ON THE USE OF SPATIAL CUES TO IMPROVE BINAURAL SOURCE SEPARATION

Harald Viste
Audiovisual Communications Lab
Swiss Federal Institute of Technology Lausanne
Switzerland
harald.viste@epfl.ch

Gianpaolo Evangelista
Dept. of Physical Sciences
University of Naples “Federico II”
Italy
gianpaolo.evangelista@na.infn.it

ABSTRACT
Motivated by the human hearing sense we devise a computational model suitable for the localization of many sources in stereo signals, and apply this to the separation of sound sources. The method employs spatial cues in order to resolve high-frequency phase ambiguities. More specifically we use relationships between the short time Fourier transforms (STFT) of the two signals in order to estimate the two most important spatial cues, namely time differences (TD) and level differences (LD) between the sensors. By using models of both free field wave propagation and head related transfer functions (HRTF), these cues are combined to form estimates of spatial parameters such as the directions of arrival (DOA). The theory is validated with the help of the experimental results presented in the paper.

1. INTRODUCTION
In life we are more or less constantly exposed to a multitude of different sound sources. In any given situation only a few of these sources are normally important, and the rest of the sources are not considered to contain relevant information. These irrelevant sources can be regarded as noise. Still, in the presence of many such noise sources, we are able to communicate and convey information by means of transmission of sound waves.

One may speculate that these capabilities of the human hearing sense are the results of the evolution of basic survival techniques that have taken place since the dawn of creation. In any case, it is clear that we are very well capable of localizing sources in the auditory space, range these according to importance, and focus on the most important sources while disregarding the rest. Actually this is a threefold problem, consisting of the localization, recognition, and separation of sound sources. Surely, these different aspects are not necessarily independent, and may be strongly interrelated. Motivated by the performance of the human hearing sense researchers have studied these fields for decades.

There is a vast literature on the many psychophysical and psychoacoustical experiments that have been performed, aiming at describing various aspects of the human binaural hearing and source localization in particular. A nice overview of this field is given in [1]. Based on some of these results, several computational models of the human auditory processing for source localization have been proposed [2, 3, 4, 5]. In general these models aim at mimicking and explaining the processing that takes place in the auditory system, and are not necessarily easy to exploit in more general applications.

The areas of source recognition and source separation have also been subject to studies. However, in these fields the focus has been on purely computational models, as opposed to psychophysical and psychoacoustical models. This is most likely due to the increased complexity such models would introduce, by the need to include high-level psychological principles such as cognition and anticipation, among others.

Such computational models are better suited for general applications. They include purely statistical/theoretical models such as blind source separation techniques [6], mathematical techniques based on simple directionality cues such as beamforming techniques [7] and the DUET method [8], as well as techniques based on more heuristic psychoacoustics and sinusoidal models [9, 10, 11, 12].

From the field of binaural hearing it is well known that the principal cues for localization of sources are the interaural time differences (ITD) and interaural level differences (ILD). Traditionally, the ITDs have been emphasized at low frequencies, and the ILDs at higher frequencies, which is also known as the duplex theory. This is a quite rude simplification, and in reality there are more complex interactions [13].

In this paper we present a basic model for localization of sound sources and its application to source separation. Motivated by the human binaural hearing we use level differences (LD) and time differences (TD) between the two sensor signals in order to localize the sources. These LDs and TDs are simple mathematical estimates based on the short time Fourier transforms (STFT) of the sensor signals. By using physical models of sound wave propagation we find simple relations between the LDs and TDs. In other words these two cues tell the same story. We devise a method for combined evaluation of the TDs and LDs. More specifically, we use the LDs in order to resolve phase ambiguities in the TDs at higher frequencies.

Even though we use some principles of binaural hearing and draw knowledge from this field, we emphasize the fact that our model is purely computational, and in no way tries to compete with any existing psychophysical models such as those referenced above. Still, we interestingly observe that for certain aspects our computational model gives qualitatively the same results as those found in psychoacoustical experiments. However, we will not speculate in this.

The organization of this paper is as follows. In section 2 we describe the simple physical models of sound wave propagation, both in free field and around the head. Then section 3 discusses the estimation of LD and TD cues and how they can be jointly evaluated, as well as experimental data. This is followed by section 4 on the application of our binaural localization model to sound source separation, as well as some experimental data. Finally, in section 5 we draw the conclusions.
2. PHYSICAL MODELS

In most real situations the TDs and LDs between two sensors are strongly interrelated. This is a natural consequence of the physics of wave propagation. In other words, a signal that arrives first at one sensor is also likely to be strongest at that sensor. We develop some simple models for these relations, first in the free field case, and then in the binaural case (the human head). In both cases we use “intersensor” polar coordinates, i.e. polar coordinates relative to the axis passing through the two sensors. Positions are given by \((\theta, \phi, \rho)\), where the elevation \(\phi\) is the rotation around this axis, the azimuth \(\theta\) is the angle from the median plane towards this axis, and \(\rho\) is the distance.

2.1. Free field

We consider a two-sensor setup, where the sensors are placed at a distance \(a\) to the left \((-\frac{a}{2}, 0, a)\) and to the right \((\frac{a}{2}, 0, a)\) of the origin. We assume that the distance \(\rho\) to any source is much larger than the distance between the two sensors, \(\rho \gg a\). In this case, the azimuths from the sensors to a source at \((\theta, \phi, \rho)\) are approximately the same \(\theta_1 \approx \theta \approx \theta_2\), and both sensors will be approximately the distance \(\Delta \rho = a \sin \theta\) closer to or farther from the source origin, \(\rho_1 = \rho + \Delta \rho\), \(\rho_2 = \rho - \Delta \rho\). The time difference between the two sensors is

\[
\Delta T = \frac{\rho_1 - \rho_2}{c} \approx \frac{2a \rho}{c} = \frac{2a \sin \theta}{c} \tag{1}
\]

where \(c\) is the wave propagation speed.

The sound intensity is inversely proportional to the square of the distance, so the sound intensity level difference (in dB) between the sources equals

\[
\Delta L \approx \log \frac{\rho_1}{\rho_2} \approx \log \frac{\rho + \Delta \rho}{\rho - \Delta \rho} \approx \sum_{n=0}^{\infty} \frac{1}{2n + 1} \left( \frac{\Delta \rho}{\rho} \right)^{2n+1} \tag{2}
\]

Clearly, when \(\Delta \rho \ll \rho\), the first order truncation of this series is a good approximation. It follows that the LDs (in dB) and the TDs are linearly related.

2.2. Head related transfer function

For applications such as in hearing aids, the sensors will typically be placed on each side of the head. In this case, the waves will propagate around the head, and a better model of the interaural time difference (ITD) is given by Woodworth's formula:

\[
\Delta T(\theta) \approx \frac{a (\sin \theta + \theta)}{c} \tag{3}
\]

The interaural level difference (ILD) is much more complex, and varies from person to person. Notably, due to head and ear shadowing effects, it depends on frequency and azimuth, and to some extent on the elevation. However, a closer inspection of different HRTFs in the CIPIC database [14] shows that the ILD (in dB) as function of azimuth can be crudely described by the following model:

\[
\Delta L(\theta, f) \approx \alpha_f \sin \theta \tag{4}
\]

where \(\alpha_f\) is a frequency dependent scaling factor. This sinusoidal model corresponds with the results of experiments on qualitative assessment of lateral localization based on ILD, as described in [1].

Figure 1 shows the estimated ILDs for one particular HRTF in the CIPIC database. The leftmost diagram shows the total ILDs (over all frequencies) as function of azimuth, and for elevations between \(-45^\circ\) and \(45^\circ\). The diagram in the middle shows the ILDs as function of azimuth (in the horizontal plane) for different frequencies. To better show the conformance to (4), these curves have been normalized (\(\alpha_f \approx 2\)). Finally, in the rightmost diagram is shown the model in (4). Clearly, this simple model is a good fit to the overall ILDs (left). For the frequency dependent ILDs (middle) the general model is still a reasonable match to the data, especially for azimuths in the range between \(-50^\circ\) and \(50^\circ\). For larger azimuths the variance of the model error is much larger. Consequently, the azimuth resolution in our method becomes coarser as the azimuth approaches the extremes (\(\pm 90^\circ\)). Qualitatively, this is comparable to the azimuth resolution obtained from pure ILDs in psychoacoustical experiments [15].

As previously mentioned, the ILDs depend on frequency. In other words the maximum ILD (at about \(\pm 90^\circ\) azimuth) is approximately \(\alpha_f\) dB. Without any knowledge about the frequency dependent scaling factor \(\alpha_f\) in (4), we can not relate ILDs and ITDs to each other. The estimated \(\alpha_f\) as a function of frequency is plotted in grey in figure 2 for each of the 45 subjects in the CIPIC database. The mean is plotted in black. We notice that for low frequencies, the scaling factor is very small (starting at about 0 dB), and in the range 3-7 kHz it is almost linear with small variance between the subjects. In the other ranges (1-3 kHz and above 7 kHz) the variance between the subjects is much bigger, and the curves are in general more complex.

3. SOURCE LOCALIZATION

Normally, the auditory space of most interest is the horizontal plane. For simplicity we therefore assume that all sources are located in the horizontal plane and in front. Consequently, we disregard any dependence on elevation and avoid all front-back confusions. Naturally, these aspects could be taken into account, e.g. by using additional cues such as head movements, envelope delays, spectral cues, etc. However, in that case our model would lose its simplicity. Under these assumptions, the problem of localizing sources is equivalent to estimating their azimuths.

For the stationary part of narrow band signals the ITDs can
only be estimated from the phase differences between the signals at the two sensors. Clearly, the phase is only correct up to an integer multiple of $2\pi$. An average head is about 15 cm thick. Using (3), the extreme ITDs are then approximately $\pm 0.55$ ms. For frequencies below 900 Hz, the distance a wave propagates during this time is less than half the wavelength. This means that below this frequency there is no phase ambiguity in the actual range of ITDs. For higher frequencies the phase is ambiguous. If the auditory horizon is restricted to $90^\circ$, i.e., if all sources lie between $-45^\circ$ and $45^\circ$ azimuth, then the corresponding frequency is approximately 1500 Hz.

Quite recently, the short-time Fourier transform (STFT) has been proposed for localization [16] and separation [8] of many sources in a two-channel mixture. From the STFTs $S_L(t, f)$ and $S_R(t, f)$ of the left and right input signals, one can compute the level differences 

$$\Delta L(t, f) = \left| \frac{S_L(t, f)}{S_R(t, f)} \right|$$

(5)

and phase delays

$$\Delta T(t, f) = \frac{1}{f} \Delta P(t, f)$$

(6)

as functions of time and frequency. Here $\Delta P(t, f) = \angle S_L(t, f) / \angle S_R(t, f)$ is the phase difference between the two sensor signals, also as function of time and frequency. We use (5) as estimates for the ILDs, and (6) as estimates for the ITDs.

Using (5) and (4) we estimate the directions of arrival from the level differences exclusively

$$\theta_L(t, f) = \sin^{-1} \left( \frac{\Delta L(t, f)}{\alpha_f} \right)$$

(7)

Inserting these estimates in (3) we get the time delays between the sensors estimated from level differences only

$$\tau_L(t, f) = \Delta T(\theta_L)$$

(8)

Similarly, the possible time delays estimated from the phase differences (6) only are given by

$$\tau_T(t, f) = \Delta T(t, f) + \frac{1}{2} 2\pi k$$

(9)

where the latter term is the phase ambiguity. For each time and frequency we choose the $k$ in (9) that gives $\tau_T(t, f)$ as close as possible to $\tau_L(t, f)$, and use the $\tau_T(t, f)$ with this $k$ as our final estimate $\tau(t, f)$ of the true delay between the sensors.

Effectively, we have applied the level differences $\Delta L$ in order to resolve the ambiguities in phase delays $\Delta T$. We notice that at low frequencies there is no phase ambiguity, so the $\Delta T$s are used exclusively for the localization. At high frequencies, the wavelengths are so short that the phase delay contains virtually no information at all, and consequently the $\Delta L$s are dominant. This is in correspondence with the duplex theory. In addition we also handle the transition between these two extreme cases gracefully.

Figure 2: Proportionality factor in our ILD model as function of frequency.
is the error between the individual $\alpha_f$ and the mean that we use. Second there is the deviation from the perfect sinusoidal shape of our model (4) for each frequency, as shown in the middle of fig. 1.

Due to these imperfections, the results do not always look as nice as in fig. 3. Still, the method gives overall valuable results. Figure 4 shows energy weighted histograms (in dB) for a mixture containing 4 sources, as functions of the delay between the sensors. The sources are located at $-45^\circ$, $-20^\circ$, $0^\circ$, and $30^\circ$ azimuth, respectively. (The corresponding delays are approximately $-17$, $-8$, $0$, and $11$ samples.) At the top is shown the histogram over the entire signal (all time and frequency bins). The grey line shows the result of the original method without phase correction, and the black line shows the result after phase correction. We see how some of the false peaks (delays about 4 and about 18) have been attenuated, and how some of the true peaks (delays about $-17$) have been enforced with our phase-correction method. In the bottom graph is shown the same case, but only for frequencies above 2 kHz. Without phase correction, the maximum delays are in the range between $-10$ and 10 samples (about half the longest wavelength). Strangely, when the phase has been corrected, the 4 peaks corresponding to the 4 source locations are even sharper than in the top graph. Part of the reason for this is that for low frequencies (below 1500 Hz), the actual ITDs of measured HRTFs are slightly larger than the model [13]. Since we did not account for this, and the signals contain significant energy at the lower frequencies, we see how the peaks are stretched out towards the extremes.

In binaural hearing the ITDs are estimated in critical bands independently. This means that for broadband signals the phase ambiguity is less problematic than in our model, which is based on narrow frequency bands in the STFT. In [13] it is argued that the ILDs are likely to be useful localization cues in individual narrow frequency bands. On the other hand there are experiments that indicate that some kind of across frequency processing takes place in the auditory processing [18]. We reemphasize that our model is purely computational, and does not try to mimic or explain aspects of binaural hearing.

This information can be applied in many of the existing source separation methods, such as e.g. methods based on sinusoidal modeling in order to improve the analysis and separation quality.

Methods based on localization cues have shown promising results for separation of sound sources, DUET [8]. Briefly explained, the original DUET method separates the signals by assigning each of the time-frequency bins in the STFT to one of the sources exclusively, based on the spatial cues. However, the original method assumes that there are no phase ambiguities, i.e. that the sensors are spaced closely enough to avoid phase ambiguities. For CD-quality audio (44100 Hz sampling rate) this corresponds to a maximum sensor spacing of less than 1 cm. Clearly, in this case the level differences between the sensors are virtually useless. In addition, the experiments shown in [8] are also flawed since several of the corresponding level and time differences used in the mixing model are highly contradictory.

We have applied our phase-correction technique in order to allow larger sensor spacing, and compared the separation result with that of the original DUET method. Figure 5 shows the spectrograms for the 4 separated sources in our example (after applying the binary masks). The left column shows the results obtained with the original DUET method. The right column shows the results with our method.

Naturally, the original method does not perform well, since the assumptions on which it is based do not hold. This method is not at all applicable to binaural signals (about 15 cm sensor spacing). This is therefore not a fair comparison of two methods. Rather the

4. SOURCE SEPARATION

The energy weighted histograms in section 3 provide convenient means for detecting the number of sources and their locations.
original method is included to visualize how our method performs, and how it can be used to extend the original method.

In the left column we see how the DUET method breaks down due to the phase ambiguities. As the frequency increases, the maximum detectable phase delay decreases, centered around 0°. Eventually, for high enough frequencies, all the time-frequency bins are assigned to the third source (since this is in the center at 0°).

With our method we see how the phase ambiguities have been resolved. All the sources contain partials in the entire frequency range. However, when several sources have overlapping energies, the spatial cues are corrupted. This is most important when strong partials overlap, and less important for overlapping in the weaker sidebands. The spatial cues estimated for time-frequency bins are consistent in these cases, but the strong partials of the center are assigned more correctly.

Listening tests reveal that the sources are well separated with little crosstalk. Still there are some distorting effects. These are mainly due to the overlapping partials, and the “on/off”-effect of the binary masks used in the DUET method. These problems have been studied earlier in [19].

5. CONCLUSIONS

We have presented simple models of wave propagation that provide us with relationships between the TDs and LDs. By jointly evaluating these cues, one can resolve phase ambiguities, effectively improving the localization and separation of sound sources. From psychoacoustic studies on source localization [1], it is well known that for low frequencies the ITD is dominant, whereas for high frequencies the ILD is most significant. Our model can be used to study the relative importance of these cues in computational models, in the transition from low to high frequencies (about 1-3 kHz).

6. REFERENCES


