A binaural model to segregate sound sources in the presence of early reflections using a multi-source precedence-effect model

Jonas Braasch\(^{(a)}\), Nikhil Deshpande\(^{(b)}\)

\(^{(a)}\)Rensselaer Polytechnic Institute, Troy, United States, braasj@rpi.edu
\(^{(b)}\)Rensselaer Polytechnic Institute, Troy, United States, deshpn@rpi.edu

Abstract

In this paper a binaural model is presented that can segregate two spatialized speech signals in the presence of early reflections and late reverberation. For each sound source, the model identifies the lateral positions for the direct sound component and an early reflection using a precedence effect model. The model then uses a filter to eliminate the reflection of the sound source it wants to extract. Afterwards the equalization/cancellation (EC) method is used to select those time/frequency bins where the binaural cues correspond to the localization cues of the desired sound source. It is shown that the model is sufficiently robust to deal with late reverberation, and it is also shown that the model performs better when the reflections are removed prior to the EC analysis.

Keywords: Sound Source Segregation, Computational Auditory Scene Analysis, Precedence Effect, Echo Cancellation
A binaural model to segregate sound sources in the presence of early reflections using a multi-source precedence-effect model

1 Introduction

The goal of the project presented here was to develop a binaural model that can operate in the presence of early room reflections by compensating for the degrading effects of reflections on the remaining signal. It is well known that the auditory system restricts the influence of early reflections in order to accurately localize a direct source in a room – a phenomenon known as the Precedence Effect (PE) [1] – and that the auditory system reduces the influence of diffuse reverberation on the auditory percept if it is decorrelated between both ears – an effect known as binaural dereverberation. Several computational precedence effect models have been developed to demonstrate how a human listener can accomplish the task of localizing a single direct sound source in the presence of one or multiple distinct room reflections [2, 3]. While it is generally believed that humans can use similar strategies in scenarios with multiple sound sources, functional models that provide strategies of how the auditory system can handle such a task do not exist yet.

Several years ago, Djelani and Blauert made an interesting discovery around the build-up of the precedence effect that will be used as the basis for a multi-source precedence effect model [4]. The build-up of the precedence effect refers to the ability of the auditory system to build up inhibition to suppress the influence of early reflections more effectively. This effect can be demonstrated by presenting a click train from one azimuthal direction and presenting a delayed copy of the click train coming from a second azimuth direction with an amplitude just above the initial detection threshold. After a few repetitions the reflections (the delayed click train) become inaudible – thus the term build-up of the precedence effect was coined. Clifton later claimed that the build-up of the precedence effect breaks down if the directions for the direct sound and its reflections are switched [5], but Djelani and Blauert revised this theory by demonstrating that the precedence effect for the original set of directions was still partially built up. They demonstrated this by reversing the positions for the direct sound and the reflections a second time – back to the original positions. Djelani and Blauert postulated that the auditory system can build up the precedence effects simultaneously for multiple spatial/temporal direct sound/reflection patterns. The sound segregation model presented here continues on this observation. The aim was to develop a sound source-segregation model that can operate under realistic reverberant conditions.

2 Model architecture

The model, depicted in Fig. 1, consists of three main stages: (i) A binaural model to localize reverberant sound sources, (ii) a mechanism to remove early reflections, and (iii) the source-segregation algorithm.
2.1 Localization model

The Binaurally Integrated Cross-correlation/Auto-correlation Mechanism (BICAM) is used to robustly localize a sound source in the presence of multiple reflections. The model also extracts the delays and lateral positions of each distinct reflection. A second-layer cross-correlation algorithm is introduced on top of the first layer autocorrelation/cross-correlation mechanism to determine the interaural time difference (ITD) of the direct sound source component. The ITD is then used to time align two auto-correlation functions obtained from the left and right ear signals to gather information about the reflections and form a binaural activity pattern. The model can accurately localize a speech signal in the presence of two or more early side reflections and additional diffuse reverberation. A detailed description of the model can be found in [6].

2.2 Reflection removal filter

In the following study, we assume that we are dealing with two stationary sound sources with simple broadband ITDs as localization cues. While the BICAM localization model and the source-segregation model can handle stimuli that have been processed with head-related transfer functions (HRTFs), the current implementation of the reflection removal filter cannot process HRTF-based stimuli yet. We further assume that the localization model can localize each of the two sound sources in isolation from one another and determine the values of the reflection amplitudes and delays for the left and right ear signals. In this study, each sound source has one early reflection, but the reflection parameters are different for both sound sources.

The early reflection is removed from the total signal, prior to the Source Segregation Algorithm as shown in Fig. 1. The filter design was taken from an earlier precedence effect model [3]. The filter takes values of the delay between the left or right channels of the direct signal and
the reflection within the same channel, \( T_{f/r} \), and the amplitude ratio between direct signal and reflection \( r_{f/r} \), which can be estimated by the BICAM localization algorithm that was described in Section 2.1 or alternatively by the precedence-effect model [3]. The lag-removal filter, \( h_d \), can eliminate the lag from the total signal:

\[
h_d(t) = \sum_{n=0}^{N} (-r)^n \delta(t - nT).
\]

It converges promptly and only a few filter coefficients, \( N \), are needed to remove the lag signal effectively from the total signal.

Instead of removing the early reflections independently for both sound sources, the reflection removal filter with the parameters for Source1 was run over the total signal to isolate only Source1, and the reflection removal filter was run with the parameters for Source2 over the total signal to isolate Source2.

Figure 2: EC-model calculations for a broadband target/masker pair for one auditory band centered at 750 Hz. The left graph shows the target-only data. Note that the target location is at the lowest point close to zero (blue zone) because the EC model was able to compensate for the target signal in this case, and no residuals remain other than internal-noise induced artifacts. The center graph shows the results for the combined target/masker presentation. Note that in this case, the signal can no longer be fully canceled out, because the EC process can only eliminate one signal at a time. The lowest point (0.15 model units [MU]) is positioned in between the two locations of target and masker – the data for the latter are shown in the right graph.

2.3 Source segregation model based on an equalization/cancellation process

To create the cue-selection map, the left and right audio channels are sent through a Gamma-tone bandpass filterbank with 36 channels [7] and then segmented in time using a 512-point Hanning window with a step size of half the window length. Durlach’s Equalization/Cancellation (EC) model [8] was used to group the analyzed time/frequency segment to individual sources. This method was found to be more effective than Faller and Merimaa’s method [9]. The EC
method uses a null-antenna approach, considering that the lobe of the 2-channel sensor that
the two ears represent is much more effective at rejecting a signal than filtering one out. In
previous literature, the EC model is mainly used to explain the detection of masked signals. It
assumes that the auditory system has mechanisms to cancel the influence of the masker by
equalizing the left and right ear signals to the properties of the masker and then subtracting
one channel from the other. Information about the target signal is obtained from what remains
after the subtraction. For the equalization process, it is assumed that the masker is spatially
characterized by interaural time and level differences. The two ear signals are then aligned
in time and amplitude to compensate for these two interaural differences. The model can be
extended to handle variations in time and frequency across different frequency bands. Internal
noise in the form of time and amplitude jitter is used to degrade the equalization process to
match human performance in detecting masked signals.

Figure 2 illustrates how this is achieved using the data in an auditory band with a center fre-
quency of 750 Hz. For each graph all possible ITD/ILD-equalization parameters were calculated
using the method by [10], and the data for each bin shows the residual of the EC amplitude
after the cancellation process. A magnitude close to zero (dark blue in the color map) means
that the signal was successfully eliminated, because at this location the true signal values for
the ITD (shown in the horizontal) and ILD were found (shown in the vertical axis). This is
only possible for the left graph which shows the case of an isolated target, and the right graph
which shows the case of the isolated masker. In case of overlapping target and masker signals,
shown in the center panel, a successful cancellation process is no longer possible because the
EC model cannot simultaneously compensate for two signals with different ILD and ITD cues.
As a consequence, the lowest point with a coherence value of 0.15 (aquamarine color) is no
longer sufficiently close to zero, and thus the magnitude of the lowest coherence point can be
used as an indicator of the presence of at least two overlapping signals in this time/frequency
bin. Our model uses the one-signal bins, groups them according different spatial locations, and
integrates over similar ITD/ILD combination to determine the positions of masker and target.
The cue selection parameter $b$ is estimated as follows,

$$b(n,m) = \frac{\sqrt{\sum (x_1(n,m) - x_2(n,m))^2}}{E(n,m)},$$  \hspace{1cm} (2)$$

with the left and right audio signals $x_1(n,m)$ and $x_2(n,m)$, and the root-mean-square energy for
the left and right audio signal for the frequency band $n$ and the time bin $m$. The cue is then
plotted as:

$$B = \max(b) - b,$$  \hspace{1cm} (3)$$
to normalize the selection cue between 0 (not selected) and 1 (selected). For the creation of
ideal maps, the threshold for $B$ was set to 0.75 to select cues.
3 Model evaluation

3.1 Test stimuli

The examples shown in these sections were created using speech stimuli from the “Music for Archimedes” CD with anechoic recordings. A female and male voice at a sampling frequency of 44.1 kHz were mixed together such that the male voice was heard for the first half second, the female voice for the second half second, and both voices were concurrent during the last 1.5 seconds. The female voice said: “Infinitely many numbers can be com(posed),” while the male voice said: “As in four, score and seven”.

For simplicity, the female voice was spatialized to the left with an ITD of 0.45 ms, and the male voice to the right with an ITD of −0.27 ms. In some examples, both sound sources (female and male voice) contain an early reflection. The reflection of the female voice is delayed by 1.8 ms with an ITD of −0.36 ms, and the reflection of the male voice is delayed by 2.7 ms with an ITD of 0.54 ms. The amplitude of each reflection is attenuated to 80% of the amplitude of the direct sound.

For the examples that included a reverberation tail, the tail was computed from octave-filtered Gaussian noise signals windowed with exponential decay set for individual reverberation times in each octave band. Afterwards, the octave-filtered signals were added together for a broadband signal. Independent noise signals were used as a basis for the left and right channels and for the two voices. In our example, the reverberation time was 1s, uniform across all frequencies with a direct to late-reverberation ratio of 0 dB.

3.2 Results for stimuli in the presence of diffuse reverberation

In the next example, the EC model will be used to determine areas in the joint time/frequency space that contain isolated target and masker components. In contrast to Fig. 2, the EC analysis is reduced to different ITD combinations and uses the second depicted dimension for time analysis instead of the ILD. Figure 3 shows the results for the EC-selection mechanism in the condition where the male voice is extracted. The top-left graph shows that the selected cues (red areas; the color bar identifies the color code for the values estimated according to Eq. 3) correlate well with the male voice signal shown in the sub panel below the graph in blue, but not with the female voice shown in red.

While the model also selects residual information from the female voice, most bins corresponding to the female voice are not selected (blue color). The top-right graph show the binary mask that was computed from the left graph using a threshold of 0.75. The white tiles represent the selected time/frequency bins corresponding to the red areas in the left graph. The sub panel of the top-right graph shows the time series of the total reverberant signal (green curve, male and female voices plus reverberation), the isolated anechoic voice signal (red curve) and the signal that was extracted from the mixture using the EC model (blue curve). In general, the model is able to both perform the task of isolation while noticeably removing the reverberation tail. The model still has issues with the onset of signals, presumably because the reverberation tail from the previous signal components interfere with the new signal components and decor-
Figure 3: Signal-selection maps (left columns) and ideal masks (right columns) based on EC analysis to extract a male-voice signal (top row) and a female-voice signal (bottom row) in the presence of diffuse reverberation.

relate the overall signal. It also has problems processing segments with overlapping voices – for time/frequency bins where the female voice dominates or both signals are equally loud.

The bottom-left panel of Fig. 3 shows the same data as the top-left graph, but this time, the EC algorithm targeted the parameters of the female voice. Now the algorithm primarily selects the time/frequency bins that corresponds to the female voice, while correctly rejecting those that belong to the male voice. For both the male- and female-voice extraction example, it is apparent that the algorithm currently performs much better for the mid-range and upper frequencies of the isolated signals parts. A fairly high percentage of time/frequency bins are missed when both signals overlap in the 1.0–1.5-s time range, while at low frequencies time/frequency bins are often over-selected.
Figure 4: EC analysis (left graphs) and binary maps (right panels) for the extraction of the male-voice signal from a female/male-voice sound mixture that contains early reflections, but no reverberant tail. The top graphs show the condition in which the early reflections were not removed prior to the EC analysis; the center row shows the data where the early-reflection for the male-voice signal was removed. The bottom graphs show the process of extracting the female-voice signal by using the correct early reflection removal filter for the female voice.
3.3 Results for stimuli with early reflections

Next, we studied how the algorithm can handle the removal of early reflections. For this purpose, we examined the test stimuli with early reflections as specified in the test stimuli section (Sect. 3.1), but without a late reverberation tail.

Figure 4 shows the results of the procedure for the extraction of the male voice (top and center row) and female voice (bottom row). The method of data representation is identical to the one applied to Fig. 3, with the exception that in the previous figure a late reverberation tail was applied with no early reflections. The top-left panel shows the test condition in which the early reflection of the male voice was not removed prior to the EC analysis. The analysis is very faulty; in particular, the signal is not correctly detected in several frequency bands, especially ERB bands #6 to #11 (220–540 Hz). At low frequencies – bands #1 to #4 – a signal is always detected, and the female voice is no longer rejected. Consequently, the binary maps contain significant errors at the specified frequencies (top right graph), and the reconstructed male-voice signal does not correlate well with the original signal (compare the blue curve in the sub-panel of the top-right figure to the blue curve in the sub-panel of the top-left figure).

The two graphs in the center row of Fig. 4 show the condition in which the filter described in Eq. 1 was applied to the total signal to remove the early reflection for the male voice. Note that the female voice signal is also affected by the filter, but in this case the filter coefficients do not match the settings of its early reflection, because both the female and male voices have early reflection with different spatial properties, as would be observed in a natural condition. Consequently, the filter will alter the female-voice signal in some way, but not systematically remove its early reflection. Since this signal is treated as background noise for now, its properties can be altered without worry as long as the signal characteristics of the male-voice signal can be improved. As the left graph of the center row indicates, the identification of the time/frequency bins containing the male-voice signal works better compared to the previous condition where no lag was removed – see Fig. 4 top-left panel. Note particularly the solid red block in the beginning, where the male-voice signal is presented in isolation. This translates into a much more accurate binary map as shown in the right graph of the center row. In this case, the extracted male voice signal (blue curve in sub-panel below) is more accurate than was the case in the last condition. Note that the amplitude of the extracted signal is lower that the male signal component of the total signal (green curve), but this is partially, because the green curve contains reflections and thus greater overall energy, while the extracted signal does not. It is important to emphasize that the application of the lag-removal filter with male-voice settings does not prevent the correct rejection of the female-voice signal. Only in a very few instances does the model select a time-frequency bin with the female voice-only region (0.5–1.0 seconds). The algorithm now also does a much better job at extracting the male voice signal from the mixture (1.0–2.5 seconds) than when no lag-removal filter was applied (compare top-right graph of the same figure). The bottom row of Fig. 4 shows how the lag-removal method works out for the female-voice extraction process. Also for this case, the female voice can be extracted when the correct reflection-removal filter is applied.
4 Conclusion and next steps

We presented a source-segregation model based on an equalization/cancellation model to separate a male and a female-voice signal from a binaural mixture. The model is robust enough to process reverberant signals. A reflection removal algorithm based on a precedence-effect model and a reflection removal filter were introduced; they show substantial improvement for the source segregation model if early reflections are correctly removed before the EC analysis. Currently, the model utilizes simple ITDs as localization cues, but we plan to extend the model for use with measured HRTFs based on an EC algorithm that has been introduced in another presentation at this conference (ICA 2016 Paper #818).

Acknowledgments

This material is based upon work supported by the National Science Foundation (IIS-1320059). This project is also affiliated with the European Research Council Project FP7-ICT-2013-C-#618075, www.twoears.eu). We would like to thank Jens Blauert for providing useful feedback on an earlier version of this manuscript.

References


