BINAURAL SOURCE SEPARATION IN NON-IDEAL REVERBERANT ENVIRONMENTS

Sylvia Schulz and Thorsten Herfet

Telecommunications Lab
Saarland University, Germany
{schulz, herfet}@nt.uni-saarland.de

ABSTRACT

This paper proposes a framework for separating several speech sources in non-ideal, reverberant environments. A movable human dummy head residing in a normal office room is used to model the conditions humans experience when listening to complex auditory scenes. Before the source separation takes place the human dummy head explores the auditory scene and extracts characteristics the same way as humans would do, when entering a new auditory scene. These extracted features are used to support several source separation algorithms that are carried out in parallel. Each of these algorithms estimates a binary time-frequency mask to separate the sources. A combination stage infers a final estimate of the binary mask to demix the source of interest. The presented results show good separation capabilities in auditory scenes consisting of several speech sources.

1. INTRODUCTION

Humans are masters in analyzing their auditory environment and in separating different sound sources. Consider the classical cocktail party example, where several people are talking simultaneously in the same room. Humans have no difficulty to attend to a single person while ignoring all the other people, additional artificial sources and background noise. Today's computational approaches for source separation – especially in reverberant environments – are far from achieving this extraordinary ability of the human brain.

When humans enter an auditory scene they first look around and estimate several features of the environment around. When source separation is required – i.e. when starting a conversation with another person – this knowledge is used to enhance the separation process. The presented source separation framework tries to model this human behavior to enhance the following source separation. To imitate the human listening situation, a robotic human dummy head, called Bob, is used. Bob resides in a normal office room of size 10×6 m and a reverberation time $RT_{60} = 0.4\ s$ and is able to move in three degrees of freedom to explore the auditory scene around him. A conventional 7.1. loudspeaker installation is utilized to construct an auditory scene consisting of several spatially separated sources by assigning each source to a specific loudspeaker. The auditory scene around is recorded via microphones in Bob's ears.

2. TIME-FREQUENCY MASKS

Rickard et al. [1] showed that speech signals are sparsely distributed in high-resolution time-frequency (TF) representations. TF representations of different speech signals overlap only in few points and so are approximately orthogonal to each other. This approximate orthogonality in the TF-domain justifies the use of TF-masks that emphasize regions of the TF-spectrum that are dominated by a specific source and attenuate regions dominated by other sources or noise. Masking effects in the human auditory system motivate the use of binary TF-masks: Within a critical bandwidth humans don't recognize sounds that are masked by louder sounds.

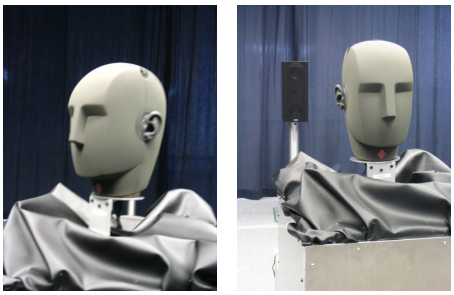

Figure 1: *Bob – The robotic head.*

Several researchers in computational source separation suggest an ideal binary mask as final goal of computational source separation algorithms (i.e.[2], [1]). Brungart et al. [3] support this goal by noting that the intelligibility of separated sounds increases if more and more energy of the ideal binary mask is reconstructed.

Assume $s_i(t,f)$ denotes the energy of the target signal$_i$ in TF-bin at time t and frequency f and $n_j(t,f)$ denotes the energy of the j-th interfering signal in this TF-bin. The ideal binary mask $\Omega_i(t,f)$ for target source$_i$ and a threshold of 0 dB is defined as follows:

$$\Omega_i(t,f) = \begin{cases} 1 & s_i(t,f) - n_j(t,f) > 0 \quad \forall j \\ 0 & \text{else} \end{cases} \tag{1}$$

2.1. Short-Time-Fourier-Transform

A commonly used TF-representation is the lossless and computationally efficient Short-Time-Fourier-Transform (STFT). The discrete STFT analyzes the time-domain signal in linearly spaced frequency channels up to the Shannon frequency. For a general discrete signal $x(n)$ and an arbitrary discrete analysis window function $w(n)$ the STFT is defined $\forall q \in \{0, 1, ..., N-1\}$ as

$$X(k,q) = \frac{1}{\sqrt{N}} \cdot \sum_{n=0}^{N-1} w(n)x(n+k)e^{-i2\pi\frac{qn}{N}}. \tag{2}$$

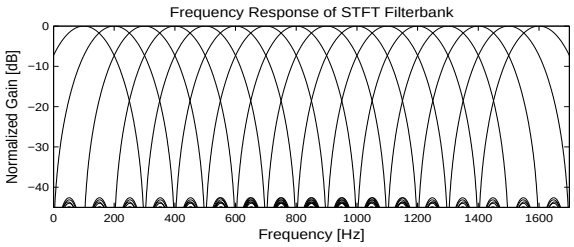

Figure 2: *Frequency Response of STFT using a Hamming window.*

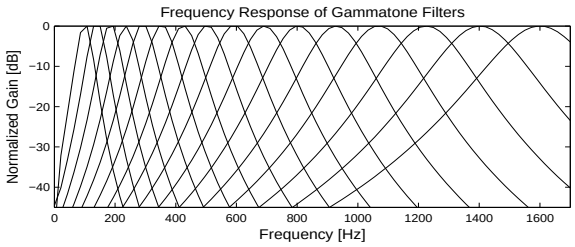

Figure 3: *Frequency Response of Gammatone Filterbank.*

Figure 2 shows the positive frequency response of the STFT for a Hamming window of length 32 using a sampling frequency of 3.2 kHz. The shape of the linearly spaced filter channels and the overlap between two consecutive channels is specified by the shape of the analyis window function.

Yilmaz et al. [1] showed that the approximate orthogonality of different speech sources in the discrete STFT representation with Hamming windows of 64 ms length is satisfied and the STFT spectrum is a suitable and easy representation for assigning complete time-frequency regions to specific sources.

Besides an amplitude and phase estimate for each bin in the spectrum, the STFT provides no further low-level information about this bin that could be used to infer the dominating source. Because of the limited time and frequency resolution the estimates are only coarse and averaged over the complete analysis window. If a specific bin is dominated by one source, there may also be energy of other sources in this bin which severely forge the amplitude and phase estimates.

Almost all energy of speech signals is distributed in frequencies up to 8 kHz. For analysing speech signals, a finer frequency resolution in the low frequency range is favorable, whereas in higher frequencies a coarse resolution is sufficient. Because the STFT analyses linearly up to the Shannon frequency, the frequency resolution in the low frequencies cannot be enhanced by increasing the sampling rate.

A source separation algorithm should use all information about a specific time-frequency region to increase the possibility of correct assignment. Using only an amplitude and a phase estimate for the assignment decision of a complete STFT-bin is quite limited and is not very reliable in reverberant mixtures. To enhance the decision process more information about each STFT-bin must be examined.

2.2. Cochleagram

Many source separation architectures try to imitate the frequency analysis of the human auditory system. The frequency analysis of the human cochlea can be approximated using a bank of gammatone filters. The impulse response of a gammatone filter is defined as the product of a gamma function and a tone [4]:

$$g_{f_c}(t) = t^{N-1} e^{-2\pi b(f_c)} \cdot \cos(2\pi f_c t + \phi) \quad \forall t \geq 0 \qquad (3)$$

where N denotes the order of the filter and f_c denotes the center frequency of the filter. The value $b(f)$ determines the bandwidth of the filter and is usually set to the equivalent rectangular bandwidth (ERB) of human auditory filters. A bank of such gammatone filters gives a good fit to experimentally derived estimates of the frequency analysis of the human cochlea and for such the TF-representation of a gammatone filter bank is commonly called cochleagram.

Figure 3 illustrates the frequency response of a bank of 16 gammatone filters in the frequency range from 100 - 1600 Hz. Consecutive filters are spaced logarithmically on the frequency scale. Filter channels in the low frequencies have fine frequency resolution, but coarse time resolution. Conversely the high frequency channels have coarse frequency resolution, but fine time resolution. The coarse time-resolution in the low frequencies is acceptable as signals consisting of low frequencies change slowly, whereas high-frequency signals need finer time-resolution to illustrate the rapid changes.

The inversion of a given cochleagram to a time-domain signal is non-trivial and lossy. There exist some approaches that yield quite good invertion results (i.e. [4], [5]), but these are complex to compute and only approximately orthogonal, which results in non-perfect reconstruction.

3. OVERALL ARCHITECTURE

The STFT is easy and lossless to compute, but the filter channels are positioned linear on the frequency scale which yields only a coarse frequency resolution in the important low frequencies. Also the amplitude and phase information are averaged over the complete analysis window and so not really reliable in reverberant environments.

The cochleagram on the other hand analyses the signal with logarithmically spaced filter channels and allows a finer frequency resolution in the low frequencies, but the invertion of a given cochleagram to a time-domain signal is quite complex.

The source separation framework presented in this paper combines the positive features of the STFT with the positive features of the cochleagram while eliminating some of the negative features. The overall goal of the source separation is to find the ideal STFT-mask. The core source separation process however is based on the analysis of the corresponding region in an additionally computed cochleagram. This way the macroscopic STFT-transform is used to define the demixing masks and to finally demix the original sources. The core assignment of each STFT-bin to a specific source is based on the corresponding region in the microscopic cochleagram and is only supported by the information gained from the STFT-spectrum.

This proceeding is analog to the approaches used in MPEG audio coders. For example MPEG Audio Layer 3 uses a FFT of 1024 samples to analyse the input signal and to apply the psychoacoustic models. The critical subsampling however is realized using only 32 subbands [6].

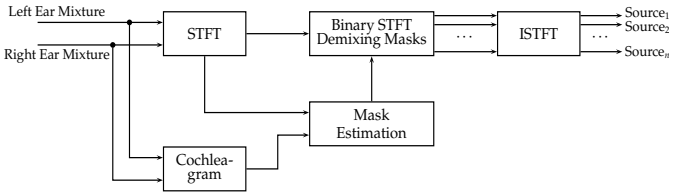

Figure 4: *Overall architecture for source separation framework.*

Figure 4 illustrates the system architecture of the source separation framework. The incoming signals of the left and right ear are STF-transformed and the respective cochleagram of each ear signal is computed. The mask estimation process computes for each source a binary STFT-mask based on the information gathered from the detailed cochleagram and supported by coarse information of the STFT spectrum. Finally the STFT spectrum is multiplied by each binary STFT mask and is transformed back to the time-domain, yielding the demixed time-domain signals.

The mask estimation stage tries to make use of all information that could be established using standard or sophisticated signal processing methods. In a first step Bob, the movable human dummy head, analyses the auditory scene and identifies the position of the preferred speaker in the room. In further steps this information is used to enhance the source separation, that uses both interaural and monaural cues to distinguish the TF-bins. Because of reverberation many of the cues used to separate bins are distorted and can only be used to some extent. To face the reflections and reverberations several algorithms compute independent estimates of the binary masks. In a final stage the estimated masks of each algorithm are combined to find a best estimate.

4. AUDITORY SCENE EXPLORATION

When humans enter an auditory scene such as a cocktail party, they automatically analyse the environment around them. Humans recognize the number of possible sound emanating sources, classify them according to speech or artificial sounds and estimate or recall from memory several expected features of each source. When a communication between the human and one of the sources begins, much information is already known to human cognition and is used to support and enhance the separation process.

The source separation architecture presented in this paper tries to mimic these cognitive abilities of the human brain. So prior to separating the speech sources, Bob analyses the auditory scene and estimates several parameters that can be used to enhance or enable later separation approaches. The following separation algorithms expect as input the position of the source to be enhanced in the azimuth plane. Furthermore some of the separation schemes require an estimate of the fundamental frequency of the desired speech source.

4.1. Source Localization

In the following sections and the source separation algorithms presented later the source of interest is assumed to be the – in some sense – strongest source in the auditory scene. The localization of the desired source is realized using an adaptive estimate of the interaural time differences between the two ears.

The interaural time difference (ITD) – the arrival time difference between the left and right ear signal – is used as localization cue and is estimated based on the correlation between the two signals. Assume x_L and x_R denote the time domain signal of the left and the right ear. The correlation function is defined as

$$R_{x_L x_R}(l) = \sum_{t=t_s}^{t_e} x_L(t+l) \cdot x_R(t). \tag{4}$$

Each source in the auditory scene contributes a peak in the correlation function. Further peaks can be introduced by reflections and reverberations. Detecting the highest peak in $R_{x_L x_R}$ yields a first estimate of the incidence direction of the strongest source, so the movable human dummy head Bob turns to this estimated position.

Because of the reverberation and interference Bob cannot rely on the validity of the estimated position. Therefore a further correlation at the new position is computed that should in the ideal case have its highest peak at the position of zero degree. To account to the reverberant environment the position is regarded to be confirmed if one of the highest peaks is located near zero degree. If this peak deviates from zero with only some degrees, Bob enhances the located position. This procedure is iterated until a stable position is reached and the regarded peak of the correlation function appears at approximately zero degree. Then Bob directly faces the source of interest which is now centered around 0° relative to Bob's facing direction.

Because the specific resonances of the human ear and head are not used yet, Bob cannot distinguish between front and back only from analysing the ear signals. Possible front-back confusions are resolved by slightly moving the head to one side at the final position and measuring the direction of change of the ITD between the two ears. If it turns out that Bob has mistaken the direction, Bob turns 180° around and faces the correct source.

For a detailed description of the design, implementation and results of the source localization scheme consider the work of Haschke [7].

4.2. Fundamental Frequency Estimation

Humans tend to emit frequencies that are an integer multiple of their own fundamental frequency (F0). Especially voiced parts of speech contain most of the energy in the harmonics of F0. Source separation approaches can use the F0 to determine those frequencies that are mainly used by a speaker.

The used fundamental frequency estimation relies on an algorithm known as "Robust Algorithm for Pitch Tracking" (RAPT) [8] and determines the F0 of spoken utterances as a function of time. The time-domain signal is split in time-frames of length 5 ms and for each frame a F0-estimate is computed based on the autocorrelation of the corresponding signal.

RAPT is originally designed to work in anechoic, single source recordings and computes reliably an estimate of the F0-track and the first harmonics. In reverberant multi source recordings – such as the recordings of Bob's ears – the estimation process severely degrades and the result forms a mixture of each F0-track and additional noise.

Assuming the source of interest is directly in front of Bob and the interfering sources are distributed at other positions, the preferred signal can be enhanced by applying simple beamforming: The right and the left ear signal are summed and divided by two.

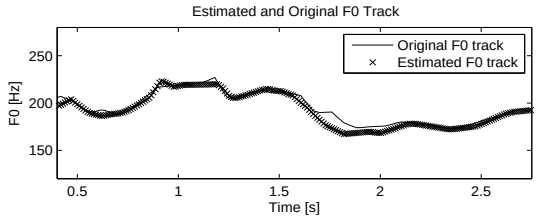

Figure 5: *Original F0-track and reconstructed F0-track for a mixture of two speech sources.*

This way the resulting signal emphasizes the preferred source and further smears the interfering sources.

The enhanced signal is used as input to the RAPT to get an estimate of the F0-track of the preferred source. For each time-frame the F0-estimate produced by RAPT is considered valid, if corresponding estimates are found in several higher harmonics. The final F0-track is constructed by linear interpolation of the valid F0s.

Figure 5 shows the result of the F0-track reconstruction for an auditory scene consisting of two speech sources at positions 0° (Bob has already geared towards the source) and an interfering source at position 45° to the right.

5. SEPARATION ALGORITHMS

Existing source separation approaches can be broadly classified into two categories:

- **Separation algorithms based on Interaural Cues** use interaural time and level differences to separate the sources. In ideal anechoic mixtures the direction of each TF-bin can be estimated and the bin is easily assigned to the correct source. Unfortunately echoic recordings blur and distort these interaural cues, so separation capabilities decrease.
- **Separation algorithms based on Monaural Cues** use only characteristics that are specific to a single signal and do not rely on the differences between the left and right ear. These algorithms mostly use the fundamental frequency of the speaker as a main feature to separate the sources.

The following algorithms assume that Bob has already analyzed the auditory scene and has turned towards the preferred source. Furthermore he has estimated the F0-track of the source of interest as described before. Imitating the human behavior, Bob automatically aligns his head to the source of interest. The goal of the following source separation algorithms is to enhance a specific source of interest, not to separate all sources.

5.1. Separation based on Interaural Time Differences

Interaural Time Differences between the left and right ear signal are used to examine the position of the respective source of each STFT-bin. Because the STFT phase value is not necessarily reliable as discussed previously, the ITD is estimated using the cochleagram. Let $X_{L_{stft}}$ and $X_{R_{stft}}$ denote the STFT-representation of the left and right ear signal and $X_{L_{co}}$ and $X_{R_{co}}$ the corresponding cochleagram representations. For each STFT-bin the corresponding left and right TF-windows $W_{L_{co}}$ and $W_{R_{co}}$ are cut out of the cochleagram. The ITD estimates of $W_{L_{co}}$ and $W_{R_{co}}$ are computed using a running cross-correlation across the time-dimension of the time-frequency regions: $\forall l \in \{-\mathrm{maxLag}, \mathrm{maxLag}\}$

$$R_{W_{L_{co}} W_{R_{co}}}(l) = \sum_{t=t_s}^{t_e} \sum_{f=f_s}^{f_e} W_{L_{co}}(t+l, f) \cdot W_{R_{co}}(t, f) \quad (5)$$

The highest peak of $R_{W_{L_{co}} W_{R_{co}}}$ yields the best estimates of the ITD for this bin. Because of reverberation and reflections there could be further peaks in the correlation function that could refer to the correct ITD and therefore should be considered. According to Faller and Merimaa [9], the height of the peak in the correlation function is a measure of reliability: The higher the peak, the more reliable the ITD estimation.

Knowing that the preferred source is at azimuth zero degree, the ITD computations offer the following three algorithms to estimate the ideal binary masks:

Algorithm 1 Assign to the source of interest all TF-bins where the estimated ITD of the highest peak of the correlation function yields an angle of incidence that deviates not more than δ° from 0° and the height of the peak is greater than h.

Algorithm 2 Assign to the source of interest all TF-bins where the estimated ITD of an existent second highest peak of the correlation function yields an angle of incidence that deviates not more than δ° from 0° and the height of the peak is greater than h.

Algorithm 3 Assign to the source of interest all TF-bins where the estimated phase of the STFT bin yields an angle of incidence that deviates not more than δ° from 0°.

Each algorithm regards only these STFT-bins that contain more energy than a specific threshold. To compare the results of each algorithm, the estimated masks are compared with an ideal mask, which is estimated from recordings of the single sources under reverberant conditions. Because reverberation differs slightly between recordings with only one source and recordings with several sources, this ideal mask is only an approximation of the real ideal mask. Using this estimated ideal mask as ground-truth there are three evaluation criteria for each algorithm:

1. *The percentage of recovered energy* of the ideal mask. The higher this percentage, the more energy of the original signal is recovered and the speech intelligibility of the desired source increases.
2. *The percentage of false estimated bins* denotes the relative number of bins that are wrongly assigned to the preferred source. According to the ideal masks, these bins should be assigned to one of the other sources of the auditory scene, as the absolute value of energy contribution to this bin of another source is larger than the energy contribution of the desired source. The lower this value, the less artifacts from other speech sources are contained in the estimated mask.
3. *The percentage of correct estimated bins* clarifies how much of the estimated bins are correctly assigned to the source of interest. The gap between this value and the percentage of false estimated bins indicates the number of those TF-bins that happen to have high energy in the recorded mixture, but none of the ideal masks of the single sources exhibit

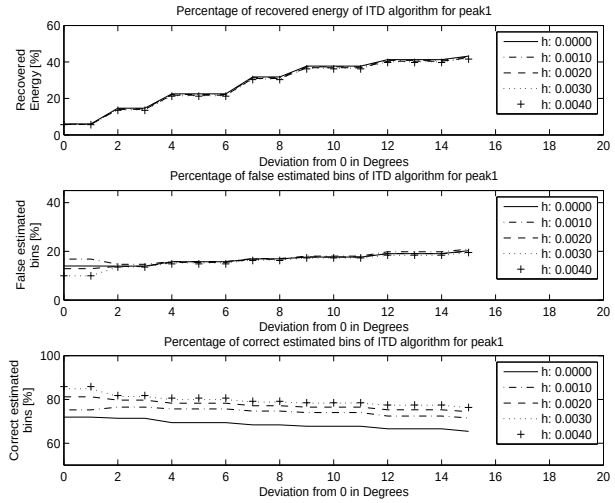

Figure 6: *Percentage of recovered energy of ideal mask, false estimated bins and correct estimated bins for the separation algorithm based on ITD and highest peak (Algorithm 1).*

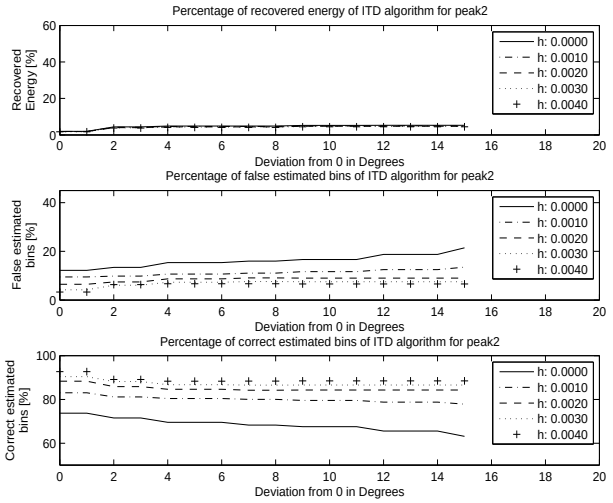

Figure 7: *Percentage of recovered energy of ideal mask, false estimated bins and correct estimated bins for the separation algorithm based on ITD and second highest peak (Algorithm 2).*

high energy in this TF-region. So these bins are likely to occur from reverberations and cannot be assigned to a specific source.

Figure 6 shows the results of algorithm 1 for different δ and h for an auditory scene consisting of two speech sources. Speech source one is located directly before the head at 0° as Bob has already geared to the source and the second source is located at 45° to the right. One can clearly see that the percentage of recovered energy increases if the deviation from zero increases. The ITDs of most of the correct TF-bins deviate considerably from the real position at 0°. In contrast to the percentage of reconstructed energy and the number of false estimated TF-bins, the percentage of correct estimated bins increases according to the peak height. TF-bins with high energy – which mostly result in high correlation peaks – are more likely to yield a correct ITD estimation as opposed to low energy bins. To achieve a good tradeoff a δ between 8 and 12 degree and high peak height h is favorable.

The same results for algorithm 2 are illustrated in figure 7. The percentage of reconstructed energy is much lower than in the case of algorithm 1. This low percentage is due to the fact that a second peak in the correlation function in most cases only exists for TF-bins at high frequencies where the correlation analysis window becomes bigger than the period of this bin. Those high frequency bins naturally include lower energy than low frequency bins, so the overall recovery is quite low.

Figure 8 plots the results for algorithm 3. The number of false estimated bins is approximately constant and the percentage of correct estimated bins decreases very slowly. As also seen in algorithm 1, the phase values of the bins deviate quite a lot from the ideal position due to reflections and interference from the other sources.

5.2. Separation based on Fundamental Frequency

If two persons speaking have a considerable different F0, their harmonics do not overlap in many frequencies. If the F0 of the preferred speaker is known in advance, this information can be used to assign the TF-bins. Each bin with a frequency value near a multiple of the fundamental frequency is more probable to belong to the preferred source, than it is to belong to one of the other sources. If additionally the F0s of the other speakers are known, the distances of the harmonics of the preferred speaker to the nearest harmonics of the other speaker can be computed and used to find the frequencies at which only the preferred speaker is present. The following F0-based algorithms are examined in the source separation architecture:

Algorithm 4 Assign to the source of interest all TF-bins where the frequency of the current STFT-bin deviates by no more than Δf Hz from the nearest harmonic of the preferred source's mean F0. If the F0 of the interfering sources is known, also the distance from the nearest interfering harmonic is used to segregate the TF-bins.

Algorithm 5 Assign to the source of interest all TF-bins where the frequency of the current STFT-bin deviates by no more than Δf Hz from the nearest harmonic of the preferred source's mean F0 and the energy in the cochleagram at the corresponding TF-unit is larger than a threshold E.

Algorithm 6 Assign to the source of interest all TF-bins where the frequency of the current STFT-bin deviates by no more than Δf Hz from the nearest harmonic of the preferred source's current F0 estimate. If the F0 of the interfering sources is known, also the distance from the nearest interfering harmonic is used to segregate the TF-bins.

Algorithm 7 Assign to the source of interest all TF-bins where the frequency of the current STFT-bin deviates by no more than Δf Hz from the nearest harmonic of the preferred source's current F0 estimate and the energy in the cochleagram at the corresponding TF-unit is larger than a threshold E.

Figure 9 displays the results of algorithm 4. The percentage of recovered energy grows as the maximal distance of the nearest harmonic of the preferred speaker grows as more and more bins

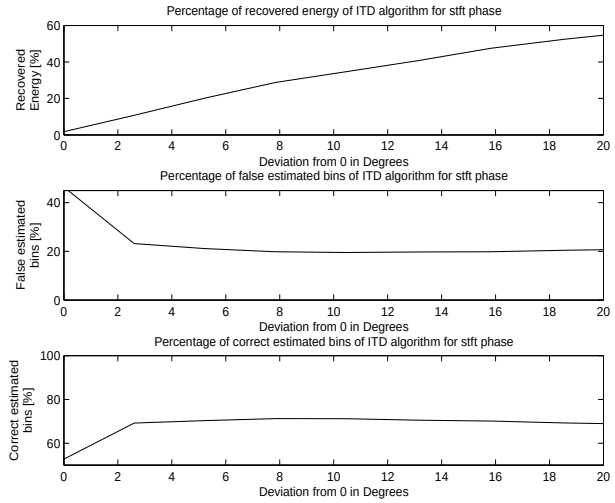

Figure 8: *Percentage of recovered energy of ideal mask, false estimated bins and correct estimated bins for the separation algorithm based on the STFT phase (Algorithm 3).*

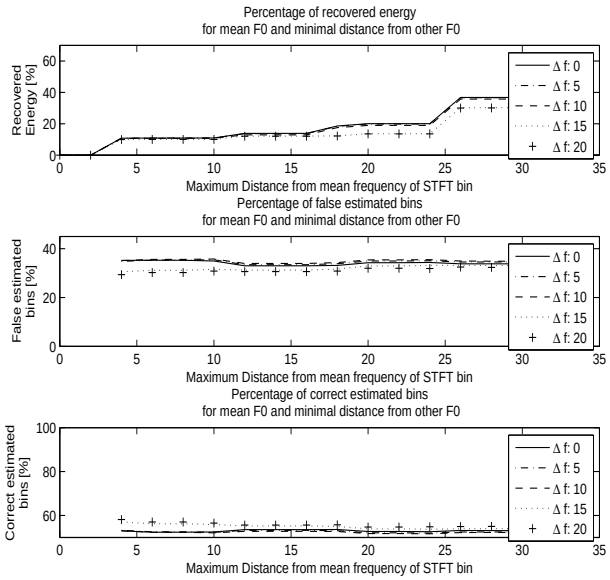

Figure 9: *Performance of the separation algorithm based on known mean F0 and the distance from the harmonic to the mean frequency of the current STFT bin (Algorithm 4). Shown are several curves for different frequency distances of the nearest interfering harmonic of the interfering speaker.*

are considered. Knowing the mean F0 of the interfering source has only minor effects that could be neglected. The rate of the correct estimated bins is constantly very low. The bad results of this algorithm arise from the fact that during a spoken word, the F0 of a human is not constant and varies about several Hz. So if the source separation relies only on the mean F0, higher order harmonics are computed incorrectly and the source separation capabilities decrease. If the complete track of the fundamental frequency is known, the separation algorithms discussed above can be enhanced.

If the absolute energy value of the regarded harmonic in the corresponding cochleagram window is used, the number of false bins and the percentage of correct bins can be slightly enhanced. Figure 10 illustrates the evaluation of algorithm 5. The higher the energy in the corresponding frequency, the higher the probability that the considered bin belongs to the preferred source. These results contribute to the reverberant environment: TF-bins with high energy and corresponding F0-characteristics are likely to originate from the main incidence direction and not from a disturbing reflection.

Figure 11 and 12 show the same results for algorithms 6 and 7, but using a complete F0 track instead of a mean value. The rate of the false estimated bins is about 10% lower than in the mean F0 case. Also the percentage of correct estimated bins is higher compared to using only an average F0. Assuming that the directly incident TF-bins have high energy, a minimum energy threshold can enhance the percentage of correct estimated bins by up to 30%. Optimal values can be achieved by using quite large maximum distances of 20 to 30 Hz.

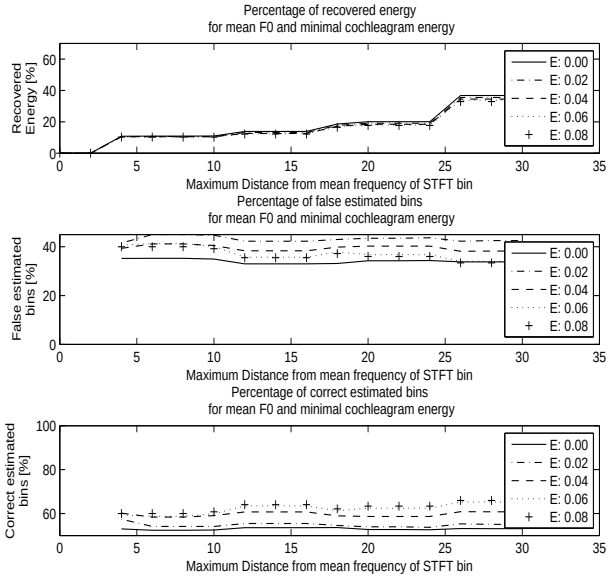

Figure 10: *Performance of separation algorithm based on known mean F0 and the distance from the harmonic to the mean frequency of the current STFT bin (Algorithm 5). Shown are several curves for different cochleagram energy levels E.*

6. COMBINING OF ALGORITHMS

Each of the introduced algorithms yields a reconstructed preferred speech source with a low to intermediate intelligibility. To enhance the separation capabilities, the algorithms work together to combine their information regarding each TF-bin. In a first stage each discussed algorithm separately estimates a STFT-demixing

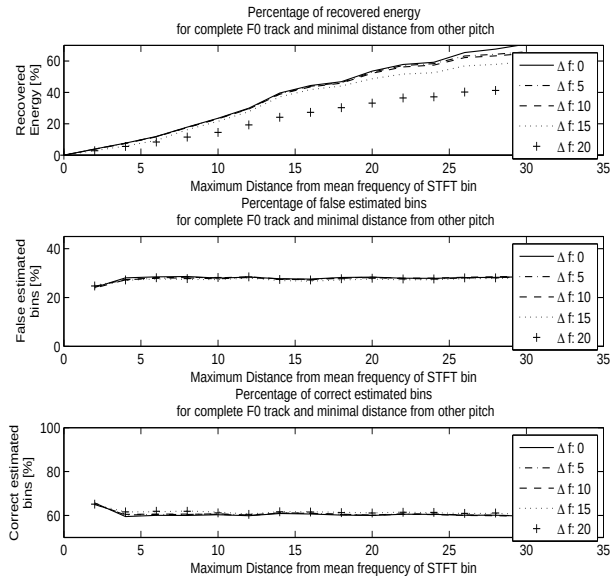

Figure 11: *Performance of the separation algorithm based on known complete F0-track and the distance from the harmonic to the mean frequency of the current STFT bin (Algorithm 6). Shown are several curves for different frequency distances of the nearest interfering harmonic of the interfering speaker.*

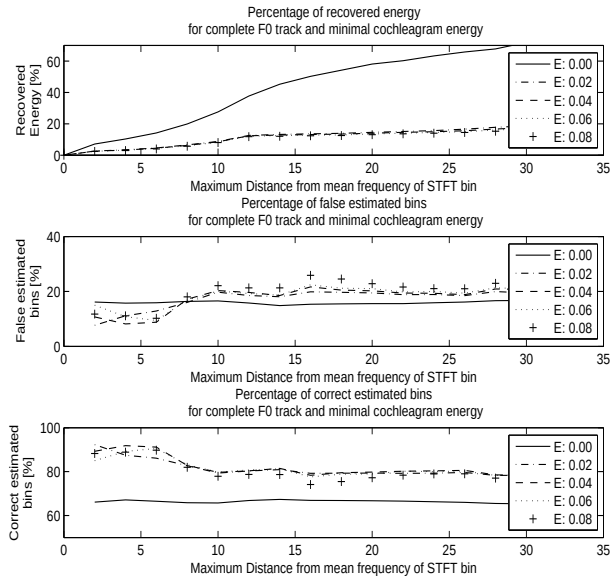

Figure 12: *Performance of separation algorithm based on known complete F0 track and the distance from the harmonic to the mean frequency of the current STFT bin (Algorithm 7). Shown are several curves for different cochleagram energy levels E.*

Algorithm Order	Recovered Energy of ideal mask [%]	Energy of estimated mask belonging to Interferer [%]
$2 \cap 6 \cup 4 \cup 3$	79.89	7.99
$6 \cap 5 \cup 7 \cup 3 \cup 2$	83.90	9.39
$1 \cup 6 \cup 7 \cup 5$	89.09	10.98
$4 \cup 1 \cup 7 \cup 2 \cup 6$	89.82	10.95

Table 1: *A selection of the best evaluation results of the sequential and parallel combining of algorithms 1-7 regarding the percentage of reconstructed energy for an auditory scene consisting of two sources.*

mask for the source of interest. A second central combining stage combines the single masks resulting in a final estimate of the ideal binary mask which is then used to demix the preferred source from the mixture.

A first separation approach combines the estimated masks in a sequential way similar to a chain of responsibility. The first algorithm in the chain assigns all bins according to its specification and passes the remainder of the bins to the second algorithm which in turn assigns those bins that match its specifications and passes the rest to the next algorithm and so on. This sequential combining of the algorithms is equivalent to computing the logical 'or' of the estimated single masks.

To further enhance the final estimated mask, a second approach additionally uses parallel combining to enhance the estimated masks. If several of the algorithms have assigned a specific bin to the preferred source, then this bin is more probable to belong to the source of interest than bins that are only assigned by a single mask. This parallel combining is realized using the logical 'and' of the single estimated masks.

Some results of the evaluation of the combining are summarized in table 1 and 2. The values are obtained by averaging over several recorded mixtures consisting of a female and male speaker positioned at 0° and 45° to the right. The English speech recordings are taken from the CMU speech database [10] and played back at the corresponding directions in a normal office room of size $10 \times 6\ m$ and $RT_{60} = 0.4\ s$. The maximum allowed deviation in degree from zero for algorithm 1 and 2 is set to 8° with a minimum correlation peak height of 0.001. The deviation of the STFT phase values used in algorithm 3 is bounded by 11°. Algorithms 5 and 7 use only TF-bins with energy higher than 0.01.

The resulting estimated masks are evaluated by noting mainly two values: The percentage of recovered energy of the preferred source declares how much of the total energy of the ideal mask of the preferred source is reconstructed. A value of 100% states that the estimated mask fully contains the ideal mask. The percentage of interference energy indicates how much energy of the estimated mask belongs to the interfering sources and noise and so is falsely assigned to the estimated mask.

The best estimated mask in terms of percentage of recovered energy is – amongst other combinations not shown for purposes of clarity – calculated using the sequential combination of algorithms 4,1,7,2 and 6. The estimated mask recovers 89.82% of the energy of the preferred source. On the other hand 10.95% of the energy

Algorithm Order	Recovered Energy of ideal mask [%]	Energy of estimated mask belonging to Interferer [%]
$3 \cup 2 \cap 1 \cup 6$	24.48	0.24
$2 \cap 3 \cup 6 \cap 7 \cup 5$	39.36	1.48
$2 \cap 3 \cup 6 \cap 5 \cup 1$	39.83	1.63
$3 \cup 4 \cap 6$	40.23	2.79

Table 2: *A selection of the best evaluation results of the sequential and parallel combining of algorithms 1-7 regarding the percentage of interfering energy for an auditory scene consisting of two sources.*

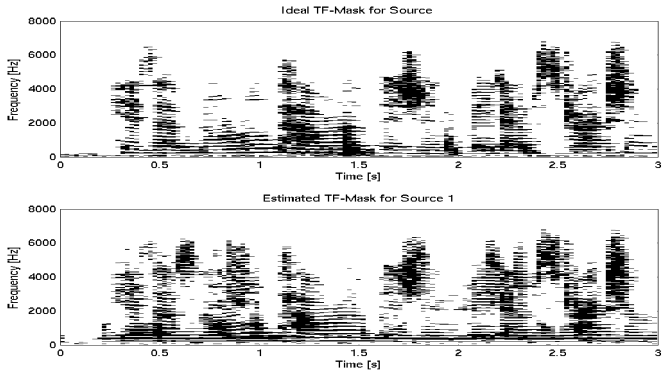

Figure 13: *The ideal mask for the source of interest and the best estimated mask regarding percentage of recovered energy. The recovered energy is* 89.82% *and* 10.95% *of the total energy belongs to the interferers.*

of the estimated mask belongs to the interfering source, yielding a value of 89.05% correct estimated energy. Listening tests result in a very good intelligibility of the source of interest, but the interfering source can be recognized as additional, but very quiet and unintelligible voice in the background.

Table 2 shows a selection of the best estimated masks regarding a minimum energy of interfering sources. Using for example the combination of algorithms $3 \cup 2 \cap 1 \cup 6$ yields estimated masks that recover 24.48% of the total energy of the preferred source while only 0.24% of the total energy of the estimated mask belong to the other source. The intelligibility of the separated speech is quite good and no interfering sources are audible. But compared to the demixed sources of table 1 the reconstructed speech is not so rich and authentic.

The strategy used for combining the masks estimated by the algorithms is dependent on the purpose of the separation infrastructure. If the source of interest is to be enhanced for better intelligibility by humans, sequential strategies should be applied. If however the framework is used as input to an automatic speech recognizer – which in most cases is very sensitive to interfering speech sources – hybrid schemes combining the parallel and sequential strategies are adequate. Other purposes could choose a combination which balances the percentage of recovered and interfering energy to gain an intermediate quality.

7. CONCLUSIONS AND FUTURE WORK

The binaural source separation architecture presented in this paper works well in non-ideal reverberant environments. Prior information regarding the auditory scene are useful to enhance the separation process. Parallel processing paths ensure that the assignment process is optimized regarding the available information at the decision process. By this means the introduced framework achieves quite good separation of speech sources.

Future work especially includes further exploration of the auditory scene. If the source separation algorithms know more characteristics such as the positions of the interfering sources and the respective fundamental frequencies in case of speech sources, the separation could be further enhanced. Additionally in mixed auditory scenes consisting of speech and artificial sources a classification and characterization of each source could assist the separation process.

On the other hand the combining of the masks is currently very rudimentary. Applying higher order inference to the estimated masks will probably further increase the source separation capabilities. Fuzzy logic systems for example can model human reasoning strategies very well and could be applied to infer the dedicated source of each STFT-bin based on all available information.

8. REFERENCES

[1] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830 – 1847, July 2004.

[2] D. L. Wang, "On ideal binary masks as the computational goal of auditory scene analysis," *In Divenyi P. (ed.), Speech Separation by Humans and Machines*, pp. 181 – 197, 2005.

[3] D. S. Brungart, P. S. Chang, B. D. Simpson, and D.L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *The Journal of the Acoustical Society of America*, vol. 120, pp. 4007 – 4018, 2006.

[4] D. L. Wang and Guy J. Brown, *Computational Auditory Scene Analysis - Principles, Algorithms, Applications*, IEEE Press, Wiley Interscience, 2006.

[5] G.J Brown and M. P. Cooke, "Computational auditory scene analysis," *Computer speech and language*, vol. 8, pp. 297 – 336, 1994.

[6] Ulrich Reimers, *Digital Video Broadcasting - The Family of International Standards for Digital Video Broadcasting*, Springer, 2005.

[7] Eric Haschke, "Sound source localization using a movable human dummy head," M.S. thesis, Saarland University, 2007.

[8] David Talkin, "A robust algorithm for pitch tracking," *Speech Coding and Synthesis*, pp. 495 – 518, 1995.

[9] Christof Faller and Juha Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, 2004.

[10] John Kominek and Alan W Black, "CMU ARCTIC databases for speech synthesis," 2003.