ON THE WINDOW-DISJOINT-ORTHOGONALITY OF SPEECH SOURCES IN REVERBERANT HUMANOID SCENARIOS

Sylvia Schulz and Thorsten Herfet

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

ABSTRACT

Many speech source separation approaches are based on the assumption of orthogonality of speech sources in the time-frequency domain. The target speech source is demixed from the mixture by applying the ideal binary mask to the mixture. The time-frequency orthogonality of speech sources is investigated in detail only for anechoic and artificially mixed speech mixtures. This paper evaluates how the orthogonality of speech sources decreases when using a realistic reverberant humanoid recording setup and indicates strategies to enhance the separation capabilities of algorithms based on ideal binary masks under these conditions. It is shown that the SIR of the target source demixed from the mixture using the ideal binary mask decreases by approximately 3 dB for reverberation times of $T_{60} = 0.6\ s$ opposed to the anechoic scenario. For humanoid setups, the spatial distribution of the sources and the choice of the correct ear channel introduces differences in the SIR of further 3 dB, which leads to specific strategies to choose the best channel for demixing.

1. INTRODUCTION

Rickard et al. [1] showed that speech signals are sparsely distributed in high-resolution time-frequency representations. Time-Frequency (T-F) representations of different speech signals overlap only in few points and so are approximately orthogonal to each other. This approximate orthogonality in the T-F-domain can be used to separate a target source out of a mixture of speech sources by defining T-F-masks that emphasize regions of the T-F-spectrum that are dominated by a specific source and attenuate regions dominated by other sources or noise.

Many speech source separation approaches are based on the assumption of approximate orthogonality of speech sources in the time-frequency domain and utilize T-F-masks to separate the single sources from a mixture (i.e. [1] [2] [3] [4] [5] [6]). Several researchers in computational source separation suggest the ideal binary mask as final goal of computational source separation algorithms (i.e. [1] [2] [3]). Each entry of the T-F-mask is set to one if the target energy in this T-F-bin is greater than the interfering energy. The binary decision is motivated by masking effects of the human auditory system: Within a critical bandwidth humans don't recognize sounds that are masked by louder sounds [7].

The orthogonality of speech sources in the time-frequency domain has been investigated in detail for anechoic speech mixtures (i.e. [1]) and most of the available source separation algorithms are only tested for anechoic and artificially mixed speech mixtures (for example [1] [2] [3] [5] [6]). To be applicable to real world scenarios (i.e. operation of a source separation algorithm as a frontend for an Automatic Speech Recognizer), source separation schemes should be able to operate also in reverberant environments. This paper therefore investigates how the orthogonality of speech sources in the time-frequency domain drops with different reverberation times of the environment and evaluates if separation schemes based on ideal binary T-F-masks can also be successfull under reverberant conditions.

Specific source separation architectures (i.e. [5], [8], [9]) use a humanoid experiment setup to imitate the excellent source separation capabilities of the human auditory system. The speech mixtures are recorded by a human dummy head that performs human-like filtering of the signals by the head and the outer ear structures. The Head-Related-Transfer-Functions (HRTFs) of the left and the right ear filter the incoming signals and disturb them. To find out if source separation schemes relying on the time-frequency orthogonality are also appropriate for such humanoid setups, this paper additionally investigates how the orthogonality of speech sources in the time-frequency domain is affected by the HRTF filtering process.

Realistic humanoid experiment setups record the speech mixtures with a human dummy head under normal reverberant conditions (i.e. [8], [9]). In these scenarios the reverberation and the HRTF filtering affects the orthogonality of the speech sources. This paper also investigates if ideal binary masks are furthermore sufficient to achieve a satisfactory source separation in reverberant environments with a humanoid recording equipment.

Section 2 first introduces the concept of ideal binary masks in the time-frequency domain and defines three values to measure the degree of orthogonality. The influence of reverberation on the orthogonality of two to five source speech mixtures in the Short-Time-Fourier-Transform (STFT) domain is investigated in section 3. Section 4 analyses to which extent the HRTF filtering of the human head and ears affects the orthogonality of speech sources. The last section combines the evaluation of the last two sections and considers the influences of reverberant humanoid recording setups.

2. WINDOW-DISJOINT-ORTHOGONALITY OF SPEECH SOURCES

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

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

An ideal binary mask Ω_i with a threshold x of 0 dB includes all time-frequency points where the energy of source$_i$ is larger than the energy of all other sources in this T-F-bin. An usual goal of source separation architectures is to maximize the Signal-to-Interference Ratio (SIR) while retaining most of the target source's energy. When using the 0-dB ideal binary mask as final goal, T-F-bins that have nearly equal energy from two or more sources are only assigned to one specific source.

Yilmaz et al. [1] investigate the quality of a separated source in the Short-Time-Fourier-Transform (STFT) domain. For a general discrete signal $x(n)$ and an arbitrary discrete analysis window function $w_a(n)$, the STFT is defined $\forall q \in \{0, 1, ..., N-1\}$ as

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

In their paper Yilmaz et. al specify three values to describe the orthogonality of speech sources in the STFT-domain:

1. The Preserved Signal Ratio (PSR) specifies how well the ideal mask preserves the energy of the target source compared to the clean target signal. The PSR is defined as the ratio of the energy of the ideal mask multiplied with the STFT-spectrum of the clean target signal and the energy of the STFT-spectrum of the clean target signal:

   $$PSR = \frac{||\Omega_i(t,f)s_i(t,f)||^2}{||s_i(t,f)||^2}$$

   where $||f(x,y)||^2$ is defined as $\int\int |f(x,y)|^2$. In the best case, when the ideal mask includes all time-frequency points of the target source with energy greater than zero, the PSR approaches one.

2. The Signal to Interference Ratio (SIR) defines how well the ideal mask attenuates the interfering sources. The SIR specifies the ratio of the remaining energy of the target source after multiplying with the ideal mask and the energy of all other sources remaining after multiplying with the ideal target source mask.

   $$SIR = \frac{||\Omega_i(t,f)s_i(t,f)||^2}{||\Omega_i(t,f)\sum_{j\neq i} s_j(t,f)||^2}$$

   High SIR values show that a high percentage of the reconstructed energy belongs to the target source and the interfering sources are suppressed very well. Ideal masks which yield low SIR values include much energy from other sources and so cannot be used to perfectly demix the target source.

3. The orthogonality of different speech sources is estimated by a value called window-disjoint orthogonality (WDO). The WDO is a combined measurement of the PSR and SIR and is specified as the normalized difference between the portion of energy remaining after demixing with the ideal mask and the portion of energy of other sources remaining after demixing:

   $$\begin{aligned} WDO &= \frac{||\Omega_i(t,f)s_i(t,f)||^2 - ||\Omega_i(t,f)\sum_{j\neq i} s_j(t,f)||^2}{||s_i(t,f)||^2} \\ &= PSR - PSR/SIR \end{aligned}$$

   A value of one defines perfect orthogonality in the STFT-domain, a value of zero specifies almost no orthogonality and so only bad demixing results can be achieved with the ideal mask.

3. WDO UNDER SIMULATED REVERBERANT CONDITIONS

Most source separation architectures that are based on the orthogonality of speech sources in the time-frequency domain are only evaluated on anechoic speech mixtures (i.e. [1] [2] [3] [4] [6]). Many practical applications of source separation architectures like i.e. frontends of Automatic Speech Recognizing Systems however require the operation of such systems in non-ideal reverberant environments like inside a car or a crowd. To estimate if ideal binary masks as final goal of source separation approaches are also applicable in reverberant scenarios, this section investigates the influence of reverberation on the window-disjoint orthogonality of speech sources in the time-frequency domain and discusses methods for increasing the SIR gains of source separation architectures based on ideal masks also in reverberant environments.

Figure 1 shows the window-disjoint orthogonality, the preserved signal ratio and the signal-to-interference ratio of mixtures of two to five speech sources for different reverberation times T_{60}. The values are obtained by computing the average values for 20000 speech mixtures of different speakers and 3-seconds duration taken from the speech database CMU Arctic [10]. The mixtures are constructed by adding together the single speech sources. To simulate the reverberation, the mixture files are filtered with Room Impulse Responses defined by FIR filters according to [11].

The influence of reverberation degrades the orthogonality of speech sources in the time-frequency domain. The room impulse responses smear the energy of specific time-frequency bins in time and in frequency and so disturb the sparseness of the speech sources in this way: the overlap of the time-frequency spectra of the single speech sources increases with increasing reverberation time. The SIR decreases for two, three, four and five source scenarios by approximately 3 dB for reverberation times $T_{60} = 0.6\ s$ compared to the anechoic case. If one considers that the maximum SIR gains for anechoic two to five source scenarios lie in the range between 16 dB and 8 dB, a decrease of 3 dB in the SIR substantially influences the quality of the separated target source. The WDO and the PSR decrease analog to the SIR with increasing reverberation time.

The analysis of figure 1 shows that the T-F-spectra of speech sources exhibit an approximate orthogonality, but also overlap in many parts, which leads to low SIR gains of only 17 dB to 8 dB also in the anechoic two to five source mixture scenarios. Source separation algorithms that segregate sources only based on a simple assigning of each T-F-point to a specific source will only achieve these SIR gains in the optimal case. To compensate for the loss in the SIR gain in reverberant environments and to increase the low SIR also in anechoic scenarios, more clever strategies in the computation of the time-frequency spectrum of the target source have to be applied.

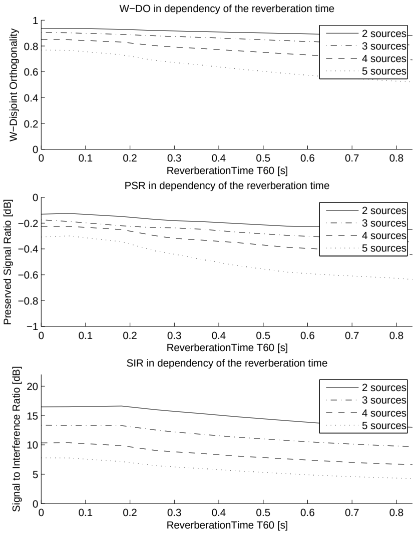

Figure 1: *Window-Disjoint Orthogonality in dependency of the reverberation time T_{60} for scenarios consisting of different sources for 0-dB ideal mask.*

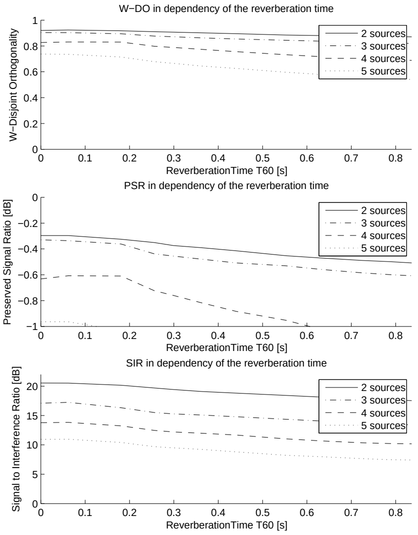

Figure 2: *Window-Disjoint Orthogonality in dependency of the reverberation time T_{60} for scenarios consisting of different sources for 6-dB ideal mask.*

One possible strategy to increase the SIR gain is to divide the separation process in two steps. In a first step, coarse binary masks are estimated that include only the time-frequency points that exhibit large orthogonality. This leads to T-F-masks that achieve high SIR-values, but the preserved signal ratio will be very low, because of the missing signal parts. In a second step these leaky masks are refilled by algorithms that apply specific cognitive models to enhance the PSR while the SIR is kept constant.

To locate the time-frequency areas of the target source spectrum that exhibit large orthogonality, the threshold of the ideal mask computation is adjusted to include only those time-frequency bins where the energy of the target source is x dB larger than the energy of the interfering sources.

$$\Omega_i(t,f) = \begin{cases} 1 & s_i(t,f) - n_j(t,f) > x \quad \forall j \\ 0 & \text{else} \end{cases} \tag{3}$$

Figures 2 and 3 show the WDO, PSR and SIR values for ideal mask thresholds of 6 dB and 9 dB. Compared to the 0-dB mask, the SIR for two source scenarios increases by 4 dB for the 6-dB mask and by 6 dB for the 9-dB mask. For mixtures of five speech sources, the SIR increases analog to the two source scenario by 4 dB respectively 6 dB. The PSR of the 6-dB and 9-dB masks decrease compared to the 0-dB mask by 0.2 respectively 0.3 dB for mixtures of two sources and by 0.7 and 0.9 dB for five source scenarios.

This decrease in the PSR inevitably leads to losses in the quality of the reconstructed target source, as many parts of the original time-frequency spectrum are missing. On the other hand, these masks reconstruct signals that include only few energy of the interfering sources, which is the final goal of source separation architectures. The high SIR values indicate the suitability of the 6-dB and 9-dB masks to extract those parts of the target source, that include only very few energy from interfering sources. To nevertheless achieve a satisfactory sound quality of the target source, the missing signal parts have to be reconstructed by postprocessing algorithms.

Assuming that an arbitrary source separation algorithm has identified the T-F points which exhibit large orthogonality and has estimated the ideal binary mask for a specific threshold (i.e. 6-dB or 9-dB like in the example above) by assigning the identified T-F-bins to each source. This could be accomplished i.e. by assigning only those bins that unambiguously belong to the target source. Possible separation schemes for example rely on the spatial position of the source (for algorithmic examples see i.e. [1], [3], [8]) or the harmonicity of speech signals (i.e. [12], [13], [14]). Then postprocessing algorithms that reconstruct the missing signal parts by either refilling the estimated binary mask or by directly manipulating the T-F-spectrum of the target source have to be applied. These postprocessing schemes include all known or inferred information regarding the sound mixture to identify the missing spectral regions. Possible approaches to refill the binary masks include for example:

Harmonic Analysis Humans tend to use frequencies that are an integer multiple of the fundamental frequency (F0) [7]. If the F0-track (the track of the fundamental frequency over time) of the target speaker is known, this information can be used to identify the higher harmonics and to determine those T-F-regions, where the target source is assumed to have energy. The binary mask of the target source can then be refilled to include all time-frequency points that lie on the harmonics of the F0-track. The F0-track can be estimated based on already identified signal parts or inferred from the complete mixture signal (see i.e. [15] for an approach of reconstructing the F0-track of the target source from the mixture).

On/Offsets Depending on the physics of the source, the start and end times of the spectral components of a speech source in a time-frequency representation are more or less the same [7]. After detecting the on- and offsets of the estimated target source (see i.e. [16] [17]), the mask can be refilled to have consistent start and end times of the segments.

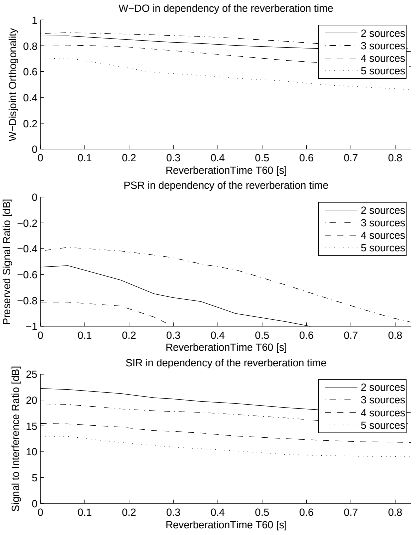

Figure 3: *Window-Disjoint Orthogonality in dependency of the reverberation time T_{60} for scenarios consisting of different sources for 9-dB ideal mask.*

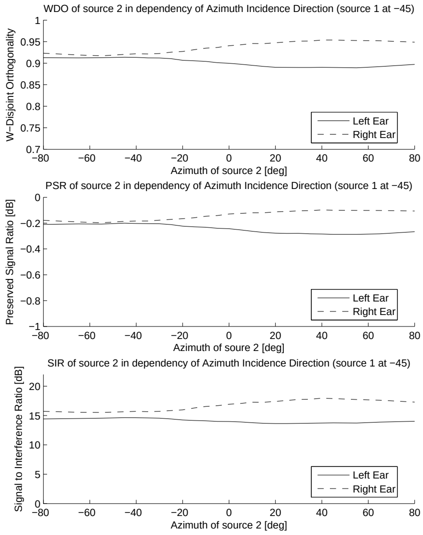

Figure 4: *Window-Disjoint Orthogonality of two speech sources in dependency of the incidence direction for left and right simulated human ear for 0-dB ideal masks (source 1 at -45 degree).*

Instead of refilling the binary mask – which inevitably leads to the inclusion of bins that exhibit only a small orthogonality and so include substantial energy from the interfering sources – postprocessing algorithms can directly manipulate the T-F-spectrum resulting from the demixing of the target source with the estimated binary mask. Demixing the target source from the mixture with a binary mask introduces sharp edges in the spectrum at transitions. This leads to artifacts in the separated source that severely affect the perceptual quality of the reconstructed signal. Postprocessing algorithms can i.e. eliminate sharp edges in the spectrum, by continuing the estimated T-F-segments in time and in frequency with smooth and decreasing coefficients.

By demixing the target source with an estimated binary mask, each T-F-bin is processed separately, but spectral coefficients are not independent of each other due to the time-frequency uncertainty principle [5]. Complete models of human speech production can be employed to describe the relation of T-F-segments among each other in time and frequency. For example the postprocessing algorithms can imitate the distribution of the energy between the fundamental frequency and the harmonics, which should be consistent over short time and frequency intervals.

4. WDO IN SIMULATED HUMANOID CONDITIONS

Specific source separation architectures (i.e. [8], [9], [5]) try to imitate the excellent source separation capabilities of the human auditory system by using a humanoid experiment setup. Auditory scenes consisting of several speech sources coming from different directions are recorded by a human dummy head to simulate the conditions humans experience in real life. Realistic outer ears, pinnae and the shape of the head perfectly imitate a real human head. The pinnae and the outer ear structures filter the incoming signals by specific filter functions – the Head-Related-Transfer Functions (HRTF). The HRTF of each ear disturbs the incoming signals in time and in frequency and so affects the time-frequency spectra of speech signals. This section investigates if the concept of ideal time-frequency masks for source separation is also suitable for such humanoid setups or if the HRTF filtering process disturbs the signals in such a way that the orthogonality of speech sources in the time-frequency domain drops drastically.

Figures 4 – 6 show the WDO, PSR and SIR values for 0-dB ideal masks for two spatially separated sources at different positions in a humanoid scenario. The values are obtained by averaging over 20000 speech mixtures of different speakers of 3 seconds duration taken from the CMU arctic database [10]. The HRTFs used to simulate the humanoid setup and the spatial positions of the sources are taken from the CIPIC HRTF database [18] and have been measured in an anechoic chamber with a KEMAR manikin [19] with small pinnae.

Figure 4 shows the results for a simulated humanoid two source scenario. Source 1 is considered to be fixed at position -45° (negative degree values are assumed to be on the left side regarding the viewing direction of the head, positive degree values on the right side). The target source (source 2) is moving from -80° to 80°. It can be seen that the orthogonality of the speech sources is dependent on the relative positions of the two sources in the auditory scene and on the considered ear. For this scenario the best values in WDO, PSR and SIR for the right ear channel are obtained if source 2 is placed far away from source 1 in the right hemisphere. Then the target source is near the right ear, while the interfering source is on the other side of the head and gets attenuated by the natural head shadow. The signal parts of the target source arrive directly at the right ear without attenuation by the head. This relative increase of the loudness of source 2 compared to source 1 leads to a higher degree of orthogonality of source 2, which is seen by high WDO values for the right ear channel in the figure at positions larger than 40°. For the left ear channel, the orthogonality is best, when the target source is assumed to be on the left side (near the left ear). Because of the interfering source on the left side, the loudness of the two sources is approximately equal and so the WDO, PSR and SIR values are lower than in the right ear channel.

Compared to the anechoic, non-HRTF filtered case of figure 1,

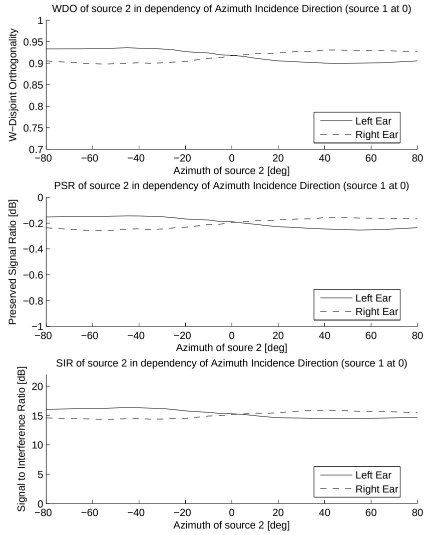

Figure 5: *Window-Disjoint Orthogonality of two speech sources in dependency of the incidence direction for left and right simulated human ear for 0-dB ideal masks (source 1 at 0 degree).*

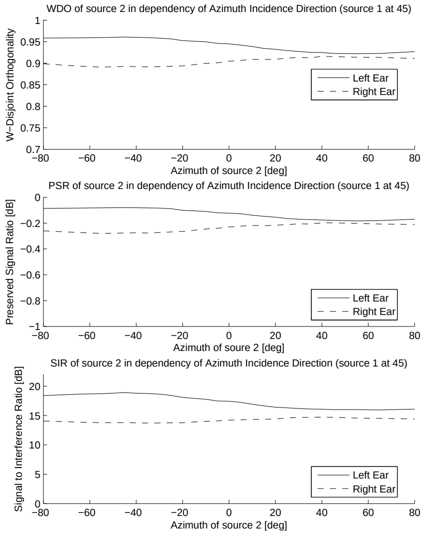

Figure 6: *Window-Disjoint Orthogonality of two speech sources in dependency of the incidence direction for left and right simulated human ear for 0-dB ideal masks (source 1 at 45 degree).*

the WDO decreases by 2 to 10 percent dependent on the source positions. The spatial HRTF filtering leads to SIR gains equal to the anechoic, non-HRTF filtered case for large spatial distances of the two sources (approximately 17 dB), if the better ear is chosen. For spatially nearby sources, SIR gains of only 14 dB can be achieved with ideal binary masks – a decrease of 3 dB compared to the anechoic case. Considering the relatively low maximal SIR gain of 17 dB, a decrease of 3 dB induced by the HRTF surely influences the quality of the separated target source.

Figure 5 evaluates the same scenario for a fixed source at position 0°. The influence of the head shadow can clearly be seen by the run of the WDO, PSR and SIR graphs. The orthogonality of the two speech signals is highest, if the sources are maximally separated in space (in this scenario, when source 2 is assumed to be at positions greater than $\pm 40^\circ$). Then the incidence direction of the target source 2 is more direct than the incidence direction of the fixed source 1 at 0°, which leads to the previously described increase in the loudness of source 2, relative to source 1.

Figure 6 shows the analog evaluation of the scenario, assuming that the fixed source is at position 45°. The WDO, PSR and SIR graphs confirm the conclusions drawn from figures 4 and 5: The orthogonality of speech sources is higher for spatially separated sources than for nearby sources, because of the HRTF filtering that is dependent on the spatial position of the sources.

This evaluation leads to specific strategies that can be applied in humanoid source separation architectures to enhance the separation capabilities. If the spatial positions of the sources in the auditory scene are known in advance, the source separation architecture can choose the better ear – the ear with the largest expected SIR – for demixing the target signal from the mixture. If the current scenario of a static auditory scene recorded by a fixed human dummy head is extended to the dynamic case, where the dummy head and potentially also the sources are able to move, the separation capabilities by ideal-mask algorithms can be enhanced by aligning the dummy head to the currently optimal position regarding the orthogonality and SIR of the target speech source. For an evaluation of the ideal head position for several source separation schemes see i.e. [9].

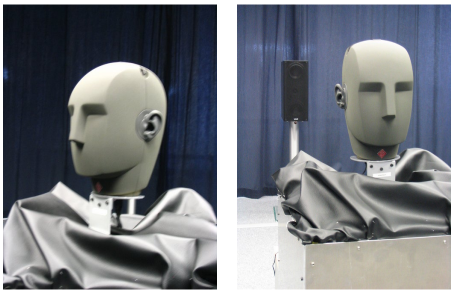

Figure 7: *The human dummy head used for the binaural recordings.*

5. WDO UNDER REAL REVERBERANT HUMANOID CONDITIONS

To examine if the simulations made in the last two sections are also valid in real reverberant humanoid scenarios, this section investigates the concept of the ideal binary mask for binaural recorded speech signals. Five scenarios are accounted to determine the relative decrease of the orthogonality of speech sources in different environments and setups:

Scenario 1 simulates the anechoic case. Speech sources are artificially mixed by adding them together.

Scenario 2 simulates the reverberant case for a normal office room with reverberation time $T_{60} = 0.4\ s$. The room impulse response is generated as described in section 3.

Scenario 3 simulates the HRTF filtering of a humanoid setup with a human dummy head. The HRTFs for the specified positions are generated as described in section 4.

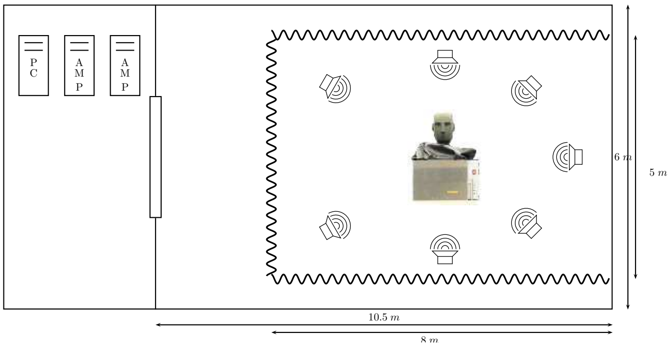

Figure 8: *Layout of the recording room and the adjacent control center.*

Scenario 4 simulates the reverberant humanoid case. The room impulse response of scenario 2 and the HRTF filtering of scenario 3 are combined to reproduce a realistic environment, where a human dummy head is situated in a reverberant office room and listens to speech sources coming from specific directions.

Scenario 5 uses real recordings of a human dummy head in a normal office room to investigate and relate the orthogonality of speech sources in real environments compared to the simulated cases. The speech signals are recorded by a human dummy head in a normal office room (see figure 7). The human dummy head (Neumann KU-100) is positioned in the center of a rectangular room like depicted in figure 8. The recording room is reduced to size $5 \times 8\ m$ by an acoustic curtain and measures a reverberation time $T_{60} = 0.4\ s$. The speech sources are played back by a conventional but high quality 7.1 surround sound system. The recording equipment like microphone amplifiers and the processing computers are placed in a neighboring control room to avoid additional noise.

Figures 9 and 10 show the mean WDO, PSR and SIR values for 200 speech mixtures of 3 seconds duration taken again from the CMU Arctic speech database [10] for speech mixtures of two and three sources and for each of the five described scenarios. For each scenario the same corpus of speech mixtures is used to make the values comparable.

Figure 9 evaluates the orthogonality of speech sources for two source mixtures. The female target speaker is assumed to be at position $0°$. The second male speaker is considered to be at position $-45°$ (in the left hemisphere of the head's viewing direction). The figure shows the WDO, PSR and SIR values for the female target speaker for the five described scenarios. For the simulated reverberant scenario 2, the orthogonality decreases opposed to the anechoic case as described in section 3. Because the second source is considered to be in the left hemisphere, the orthogonality of the right ear channel is slightly higher than in the left ear channel. For scenario 3 and 4 the choice of the correct ear channel increases the SIR gain by approximately 3 dB. For the real binaural recording of scenario 5, the differences between the two ears only amount to approximately 1 dB. Opposed to the anechoic case, the SIR decreases by 5 dB compared to the unprocessed anechoic scenario 1. Nevertheless the choice of the correct ear channel in scenario 5 achieves in the mean of 200 speech mixtures about 1 dB SIR gain.

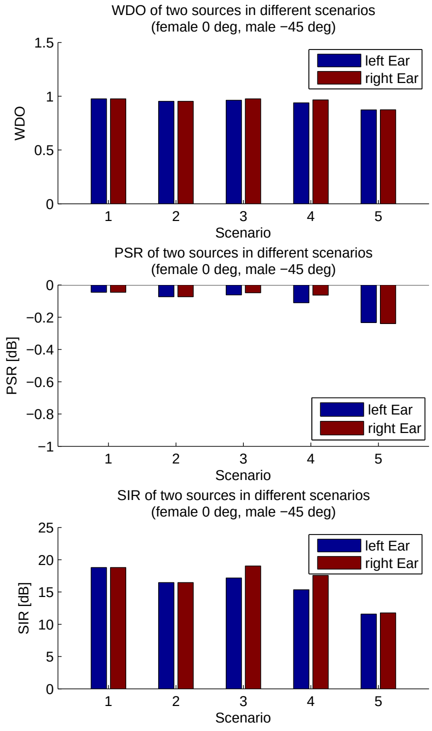

Figure 9: *Window-Disjoint Orthogonality of two speech sources (female at* $0°$ *and male at* $-45°$*) for the five different reverberation scenarios.*

Compared to the simulated reverberant humanoid conditions, the real reverberant humanoid scenario performs about 3 dB worse.

Figure 10 shows the WDO, PSR and SIR values for three source speech mixtures. The female target source is again considered to be at position $0°$. The two interfering male sources emanate at positions $\pm 45°$. The right ear channel shows slightly better WDO and SIR values. If the interfering sources would be equal and so have equal energy at all time instances, than the left and right ear channel should perform equal. In this evaluation the two interferers are two different male speech sources. Due to nature of the specific speech signal, the right ear channel has better SIR and WDO values than the left ear channel. Similar to the two source mixture, the real reverberant scenario 5 looses about 6 dB SIR against the anechoic scenario and about 5 dB against the simulated reverberant humanoid scenario 4.

Source separation methods based on ideal binary masks perform worse in real reverberant humanoid scenarios than in simulated reverberant humanoid scenarios. The decrease in SIR gain between simulated humanoid and real reverberant humanoid is about 3 dB for two source scenarios and about 5 dB for three source scenarios. The real reverberant scenario has a more complex room impulse response than the simulated room impulse response, which leads to stronger perturbations of the energy in the time-frequency spectrum. The simulated room impulse responses only take into a account a limited number of reflections [11] and so also limit the number of disturbed time-frequency bins.

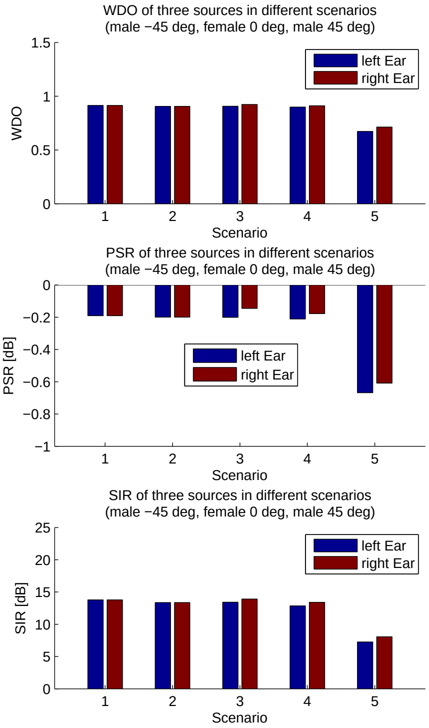

Figure 10: *Window-Disjoint Orthogonality of three speech sources (female at* 0° *and male at* 45° *and* -45°*) for the five different reverberation scenarios.*

6. CONCLUSIONS AND FUTURE WORK

Reverberation and the HRTF filtering of the human head influences the orthogonality of speech sources in the time-frequency domain. The SIR decreases by approximately 5 dB for a two-source humanoid reverberant scenario compared to the anechoic case. For three source scenarios, the decrease in SIR gain amounts up to 6 dB.

To account for the decrease in orthogonality in humanoid reverberant scenario compared to the anechoic case, source separation algorithms based on ideal binary masks have to apply more clever strategies than a simple assigning of the T-F-bins to specific sources. One possible strategy for increasing the separation capabilities is to construct coarse binary masks by localizing those areas in the time-frequency spectrum that exhibit a high degree of orthogonality. Then these coarse binary mask can be refilled by specific cognitive models that include the harmonicity of speech, consistent onsets and offsets or smoothing of the binary masks to avoid sharp edges in the spectrum. In humanoid scenarios the ideal head position and the choice of the correct ear can gain up to 3 dB in SIR.

Future work especially includes the implementation and testing of the proposed guidelines for enhancing the SIR of the separated sources.

7. 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.L Wang, N. Roman, and G. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, pp. 2236 – 2252, 2003.

[4] T. Melia, S. Rickard, and C. Fearon, "Extending the duet blind source separation technique," in *Proceedings of Signal Processing with Adaptative Sparse Structured Representations Workshop (SPARS '05)*, 2005.

[5] Harald Viste, *Binaural Localization and Separation Techniques*, Ph.D. thesis, Swiss Federal Institute of Technology, Lausanne, June 2004.

[6] B. Kollmeier, J. Peissig, and V. Hohmann, "Real-time multiband dynamic compression and noise reduction for binaural hearing aids," *Journal of Rehabilitation Research and Development*, vol. 30 (1), pp. 82–94, 1993.

[7] Albert S. Bregman, *Auditory Scene Analysis – The Perceptual Organization Of Sound*, MIT Press, 1990.

[8] Sylvia Schulz and Thorsten Herfet, "Binaural source separation in non-ideal reverberant environments," in *Proceedings of 10th International Conference on Digital Audio Effects (DAFx-07), Bordeaux, France*, September 2007.

[9] Sylvia Schulz and Thorsten Herfet, "Humanoid separation of speech sources in reverberant environments," in *Proceedings of 3rd International Symposium on Communications, Control and Signal Processing (ISCCSP 2008), St. Julians, Malta*, March 2008.

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

[11] Stephen G. McGovern, "A model for room acoustics," http://www.2pi.us/rir.html, 2004.

[12] G. Hu and D.L.Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, pp. 1135 – 1150, 2004.

[13] N. Roman and D.L Wang, "Pitch-based monaural segregation of reverberant speech," *Journal of the Acoustical Society of America*, vol. 120, pp. 458 – 469, 2006.

[14] M. Wu and D.L.Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 774 – 784, 2006.

[15] Jochen Kraemer, "Multiple fundamental frequency estimation for cognitive source separation," Bachelor's thesis, Saarland University, 2008.

[16] G. Hu and D.L.Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 396 – 405, 2007.

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

[18] Ralph Algazi, Richard O Duda, Dennis M. Thompson, and Carlos Avendano, "The cipic hrtf database," in *WASSAP '01*. 2001 IEEE ASSP Workshop on Application of Signal Processing to Audio and Acoustics, 2001.

[19] M. D. Burkhard and R. M. Sachs, “Anthropometric manikin for acoustic research,” *Journal of the Acoustical Society of America*, vol. 58 (1), pp. 214 – 222, 1974.