sound mixtures, the main difficulty lies in the estimation of these masks, which means identifying T-F units that are dominated by the target source. Ideally, the IBM estimation makes the best use of all available features and uses information from a larger spectro-temporal context, while at the same time classifying small spectro-temporal units as belonging to either target or interfering source. A major challenge in the estimation, is to remain source independent, meaning that no oracle knowledge of the present sources is needed in order to perform the segregation, and even further, to find a general solution to the segregation problem which can work in acoustic conditions not seen during training.

Although a general source separation is the focus of this study, a number of deep neural network (DNN)-based approaches will be presented here, that focus on the enhancement of a single target speaker but perform source separation as a part of this task. A common strategy to estimate T-F units dominated by the target is by learning the characteristics of the considered source and to distinguish its components from the background. The speaker segregation system presented in [2] uses a DNN which is trained with two-talker mixtures to estimate the IRMs of a specific male target speaker. The algorithm can then estimate the IRM for the trained speaker given the two-talker mixture, which was shown to improve speech intelligibility for the specified speaker. However, this type of system is highly specialized since it was trained and tested with the same speaker. This means that the system will not generalize well to unseen talkers which makes this model-based approach less suitable for speaker-independent applications.

An alternative approach which achieves a higher degree of source independence is presented in [14] and uses discrimi-
native embeddings to separate mixed speakers. In this system, a DNN is trained to project each T-F unit into an embedding space, where elements belonging to the same source lie close together and can be clustered using a K-means algorithm, and eventually be used to construct a binary mask. In contrast to the previously described system [2], this method uses the distinction between active sources to perform the grouping, which makes it less dependent on knowing the characteristics of the target source. The system was further extended, as for example to incorporate spatial features [15] which improved the separation performance over the monaural baseline system, and has been combined with permutation invariant training [16] to further increase its source-independence. This combination has been successfully used, along with complementary spectral and spatial features, to perform multi-channel speaker separation in reverberant environments [17].

Besides these approaches, which strongly rely on deep learning methods, another type of source separation systems comes from the field of computational auditory scene analysis (CASA). Unlike the previously-described systems, which automatically learn characteristic features by training on spectral magnitudes or time signals, CASA-based systems are more explicit about the underlying grouping principles. Several CASA systems are motivated by principles of human auditory perception, such as studies on the effect of single features on auditory scene analysis in human listeners [18] that show how listeners exploit these features to distinguish sources.

One example of a CASA system is presented in [19] and combines pitch cues with an analysis of onsets and offsets to form segments of voiced and unvoiced speech components. These segments are combined to create a binary mask using a separate simultaneous and sequential grouping stage at different time scales. Although the system does not apply any pre-trained speaker models, it achieves comparable segregation performance to model-based systems for co-channel speech ([20], [21], [22]). Similarly, the system in [23] uses pitch combined with spatial features for mixtures of two and three speakers. Hidden Markov Models (HMMs) are applied to estimate the number of active sources taking the system a step further towards actual scene independence. A comparison between the estimated IBMs and a baseline system [24], which solely relies on spatial cues for the segregation, shows that adding pitch features improves the segregation, which shows that it is possible to benefit from the complementary nature of different grouping principles.

Previous work by [25], which will be used as the baseline system in this study, performs source separation based on common spatial location. This means that each T-F unit is assigned to a source, based on the estimated azimuth direction within the specified unit. An appealing aspect of this system is its relatively simple working principle and the fact that it only requires spatially separated sources to perform the segregation. One limitation is the sparseness of the estimated IBMs which can be caused by locally masked features in case of low SNRs, or ambivalent feature estimates if the sources are co-located. The integration of context information and other complementary features to overcome some of the aforementioned limitations is one goal of the current study.

Since the previously-discussed CASA systems are designed for specific features and not designed to integrate arbitrary feature types or exploit spectro-temporal context, one goal of this study was to develop a more general framework which can exploit a number of different features inspired by auditory grouping without changing the topology of the system. The grouping process in the presented algorithm is based on auditory features which can be extracted locally for individual T-F units. Information is combined across different feature domains and the near spectro-temporal context using a DNN which is trained to detect boundaries between different sources within the T-F plane. To maintain a high degree of source independence, the initial segmentation of the scene is based on source distinction instead of source recognition and forms spectro-temporal segments that are dominated by a single source. To achieve this, the absolute features are replaced by relative features which represent contrasts instead of absolute source characteristics, and are less dependent on the familiarity of the system with a specific sound source. During inference, this DNN is used to convert the feature observations into contrast values, which allows to identify regions of slowly varying or homogeneous features. Such regions can be considered as spectro-temporal glimpses [26] which are dominated by a single source and have been shown to be crucial elements for the human auditory system to cope with challenging acoustic conditions such as a social gathering with many talkers, which is commonly referred to as the cocktail party problem. Psychoacoustic studies such as [27] or [28] indicate that even in adversary listening conditions, human subjects extract and use such glimpses, suggesting that they convey essential information for e.g. speech intelligibility. The glimpses are extracted by clustering T-F units with similar acoustic properties that are spectro-temporally close to each other. Other than the approach of [14], which also uses discriminative properties to cluster parts of the spectrogram, the discrimination between sources is here performed in the spectro-temporal plane instead of an embedding space.

The spectro-temporal glimpses which are obtained with this rather general process can then be assigned to separate sources e.g. based on their spatial properties. The major advantage of assigning larger glimpses to a certain source, as compared to assigning individual T-F units, is the possibility to consider a larger spectro-temporal region for the feature estimation. When using information from an individual T-F unit only, the amount of available information is fundamentally limited, and thus, a limited IBM estimation performance can be expected. One limitation in [25] was the sparseness of the estimated IBMs in adverse acoustic conditions with low SNRs. Especially a larger number of interfering sources and increasing spectro-temporal overlap inherently cause uncertainty in the observable features. By averaging the evidence across a spectro-temporal glimpse, the amount of reliable information is increased, which is expected to improve the final IBM.

Furthermore, it is important to note that not all features within a single T-F unit are informative about which source is active. Whereas a local azimuth estimate allows a rough allocation of the T-F unit to a source, a local pitch estimate is not characteristic for a source itself as a voice changes its
pitch over time. Similarly, local spectral power will reveal little information about which source is active. The intermediate glimpse formation based on feature contrasts, nevertheless allows to make use of features as periodicity and spectral power in a single T-F unit before allocating it to a source.

The goal of the current study is to investigate, whether a general framework based on glimpse formation using multiple relative auditory features can improve an IBM estimation in comparison to a baseline system that utilizes spatial cues only [25]. Such a general framework can then be extended straightforwardly in the future with additional features. Furthermore, the present study focuses on the choice of features and methods used to extract the glimpses from a signal mixture. The two guiding questions with respect to the choice of methods and features are: 1) whether the contrast estimation benefits from feature combination and context integration, and 2) which out of three image segmentation approaches is most suitable for converting the estimated spectro-temporal contrasts into the best possible glimpses.

II. SYSTEM OVERVIEW

The presented algorithm estimates the IBM of a target speaker located at a fixed spatial position out of a mixture of up to 5 different speakers. To facilitate the IBM estimation, the system divides the input signal into a glimpse estimate based on contrasts that are derived from feature maps. An illustration of the main processing steps and intermediate representations is given in Fig. 1. The individual building blocks are described separately in the following subsections.

A. Auditory Front-End

The input to the system is a binaural mixture sampled at 16 kHz, which is first decomposed into 32 subbands using a Gammatone filterbank with center frequencies ranging from 200 Hz to 7 kHz with 1.25 filters per ERB (Equivalent Rectangular Bandwidth, [29]). In order to extract periodicity features, subbands are input to an inner-hair-cell model which consists of a half-wave rectification [30] and a 4th order lowpass Butterworth filter with a cutoff frequency of 400 Hz [31].

The periodicity within a T-F unit is estimated by calculating the normalized autocorrelation (NAC) of the hair-cell model output within individual T-F units according to Eq. (1).

\[
NAC(t, f, \tau) = \frac{\sum_{n=0}^{N-1-\tau} [x(t, f, n) \cdot x(t, f, n + \tau)]}{\sqrt{\sum_{n=0}^{N-1-\tau} x(t, f, n)^2} \sqrt{\sum_{n=0}^{N-1-\tau} x(t, f, n + \tau)^2}}
\]

where \(t, f\) and \(n\) denote time frame, subband index and sample. \(N\) indicates the frame duration in samples. The time lag \(\tau\) ranges from 40 to 267 samples which corresponds to pitch values between roughly 500 and 60 Hz. This produces a vector that indicates the periodicity strength at different time lags. The obtained autocorrelation is then divided by the autocorrelation of the window to remove the influence of the window on the autocorrelation function [32].

The logarithmic power within each unit \(E(t, f)\) is obtained by summing the squared absolute values of the waveform over the length of one frame and applying a logarithm.
The binaural localization model suggested in [25] was used to extract spatial features. The model extracts interaural time differences (ITDs) and interaural level differences (ILDs) for each T-F unit, and jointly analyzes them to estimate the azimuth of the acoustic source. ITDs and ILDs are directly calculated from the subband outputs without applying the haircell model.

A previously trained Gaussian Mixture Model is used to predict azimuth probabilities $P(\phi)$ for each ITD/ILD combination. In addition to the azimuth probability $P(\phi)$, the logarithmic azimuth probability $\log P(\phi)$ is regarded, as it shows enhanced peaks and can be more conveniently averaged over a larger spectro-temporal region than the probability itself.

### Contrast Maps

The segmentation algorithm is focused on detecting boundaries of glimpses for different acoustic sources as indicated by a change in the observable acoustic features. To capture these changes, the observed absolute features relating to pitch, log. power and spatial location, are converted into relative features which reflect the similarity of the acoustic properties in neighboring T-F units. The relative features are obtained by setting the absolute features of neighboring T-F units, both across time and across frequency, in relation to each other according to the calculations given in Tab. 1.

As displayed in Tab. 1 the NAC vector, representing pitch strength at various time lags, serves as basis for two different relative features. Shown first, is the correlation of the NACs in two neighboring T-F units (here denoted as $NAC(\tau)_x$ and $NAC(\tau)_y$). The correlation coefficient is then used as the relative feature which determines how similar the T-F units are with respect to periodicity. Secondly, the mean of the two maximum values is used to indicate pitch salience. The log. power ($E_x$ and $E_y$ in Tab. 1) for two neighboring T-F units $x$ and $y$ is used to derive contrasts in log. power which reflect energetic on- or off-sets and the absolute sum of both values, indicating spectral features. The vectors of azimuth probabilities of two neighboring T-F units ($P(\phi)_x$ and $P(\phi)_y$ in Tab. 1) are correlated to reflect the spatial proximity. This is done for the azimuth probabilities as well as the logarithmic azimuth probabilities. In Fig. 1, three of the six relative features, log. power deltas, mean maximum pitch and azimuth probability correlation, are shown after transforming them into feature-specific contrast maps.

### D. Combined Contrast

In order to exploit information across different feature domains as well as the near spectro-temporal context, the feature-specific contrast maps are fed into a DNN classifier which estimates a combined contrast map. To combine the different feature domains, the DNN is trained to approximate the expected feature contrast based on ground truth knowledge of changes in the local signal-to-noise ratio (SNR) in each T-F unit, which correspond to changes of the dominant source. The local SNR values for obtaining the ground truth are calculated based on 20 ms time frames with a step size of 10 ms. For each transition between two neighboring T-F units $x$ and $y$, the ideal contrast label $P_{\text{contrast}}$ is calculated according to Eq. (2):

$$P_{\text{contrast}} = 1 - \sum_s p_s(x) \cdot p_s(y)$$

and serves as training target for the DNN. $P_{\text{contrast}}$ ranges from 0 to 1 and is derived from a heuristically determined probability function $p_s(k)$ of the source $s$ being dominant in the T-F unit $k$:

$$p_s(k) = \frac{w_s(k)}{\sum_s w_s(k)}$$

$$w_s = \frac{1}{1 + \exp(-SNR_s)}$$

$SNR_s$ signifies the SNR of the source $s$ in relation to all other present sources within $k$. If $p_s(k)$ of the same source $s$...
is high in both T-F units, the $P_{\text{contrast}}$ value is low, hence, only small feature changes are expected. In the case of two different sources being dominant in the two units, $P_{\text{contrast}}$ increases.

Besides combining different feature domains, the DNN also allows to integrate spectro-temporal context by including neighboring time frames and frequency channels. The input vector consists of the relative features determining the size of the input layer to be equal to the number of trained features. For the integration of spectral context, the feature vectors of the 32 available subbands are concatenated, which means that the size of the input layer is multiplied by 32. In the same manner, the integration of temporal context is achieved by concatenating the feature vectors of the regarded time frames, multiplying the size of the input vector by the number of integrated frames.

The DNNs used in this framework are feed-forward neural networks with 2 hidden layers, which are of the same size as the input layer. For the hidden layers, a ReLU (rectified linear unit) activation function is used and a sigmoid function in the output layer. To increase the generalization capabilities of the DNN, dropout was used during training [33]. Specifically, a dropout rate of 0.2 was applied to the hidden layers which means that 20% of neurons in the hidden layers are ignored during each update cycle. Before activation, batch normalization was applied to the hidden layers [34]. The Adam optimizer with a learning rate of 0.001 was used for training [35]. Since the distribution of the training data is highly imbalanced and contains predominantly low contrasts, the loss function was modified to emphasize errors in the estimation of high contrast values. The modified loss function $L$ is calculated according to:

$$L = \frac{0.5}{M} \sum_{m=1}^{M} (P_{\text{contrast}}(m)^3 - \text{DNN}(m)^3)^2$$

with $M$ being the number of datapoints, DNN$(m)$ being the predicted contrast value and $P_{\text{contrast}}(m)$ being the respective label for datapoint $m$. By setting $P_{\text{contrast}}(m)$ and DNN$(m)$ to the power of 3 before calculating the loss, the training becomes more sensitive to prediction errors for high contrast values ($P_{\text{contrast}}(m)$ close to 1), as compared to low contrasts. Spectral context is added by jointly training all subbands, whereby 2 DNNs are trained separately, one for transitions across subbands and one for within subband transitions. Temporal context is integrated by stacking the feature vectors of neighboring time frames (e.g. 2 frames in the past and in the future). Prior to training, the input features are normalized separately for each file using histogram equalization [36].

The speech material used for training is taken from the TIMIT database [37] and contains roughly 3 hours of speech consisting of 2, 3, and 4 speaker mixtures. A total of 22 different speakers was used for training, out of which 8 were female and 14 were male speakers. The mixtures are produced by mixing between two and four different sentences spoken by individual speakers at a long-term SNR of 0dB, resulting in all speakers being equal in level. Diffuse pink noise is added to the speaker mixtures at an SNR of -3dB, 0dB, or +3dB as compared to a single speaker. The diffuse noise was generated by convolving and adding different realizations of a monaural pink noise signal with head-related-transfer functions (HRTFs) corresponding to azimuth position ranging from $-180^\circ$ to $+180^\circ$ (in $5^\circ$ steps) in the frontal horizontal plane. Male and female speakers are mixed in various constellations, with either one or two speakers of the same gender being simultaneously active. All sentences are rendered with HRTFs to imprint spatial properties on the mixtures [38]. Different spatial conditions are trained with separation angles between the speakers ranging from $0^\circ$ to $50^\circ$ azimuth (all speakers have equal spatial distance to each other in the frontal horizontal plane).

An example of the contrast estimation for a mixture consisting of three competing speakers in the presence of diffuse...
pink noise at an SNR of -3dB is shown in Fig. 2. The separate speakers in a) are convolved with HRTFs to assign them to different locations in the azimuth plane and are then mixed with diffuse pink noise. From the mixture, a map of local contrasts is estimated as represented in b). Shown are three alternatives. The first panel from the top shows a contrast map based on a single relative feature, which is here the correlated azimuth probability, without any further training. It can be observed that the map is fairly noisy and little structured, however, regions of high contrast are identifiable. The second alternative shows the estimate after integrating spectral and temporal context into the training. The third option does not only consider spatial features but includes the two periodicity and log. power based features as well as spectro-temporal context. From top to bottom the amount of integrated information increases, visibly improving the reliability and structure of the contrast estimate. For comparison, c) shows the corresponding ground truth representations. Displayed on the top panel are the ideal binary masks for the 4 sources (3 speakers and noise) which are obtained after averaging each separate source signal across both ear channels and determining the dominant source in each T-F unit. Shown below are the ideal contrast values that are derived from the ground truth local SNR values of each source according to Eq.(2)-(4).

E. Glimpse Estimate

The combined contrast map is used to derive the glimpse estimate (i.e. glimpse boundaries) by applying one of three image segmentation methods, which identify regions with homogeneous features that are enclosed by high feature contrasts. These enclosed regions are representing spectro-temporal glimpses that are not yet assigned to a specific source. A straightforward method to convert the continuous contrast values into boundaries, is to apply a threshold to the contrast map and consider each value that exceeds this threshold as boundary. Afterwards, the boundaries are filled with a regiongrow algorithm [39] to obtain the final glimpse estimate. Although this regiongrow method is fairly simple, it is expected to yield good results if the estimated contrast value is reliable and unambiguous. In case of blurred areas within the contrast map however, the regiongrow method may cause the glimpses to either leak and grow over a large area of the spectro-temporal plane, or to become rather small when using a more conservative criterion. Another method, referred to as graph-based superpixels, is described by [40] and comes from the field of image segmentation. It takes into account the homogeneity within a certain spectro-temporal region by using the averaged contrast within an existing cluster of pixels (i.e. a superpixel) as a criterion for the integration of neighboring pixels. This criterion becomes more conservative with the increasing size of the regarded superpixel, preventing it from extensive growth. At the same time, it allows larger glimpses to grow in areas with an overall high average contrast. By varying the parameter $\tau$, the tolerance of the superpixel growth can be controlled which allows influencing the degree of over- or undersegmentation.

The third method combines the simple regiongrow algorithm with a process similar to a dilation & erosion approach [39]. The contrast map can be regarded as a grid of pixels, in which each T-F unit (represented by one pixel) is separated from the neighboring T-F unit by a pixel which contains the estimated contrast value. This map is first converted into a binary map, based on thresholding. The boundary elements of the binary map are dilated by 2 pixels (in x- and y- direction) to close potential gaps in glimpse enclosures. Afterwards, regiongrow is used to form glimpses in the remaining open regions of the map. The boundaries are thereafter eroded by 2 pixels, followed by dilating the previously formed glimpses by 2 pixels. Since a portion of pixels remain unassigned to any glimpse, regiongrow is used in a final step to form glimpses in the empty spaces of the boundary map.

F. IBM Estimate

To derive the final IBM estimates that are source specific, the previously obtained glimpses are assigned to the spatial locations of active sources. In order to determine these locations, the azimuth estimates are integrated across the entire T-F plane. Specifically, a histogram is created based on the highest averaged logarithmic azimuth probability for each time frame. The most prominent peaks in the histogram indicate the azimuths of active sources which the estimated glimpses are then assigned to. At this point, the number of active sources in the mixture is assumed to be known a priori. The peaks in the histogram are identified and according to the number $N$ of direction sources (speakers), the highest $N$ peaks are set as source locations. One further source is assumed to be background noise with no specified location. For each glimpse, the most probable azimuth is obtained by averaging all logarithmic azimuth probabilities of the T-F units within the glimpse. If the averaged azimuth value agrees with any of the assumed source locations, the glimpse is assigned to the respective source. Any glimpse that cannot be assigned to a peak, is classified as background noise.

An example of an estimated IBM is shown in Fig. 7 panel (f), along with the output of the baseline system (panel (d)), in which the T-F units are assigned to the azimuth peaks based on the most probable azimuth value within individual T-F units. (Code is available online\(^1\).)

III. TEST PROCEDURE

In addition to the final output of the system, which is an IBM estimate, two intermediate representations of the system are evaluated. These two intermediate steps are the contrast map and the estimated glimpse map prior to the labeling stage.

A. Evaluation Methods

To evaluate the estimated contrast maps, receiver operating characteristics (ROCs) were used, which are a common tool to assess classification performance. The estimated contrasts are compared to the ideal boundaries by quantifying true-positive and false-positive rates as a function of the decision criterion.

\(^1\)Source code: https://github.com/SSutojo/spatial-glimpse-estimation
applied to the estimated contrasts. The true-positive and false-positive rates are displayed in a ROC-curve. If the estimated contrasts are similar to the ideal boundaries, the area under the ROC-curve (AROC) is large, indicating the suitability of the regarded feature for a correct classification. The ROC analysis will be used to assess the contribution of individual relative features with and without DNN training, as well as various feature combinations and degrees of context integration.

The second intermediate representation that will be evaluated is the glimpse map. One reason to assess the correctness of the glimpse map prior to the final stage is because errors could be introduced during labeling. The guiding question for this evaluation is, which method and threshold allow the best IBM estimate, given an ideal labeling stage. Therefore, the extracted glimpses are assigned to active sources based on oracle information. The ground truth labels of all T-F units within one glimpse are considered and the predominant label is used to assign the entire glimpse to the respective source. The metric used for this assessment is the labeled accuracy $ACC_l$. It is calculated by comparing the T-F units of the ideal source map and the matched glimpses (Eq. (6)).

The estimated glimpse map is considered as a segmentation $U = \{U_k\}_{k=1}^K$ of the image $I$ consisting of $N$ pixels. $U$ is the array of estimated segments (i.e. glimpses) and $K$ the number of estimated segments. $T = \{T_s\}_{s=1}^S$ signifies the ideal source map (i.e. every T-F unit is labeled according to the dominant source) and $S$ the number of sources.

$$ACC_l = \frac{1}{N} \sum_k \max_{T_s} \{|U_k \cap T_s|\}$$  \hspace{1cm} (6)

One property of $ACC_l$ is that it reaches its maximum value 1 in the case of complete oversegmentation. Therefore, the three methods for glimpse formation are applied for a range of parameters that result in different degrees of over- i.e. oversegmentation. Potentially, oversegmentation errors at this stage can be repaired in the final glimpse labeling whereas undersegmentation errors (e.g. the connection of two glimpses that should ideally belong to different sources) will affect the final output (IBM estimate) of the system.

Eventually, the estimated IBMs are evaluated by first matching each estimated IBM to the ground truth IBM of the source with which it has the largest overlap. This yields the estimated source map $G = \{G_l\}_{l=1}^L$ with $L$ being the number of estimated sources. The percentage of correctly identified T-F units $%_{corr}$ is then calculated by comparing the ideal source map $T = \{T_s\}_{s=1}^S$ to the estimated source map $G = \{G_l\}_{l=1}^L$. Assuming oracle knowledge about the number of active sources, $L = S$ in this calculation, which gives $T = \{T_s\}_{s=1}^S$.

$$%_{corr} = \frac{\sum_s |T_s \cap G_s|}{N} \times 100$$  \hspace{1cm} (7)

IV. RESULTS AND DISCUSSION

A. Contrast Features

The three absolute features are extracted on different time scales to analyze the influence of temporal resolution on the effectiveness of the different contrast features. For this purpose, the three relative features; azimuth correlation, pitch correlation, and log. power deltas are evaluated for a range of different frame durations. Regardless of the frame duration, the stepsize between consecutive frames was kept constant at 10 ms ensuring alignment of the frames in the following stages. To guarantee a minimum overlap of 50% of the total frame duration, the shortest evaluated frame duration in this study was 20 ms which was also used for the calculation of the ideal boundaries.

Speech material for the evaluation were 60 sentence mixtures with one third of them being 2, 3, and 4 speaker mixtures. All speakers are present at equal levels, with a SNR of 0dB as compared to each other, and with diffuse pink noise at -3dB, 0dB, and +3dB SNR as compared to a single speaker. If the goal is to extract a single target speaker from the mixture, the signal-to-interference ratio (SIR) can be calculated by summing the remaining speakers and noise. For example, in a 4 speaker mixture, with the diffuse noise being presented at an SNR of 0dB, the SIR lies at -6dB. The separation angle between speakers was kept constant at 40° azimuth. The sentence material is taken from the TIMIT database but was not used during training. All sentence mixtures are cut to silent time frames, resulting in average lengths between 2 and 3 seconds per recording.

Results of the ROC analysis for all three feature types can be seen in Fig. 3. Displayed are the mean AROC values, averaged over speaker numbers and SNRs along with standard errors. The evaluation reveals that a 20 ms time frame is best suited for the extraction of the spatial cues, while pitch and log. power features show better performance for 30 ms frames. Based on this evaluation, the frame durations for the following DNN training with feature combination and context integration were determined at 20 ms for spatial cues and 30 ms for pitch and log. power cues. To observe the effect of frame duration after DNN training, the model which combines all
feature types and context information was once trained with features being extracted from different frame durations (20 ms and 30 ms), and once trained with all features being extracted from 20 ms time frames. In Tab. II the DNN trained on 20 ms frames only is marked with the note (20 ms frames) and yields worse results than the DNN which is trained with 20 ms and 30 ms frames.

B. Context and Feature Integration

To investigate the benefit of context and feature integration for the contrast estimation, a set of DNNs was trained as described in Sec. II-D. All trained and evaluated DNNs are listed in Tab. II with the model specifications and results of the ROC analysis. The evaluations shown in Tab. II and Fig. 4 were performed for a set of 60 sentence mixtures out of which a third are 2, 3, and 4 speaker mixtures. Each of the mixtures is presented in three different spatial conditions and three different noise conditions. The spatial conditions were defined as 5°, 20°, and 50° azimuth separation angle. Diffuse pink noise was added at SNRs of -3dB, 0dB, and +3dB compared to a single speaker. Four different constellations of context integration are considered in this evaluation. As condition without any context integration, a subband specific DNN is trained for each subband and separately for across and within subband transitions. For the considered 32 channel gammatone filterbank, this results in a set of 63 DNNs (32 models covering all within subband transitions plus 31 models for across subband transitions). During inference, these subband models are applied independently to each subband (subband only). The next stage is integrating spectral context by replacing the set of 32 (or 31) independent models by a single broadband model which jointly analyzes all frequency channels at one time step (broadband only). The broadband model is further extended with temporal context by including either one (spectral+1frame) or two (spectral+2frames) neighboring frames before and after the regarded time frame. The four levels of spectro-temporal context integration are shown in Fig. 4. While the leftmost condition (untrained) displays the AROC values after estimating the contrast by using the relative features without any further training, the rightmost condition represents the highest degree of spectral and temporal context integration.

The only two features which benefit from training a subband only model as compared to the untrained condition, are log. power sum and (to a lesser degree) pitch salience. Other than the other four relative features for which we expect that, e.g. a low correlation coefficient coincides with a high contrast value (similarly for a large log. power difference), the conversion of log. power sum and pitch salience into high or low contrast values is more complex. With regard to the log. power sum, for example, it could be expected that both a markedly high and a markedly low sum indicates that the same source is present in both T-F units, which would suggest a low contrast. High contrasts, however, would probably be indicated by medium log. power sums, since in most cases only one of the two T-F units is dominated by the regarded source. By training a subband DNN, the observed feature can be mapped to the expected contrast value which produces a performance gain for log. power sum and pitch salience. This improvement can not be observed for the other four features whose mapping to the contrast scale is more straightforward.

A more pronounced performance step for all features can be observed between the subband only and the broadband only models. Clearly, the spectral context boosts the distinctiveness and reliability of all regarded features, probably due to the fact that certain acoustic attributes such as timbre are only observable when regarding the other subbands as well. For the six features, the AROC increase, ranges between 0.0514 for log. azimuth correlation and 0.1009 for the log. power sum.

A slightly smaller AROC increase can be observed after including one frame temporal context in the future and the past. Here, the gain ranges between 0.0189 for azimuth correlation and 0.0598 for the log. power sum. This gain is decreased overall when including a second frame before and after the regarded time step. Pitch salience shows the largest benefit with 0.0229 AROC increase and the smallest gain is again observed for the log. azimuth correlation (0.0069). Furthermore, the inclusion of temporal context has a varying effect on the different feature types. While log. power sum and pitch salience draw a comparatively large advantage from including the two frames temporal context (0.0759 and 0.0675), the benefit is rather moderate for the two models trained with spatial cues (0.0271 for azimuth correlation and 0.0298 for log. azimuth correlation).

Other than the integration of spectro-temporal context, Fig. 4 also displays a DNN which combines all six relative features. Since the feature combination is not possible without the application of a previously trained model, there is no untrained condition for this setting. The benefit from feature integration is striking, although the subband version of this combined model is slightly weaker than the single feature broadband DNNs of spatial and log. power features, suggesting

![Fig. 4: Influence of spectro-temporal context and feature combination on contrast estimates. Shown are mean values and 95% confidence intervals](image-url)
that context integration is of similar importance as feature combination. However, the broadband-combined model lies clearly above the single feature models for all other settings, with an AROC gain between 0.0605 (for broadband with 2 time frames context) and 0.0712 (for broadband without temporal context). Overall, the largest benefits originate from the combination of different feature types and the integration of spectral context. To a lesser degree, temporal context also improves the contrast estimation.

A choice of other feature combinations was evaluated, using 2, 3, 4, and 5 of the available features to train broadband DNNs. The mean values and standard deviations of the ROC results are displayed in Tab. II, along with the results plotted in Fig. 4. The broadband DNN which combines 5 of the 6 available relative features, along with 2 frames temporal context shows the overall best performance and is therefore marked in bold font. This observation is not entirely unexpected, since the 6th relative feature is the logarithmic azimuth probability, which does not add new information considering that the azimuth probability is already included in the 5-feature DNN. However, it also suggests that adding redundant features to the training may have a negative effect on the model performance, although the AROC difference between the 5- and 6-feature DNN is minor (0.0031). For the following evaluation in unknown scenarios, as well as glimpse formation and final IBM estimation, the 5-feature model is used to obtain the contrast maps.

To observe how well the models generalize to conditions that have not been seen during training, a choice of four models was tested on an extended evaluation set. In addition to the tested conditions described above, the evaluation set includes -6dB and +6dB SNR noise conditions, as well as

| trained features | context dimensionality | results |
|------------------|------------------------|---------|
|                  | spectral | temporal | size input layer | size output layer | size hidden layers | mean AROC | std |
| pitch correlation | none | broadband | 0 | 32 | 32 | 2 | 0.5653 | 0.0342 |
| pitch correlation | broadband | 1 | 96 | 32 | 2 | 0.6536 | 0.0304 |
| pitch correlation | broadband | 2 | 160 | 32 | 2 | 0.6898 | 0.0314 |
| pitch salience | none | broadband | 0 | 32 | 32 | 2 | 0.6888 | 0.0303 |
| pitch salience | broadband | 1 | 96 | 32 | 2 | 0.7097 | 0.0227 |
| pitch salience | broadband | 2 | 160 | 32 | 2 | 0.7326 | 0.0232 |
| log. power delta | none | broadband | 0 | 32 | 32 | 2 | 0.5957 | 0.0173 |
| log. power delta | broadband | 1 | 96 | 32 | 2 | 0.6888 | 0.0220 |
| log. power delta | broadband | 2 | 160 | 32 | 2 | 0.7118 | 0.0232 |
| log. power delta | broadband | 3 | 160 | 32 | 2 | 0.7244 | 0.0238 |
| log. power delta | broadband | 4 | 160 | 32 | 2 | 0.6153 | 0.0605 |
| log. power delta | broadband | 5 | 160 | 32 | 2 | 0.7695 | 0.0407 |
| log. power delta | broadband | 6 | 160 | 32 | 2 | 0.7760 | 0.0293 |
| log. power delta | broadband | 7 | 160 | 32 | 2 | 0.7884 | 0.0340 |
| log. power delta | broadband | 8 | 160 | 32 | 2 | 0.7921 | 0.0313 |
| log. power delta | broadband | 9 | 160 | 32 | 2 | 0.7960 | 0.0322 |
| log. power delta | broadband | 10 | 160 | 32 | 2 | 0.7598 | 0.0316 |
| log. power delta | broadband | 11 | 160 | 32 | 2 | 0.7872 | 0.0343 |
| log. power delta | broadband | 12 | 160 | 32 | 2 | 0.7896 | 0.0343 |

TABLE II: Model specifications and results of the ROC-analysis for the evaluation of context influence and feature combination.
a single speaker and a 5 speaker setting. Furthermore, the models are evaluated in a case with 0° azimuth separation angle, which was seen during training but represents a highly challenging condition since a spatial separation of the speakers is made impossible.

The four tested models are the 5-feature model which showed the best overall performance and three DNNs which are each trained with the two relative features derived from the same absolute feature, and including spectral context and 2 frames temporal context. For example, the pitch model is trained with both pitch salience and pitch correlation.

In Fig. 5 the mean AROC values and standard errors for the three 2-feature models and the 5-feature model are displayed in dependence of SNR, speaker number and separation angle between the speakers. The untrained conditions are shaded in gray and the collocated condition in hatched lines.

Separation angle, as expected, has no influence on log. power and pitch features. It has a slight effect on the combined model, and a more significant impact on the azimuth-only model. Although the condition with collocated speakers does not allow a spatial separation of the speakers, it still allows a spatial separation of speakers against the diffuse background noise which is the reason for the above chance-level performance of the azimuth-only model in this condition.

SNR has a strong effect on the pitch-only model and all models, except for the azimuth-only model, suffer from the untrained condition. With increasing SNR, the azimuth-only model shows worsening performance which seems counter intuitive. However, this effect was also observed during other parts of the study and may be explained with the diffuse background noise being comparatively easy to distinguish from the spatially defined speakers. Consequently, an increase in the noise energy can improve the AROC value. For log. power, the same reason may be valid since the background noise is stable in power and creates a source that is relatively easy to identify.

The effect of speaker number is according to the expectations, since a larger number of speakers decreases the performance while fewer speakers are easier to distinguish. This trend is continued for the single speaker conditions, as well as the 5 speaker condition.

In all conditions, the combined model performs best. Furthermore, it can be observed that its performance in untrained conditions does not drop by more than 4% compared to the neighboring trained condition, indicating an increased robustness due to the feature combination.

C. Glimpse Formation

The three image segmentation methods can each be set to create either fewer large glimpses, or plenty of small glimpses, e.g. by varying the contrast threshold. At this stage, the glimpses are still unlabeled, which means, they are not yet assigned to a specific source. Consequently, plenty of small glimpses that consist of a single T-F unit each, can theoretically be labeled 100% correct, given an optimal glimpse labeling stage. However, the benefit of the glimpse formation stage stems from the ability to average features over an extended spectro-temporal context which requires segments to be larger than single T-F units. To assess the potential accuracy in terms of overlap with the IBM, whilst retaining as large glimpses as possible, all three methods are applied with varying parameters to create glimpse maps with different degrees of oversegmentation. Ideally, a high accuracy can be achieved with as little oversegmentation as possible.

Fig. 6 shows this process for an example sentence. Panel (a) displays the initial contrast map, which is supposed to be segmented into a glimpse map that approximates the ideal glimpse map as shown in panel (c). For the regiongrow method, the contrast threshold can be varied to create different amounts of glimpses as displayed in panels (b) and (d). Displayed in panel (b) is the resulting glimpse map when using a relatively low threshold (0.14). This creates more boundaries, and therewith, a larger amount of small glimpses as compared to a high threshold as shown in (c), resulting in larger and fewer glimpses. The dilation & erosion method also varies the contrast threshold
to create different sizes (i.e., numbers) of glimpses. For graph-based superpixels, the glimpse number can be controlled using the parameter $\tau$.

The right panel in Fig. 6 shows the labeled accuracy for this example sentence in dependence of the estimated glimpse number. The results suggest that the dilation & erosion approach allows for the best accuracy while causing comparatively little oversegmentation and is therefore expected to generate the best IBM estimates.

**D. IBM Estimation**

The final IBM estimation (glimpse labeling) stage is illustrated in Fig. 7 using the same example mixture as above (Fig. 6). Panels (a) and (c) illustrate the active source localization which is based on the prominent logarithmic azimuth probabilities for all available time frames. The ground truth locations of the three speakers in the mixture are -90°, -50°, and -10° azimuth and seem to be well approximated in the estimation, as the histogram in panel (c) indicates. Panel (d) illustrates the output of the baseline system in which each T-F unit is assigned to one of the peaks in the histogram, granting a tolerance radius of 3° azimuth. For the regarded three speaker mixture with separation angles of 40° azimuth, the estimated IBM is fairly sparse. For comparison, the IBMs of all three speakers and background noise are shown in panel (b). The output of the current system is depicted in panel (f), along with the glimpse map, obtained from the previously described contrast estimation and the dilation & erosion segmentation method (panel(e)). Apparently, the glimpse formation does not improve the labeling of T-F units in all cases, as in some regions a whole glimpse is labeled wrongly, causing a larger error than the false labeling of a single T-F unit. Nevertheless, the overall similarity to the true IBMs seems improved as compared to the output of the baseline system.

Figure 8 shows the percentage of correctly identified T-F units after labeling the estimated glimpses, in comparison to the baseline system. The parameters to control the degree of oversegmentation for the three image segmentation methods were set to fixed values (contrast threshold for regiongrow 0.14, contrast threshold for dilation & erosion 0.1, and graph-based superpixels $\tau = 0.3$). The evaluation set consists again of 60 mixtures with one third 2, 3, and 4 speaker mixtures, three different spatial conditions (5°, 20°, and 50° separation angles), and three noise conditions (-3dB, 0dB, and +3dB SNR). Displayed is the correct percentage in dependence of speaker number which has a striking but relatively consistent impact on the performance on all three methods. According to this evaluation, all three methods which make use of the previously formed glimpses yield better results than the
than global feature values is the possibility to exploit features
benefit of exploiting feature contrasts in the first stage rather
and should generalize better to unseen conditions. Another
stage lies in replacing the absolute features with relative
scene.

The three different glimpsing methods are compared to a
baseline method that relies on spatial cues only. Shown are
mean values and 95% confidence intervals

glimpse boundaries using ideal contrasts as training targets
in the estimated IBMs shows a clear advantage from the
same source is lower compared to a two-talker mixture.

The addition of spectral context has been shown to generate
a strong advantage for all feature types while temporal context,
added a smaller overall performance gain. This may be at-
tributed to the fact that the relative features inherently capture
the temporal progression of features in two consecutive time
frames. Partly, the advantage of temporal context may also
depend on the acoustic scene, specifically the rate at which the
dominant source switches. Mixtures of up to 5 simultaneous
speakers in noise were evaluated in this study which means
that the dominant source switches at a relatively high rate and
the chance of consecutive time frames being dominated by the
same source is lower compared to a two-talker mixture.

The final evaluation of percent correctly labeled T-F units
in the estimated IBMs shows a clear advantage from the
glimpse formation as compared to using spatial cues only.
This indicates that the segments based on contrast features
are useful to obtain a more reliable assignment of the T-F
plane to specific sources as compared to estimating IBMs on
the basis of individual T-F units. Although this stage uses
oracle knowledge by assuming the correct number of active
sources while the number of sources in real-world scenarios
is unpredictable, it is in principle possible to estimate the
number of active sources based on the number of salient peaks
in the azimuth histogram. Another modification that is likely
to improve future implementations of the presented system,
is the integration of further top-down knowledge for the
grouping of glimpses. For example, the fragment-based speech
recognition system used in [43] and [44] applies an auditory
front-end to identify spectro-temporal source fragments using
a source-driven method. Subsequently, the selection of target
source fragments is achieved with a model-driven technique,
combining the fragments to generate the most likely word
sequence and effectively incorporating top-down processing

V. CONCLUSIONS AND FUTURE WORK

In the current study, a general framework for the automatic
segmentation of auditory scenes is presented. Specifically,
the first stage creates unlabeled segments based on feature
contrasts, which resemble auditory glimpses, and allows to
integrate information across spectro-temporal context as well
as different feature domains. During the second stage, the
previously formed segments are assigned to active sources
based on their prominent spatial properties which provides
a set of estimated IBMs for all identified sources within the
scene.

The advantage of estimating feature contrasts in the first
stage lies in replacing the absolute features with relative
features that are less source-specific than absolute features
and should generalize better to unseen conditions. Another
benefit of exploiting feature contrasts in the first stage rather
than global feature values is the possibility to exploit features
for which the global value may not be informative about the
source origin, but can still be informative about T-F units
belonging to the same source or not. Examples of such features
are pitch and energy features.

The main objective of this study was to apply a source-
independent grouping principle by training a DNN to detect
glimpse boundaries using ideal contrasts as training targets
instead of IBMs and therewith support source distinction,
rather than source recognition. Another goal was to construct a
general framework that can flexibly integrate different feature
types as well as context information into the contrast estima-
tion. It was shown that complementary grouping principles can
be exploited by a DNN and that analyzing spectro-temporal
context improves estimated contrast maps significantly. A clear
advantage was observed from integrating five of the six tested
features as compared to using the strongest single feature.
Besides evaluating the influence of additional feature types
on the contrast estimation in a future study, it would also be
interesting to observe the effect of including the already tested
features in a different manner, for example, by using ITD and
ILD directly as a feature and omitting the intermediate GMM
classifier. It has already been shown that the performance of
sound source localization [41] and sound source separation
[42] can be improved by training a DNN with spatial features
instead of using a GMM.

The addition of spectral context has been shown to generate
a strong advantage for all feature types while temporal context,
added a smaller overall performance gain. This may be at-
tributed to the fact that the relative features inherently capture
the temporal progression of features in two consecutive time
frames. Partly, the advantage of temporal context may also
depend on the acoustic scene, specifically the rate at which the
dominant source switches. Mixtures of up to 5 simultaneous
speakers in noise were evaluated in this study which means
that the dominant source switches at a relatively high rate and
the chance of consecutive time frames being dominated by the
same source is lower compared to a two-talker mixture.

The main objective of this study was to apply a source-
independent grouping principle by training a DNN to detect
glimpse boundaries using ideal contrasts as training targets
instead of IBMs and therewith support source distinction,
rather than source recognition. Another goal was to construct a
general framework that can flexibly integrate different feature
types as well as context information into the contrast estima-
tion. It was shown that complementary grouping principles can
be exploited by a DNN and that analyzing spectro-temporal
context improves estimated contrast maps significantly. A clear
advantage was observed from integrating five of the six tested
features as compared to using the strongest single feature.
Besides evaluating the influence of additional feature types
on the contrast estimation in a future study, it would also be
interesting to observe the effect of including the already tested
features in a different manner, for example, by using ITD and
ILD directly as a feature and omitting the intermediate GMM
classifier. It has already been shown that the performance of
sound source localization [41] and sound source separation
[42] can be improved by training a DNN with spatial features
instead of using a GMM.

The addition of spectral context has been shown to generate
a strong advantage for all feature types while temporal context,
added a smaller overall performance gain. This may be at-
tributed to the fact that the relative features inherently capture
the temporal progression of features in two consecutive time
frames. Partly, the advantage of temporal context may also
depend on the acoustic scene, specifically the rate at which the
dominant source switches. Mixtures of up to 5 simultaneous
speakers in noise were evaluated in this study which means
that the dominant source switches at a relatively high rate and
the chance of consecutive time frames being dominated by the
same source is lower compared to a two-talker mixture.

The final evaluation of percent correctly labeled T-F units
in the estimated IBMs shows a clear advantage from the
glimpse formation as compared to using spatial cues only.
This indicates that the segments based on contrast features
are useful to obtain a more reliable assignment of the T-F
plane to specific sources as compared to estimating IBMs on
the basis of individual T-F units. Although this stage uses
oracle knowledge by assuming the correct number of active
sources while the number of sources in real-world scenarios
is unpredictable, it is in principle possible to estimate the
number of active sources based on the number of salient peaks
in the azimuth histogram. Another modification that is likely
to improve future implementations of the presented system,
is the integration of further top-down knowledge for the
grouping of glimpses. For example, the fragment-based speech
recognition system used in [43] and [44] applies an auditory
front-end to identify spectro-temporal source fragments using
a source-driven method. Subsequently, the selection of target
source fragments is achieved with a model-driven technique,
combining the fragments to generate the most likely word
sequence and effectively incorporating top-down processing

Fig. 8: Percent correctly labeled T-F units, after assigning
either glimpses or single T-F units to active source locations.
The three different glimpsing methods are compared to a
baseline method that relies on spatial cues only. Shown are
mean values and 95% confidence intervals.
into the segregation.

Potentially, the framework presented in this paper can be applied for speech enhancement as well, by using the estimated masks to obtain attenuation factors. In this case, the estimated IBMs can be replaced by IRMs which have been shown to improve speech recognition [11] and cause less degradation of sound quality [45]. The estimation of IRMs can possibly be achieved by taking into account the uncertainty with which individual glimpses are assigned to the considered target source, similar to the estimation of source occupation probabilities per T-F unit implemented in [42] or [46].

One objective of this study was to observe the ability of the contrast estimation to generalize to unknown scenarios which was achieved by evaluating unknown speakers and noise conditions showing an improved performance for the untrained single speaker condition and an expectable but not dramatic decline for the untrained 5 speaker condition. For none of the tested conditions, performance fell below chance level.

In contrast to other systems, such as [47] which uses end-to-end training and achieves real-time separation, the separation system presented in this study can neither perform end-to-end training nor real-time separation. In exchange, it provides intermediate representations which give an opportunity to observe the influence of specific features and glimpse formation. This grants more interpretability and allows to make links to human auditory scene analysis, such as the exploitation of better-ear glimpses in multi-talker scenarios [28].

One aspect to be evaluated in future work is the applicability of this method to non-speech sounds and other scenarios such as reverberant environments. As pointed out in a study on universal sound separation in [48], most approaches for source separation are geared towards speech separation. Since one of the major assumptions in the present study was, that any onset of a new acoustic source should be reflected in observable feature contrasts, it would be interesting to study how well the relative features indicate contrasts for non-speech sounds.

ACKNOWLEDGMENT
This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 352015383 – SFB 1330. Simulations were performed on the the HPC Cluster CARL funded by the DFG under INST 184/157-1 FUGG.

REFERENCES

[1] S. J. Rennie, J. R. Hershey, and P. A. Olsen, “Single-channel multitalker speech recognition,” IEEE Signal Process. Mag., vol. 27, no. 6, pp. 66–80, 2010.
[2] E. W. Healy, M. Delfarah, J. L. Vasko, B. L. Carter, and D. Wang, “An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker,” J. Acoust. Soc. Am., vol. 141, no. 1, pp. 5254–5256, 2017.
[3] Z.-Q. Wang and D. Wang, “Deep neural networks for single-channel multi-talker speech recognition,” IEEE Trans. Audio, Speech, Language Process., vol. 23, no. 10, pp. 1670–1679, 2015.
[4] T. May, “Influence of binary mask estimation errors on robust speech identification,” Speech Comm., vol. 87, pp. 40 – 48, 2017.
[5] T. Bentzen, T. May, and T. Dau, “The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility,” PLoS ONE, vol. 13, no. 5, 2018.
[6] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in Speech separation by humans and machines, P. Dienveni, Ed. Kluwer Academic, Norwell, 2005, ch. 12, pp. 161–197.
[7] J. Barker, “Missing data techniques: Recognition with incomplete spectrograms,” in Techniques for Noise Robustness in Automatic Speech Recognition, T. Virtanen, R. Singh, and B. Raj. Eds. Wiley, 2012, ch. 5, pp. 371–398.
[8] T. May, S. van de Par, and A. Kohlrausch, “Noise-robust speaker recognition combining missing data techniques and universal background modeling,” IEEE Trans. Audio, Speech, Language Process., vol. 20, no. 1, pp. 108–121, 2012.
[9] G. J. Brown and M. Cooke, “Computational auditory scene analysis,” Computer Speech and Language, vol. 8, pp. 107–136, 1994.
[10] A. Narayanan and D. Wang, “The role of binary mask patterns in automatic speech recognition in background noise,” J. Acoust. Soc. Am., vol. 133, no. 5, pp. 3083–3093, 2013.
[11] ——, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” in Proc. ICASSP, 2013, pp. 7092–7096.
[12] C. Hummersone, T. Stokes, and T. Brooks, “On the ideal ratio mask as the goal of computational auditory scene analysis,” in Blind Source Separation. Advances in Theory, Algorithms and Applications, G. Nætt and W. Wang, Eds. Springer, 2014, ch. 12, pp. 349–368.
[13] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” IEEE Trans. Audio, Speech, Language Process., vol. 22, no. 12, pp. 1849–1858, 2014.
[14] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in Proc. ICASSP, 2016, pp. 31–35.
[15] Z.-Q. Wang, J. L. Vasko, and R. J. Hershey, “Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation,” in Proc. ICASSP, 2018, pp. 1–5.
[16] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” IEEE Trans. Audio, Speech, Language Process., vol. 25, no. 10, pp. 1901–1913, 2017.
[17] Z.-Q. Wang and D. Wang, “Combining spectral and spatial features for deep learning based blind speaker separation,” IEEE Trans. Audio, Speech, Language Process., vol. 27, no. 2, pp. 457–468, 2019.
[18] B. C. Moore and H. E. Gockel, “Properties of auditory stream formation,” Phil. Trans. R. Soc. B, vol. 367, pp. 919–931, 2012.
[19] K. Hu and D. Wang, “An unsupervised approach to cochannel speech separation,” IEEE Trans. Audio, Speech, Language Process., vol. 21, no. 1, pp. 122–131, 2013.
[20] Y. Shao and D. Wang, “Sequential organization of speech in computational auditory scene analysis,” Speech Comm., vol. 51, pp. 657 – 667, 2009.
[21] P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” IEEE Trans. Audio, Speech, Language Process., vol. 15, no. 1, pp. 1–12, 2007.
[22] Y. Shao, S. Srinivasan, J. Z. Jin, and D. Wang, “A computational auditory scene analysis system for speech segregation and robust speech recognition,” Computer Speech and Language, vol. 24, pp. 77 – 93, 2010.
[23] J. Woodruff and D. Wang, “Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues,” IEEE Trans. Audio, Speech, Language Process., vol. 21, no. 4, pp. 806–815, 2013.
[24] Q. K. Duong, E. Vincent, and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” IEEE Trans. Audio, Speech, Language Process., vol. 18, no. 7, pp. 1830–1840, 2010.
[25] T. May, S. van de Par, and A. Kohlrausch, “A probabilistic model for robust localization based on a binaural auditory front-end,” IEEE Trans. Audio, Speech, Language Process., vol. 19, no. 1, pp. 1–13, 2011.
[26] M. Cooke, “A glimpsing model of speech perception in noise,” J. Acoust. Soc. Am., vol. 119, no. 3, pp. 1562–1573, 2006.
[27] E. Schoonmaker, S. Sutojo, and S. van de Par, “Better-ear rating based on glimpsing,” J. Acoust. Soc. Am., vol. 142, no. 3, pp. 1466–1481, 2017.
[28] D. S. Brungart and N. Iyer, “Better-ear glimpsing efficiency with symmetrically-placed interfering talkers,” J. Acoust. Soc. Am., vol. 132, no. 4, pp. 1466–1481, 2012.
[29] B. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes symmetrically-placed interfering talkers,” J. Acoust. Soc. Am., vol. 119, no. 3, pp. 1562–1573, 2006.
[30] E. D. Young, “Neural representation of spectral and temporal information in speech,” Phil. Trans. R. Soc. B, vol. 363, pp. 923–945, 2008.
Z. Chen, Y. Luo, and N. Mesgarani, “Deep attractor network for single-microphone speaker separation,” in Proc. ICASSP, 2017, pp. 246–250.

I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. L. Roux, and J. R. Hershey, “Universal sound separation,” in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019, pp. 175–179.

Steven van de Par studied physics and mathematics at the University of Oldenburg, Germany, and received a B.A. degree in 2012 and a M.Sc. degree in engineering physics from the University of Oldenburg, Germany in 2015. In 2021 she completed a Ph.D. degree on the topic of computational auditory scene analysis in the Acoustics Group, Department of Medical Physics and Acoustics at the University of Oldenburg where she is currently working as a Postdoctoral Researcher.

Tobias May studied hearing technology and audiology and received his M.Sc. degree in 2007 from the University of Oldenburg, Germany. In 2012, he obtained the binational Ph.D. degree from the University of Oldenburg in collaboration with the Eindhoven University of Technology, The Netherlands. From 2013 to 2017 he was a postdoctoral researcher at the Technical University of Denmark. He joined the faculty in 2017 as an Assistant Professor in speech signal processing and hearing technology and was promoted to Associate Professor in 2020. His research interests include computational auditory scene analysis, binaural signal processing, noise-robust speaker identification and signal processing strategies for hearing instruments.

Sarinah Sutojo studied Physics and Music at the University of Oldenburg, Germany and received a B.A. degree in 2012 and a M.Sc. degree in engineering physics from the University of Oldenburg, Germany in 2015. In 2021 she completed a Ph.D. degree on the topic of computational auditory scene analysis in the Acoustics Group, Department of Medical Physics and Acoustics at the University of Oldenburg where she is currently working as a Postdoctoral Researcher.

Authorized licensed use limited to: Danmarks Tekniske Informationscenter. Downloaded on March 14,2022 at 12:23:50 UTC from IEEE Xplore. Restrictions apply.