Modeling the Precedence Effect for Binaural Sound Source Localization in Noisy and Echoic Environments

Martin Heckmann[1], *Tobias Rodemann*[1], *Björn Schölling*[1,2], *Frank Joublin*[1], *Christian Goerick*[1]

[1]Honda Research Institute Europe GmbH
Carl-Legien-Strasse 30, D-63073 Offenbach/Main, Germany
{martin.heckmann, tobias.rodemann, frank.joublin, christian.goerick}@honda-ri.de

[2]Darmstadt University of Technology, Institute for Automatic Control, Control Theory and Robotics Lab
Landgraf-Georg-Str. 4, D-64283 Darmstadt, Germany
bjoern.schoelling@rtr.tu-darmstadt.de

Abstract

We present a new way of modelling the Precedence Effect to enable the robust measurement of localization cues (ITD and IID) in echoic environments. Based on this we developed a localization system which is inspired by the auditory system of mammals. It uses a Gammatone filter bank for preprocessing and extracts the ITD cue via zero crossings (IID calculation is straight forward). The mapping between the cue values and the different angles is learned offline which facilitates the adaptation to different head geometries. The performance of the system is demonstrated by localization results for two simultaneous speakers and the mixture of a speaker, music, and fan noise in a normal meeting room. A real-time demonstrator of the system is presented in [1].

Index Terms: sound source localization, binaural, precedence effect, reverberant, echoic.

1. Introduction

In real world scenarios noise and echoes are ubiquitous and make sound source localization on a robot a difficult task. Most systems for source localization are based on an autocorrelation. In order to deal with echoes they perform a weighting of the correlation function [2] or select measures based on a reliability criterion [3, 4]. A different approach to overcome the echoes is inspired by psychoacoustics, more precisely the *Precedence Effect*, and only uses the onsets of the signals to measure the localization cues[3]. Since the task gets easier as the number of microphones and their distance is increased a multitude of systems uses arrays of microphones[5]. For sound source localization on a robot like Asimo the dimensions of the robot restrict the size of the array and therefore make the problem more difficult. Furthermore biological systems are still far better in localizing sound sources in noisy environments than technical systems and therefore better performance for technical systems which try to understand and implement solutions found in biology can be expected. For these reasons we are investigating binaural source localization. The number of systems performing binaural localization is much more limited [6, 7] especially of those which work in echoic environments [8].

Binaural systems commonly work in the frequency domain (either via FFT or as in our case by using a band pass filterbank) and use the following cues:

Interaural Time Difference (ITD): The time delay between the left and right signal.

Interaural Intensity/Level Difference (IID/ILD): The intensity difference between the left and right signal.

These cues are known to be also responsible for the sound source localization capabilities of humans [9].

In the following we will first detail our echo suppression mechanism based on the Precedence Effect which enables robust measurements in echoic environments. Then we introduce our basic localization system and finally present some results.

2. Modeling the precedence effect

It is known that the Precedence Effect makes localization in echoic environments possible for humans. The main findings relevant for our modeling are that a leading sound suppresses localization of a shortly following sound ($\approx 40\,\mathrm{ms}$) and that a lagging sound sufficiently more intense than the leading sound ($10 - 15\,\mathrm{dB}$) overwrites the precedence effect [9]. The most basic model is to perform the localization only in the onsets of a signal and inhibit following onsets for a fixed time span determined a priori [3]. The motivation behind this is that with the onsets only the direct path is captured and the measurement is stopped when the echoes arrive and hence implements the first aspect of the precedence effect. In our model we also included the second aspect that a loud signal triggers again the measurement process even if the inhibition time is not over. Additionally we changed the measurement point and do not use the onsets of the signal but the maxima. A first reason for doing so is that the onsets are difficult to determine reliably and a threshold is necessary to make the decision if the current rise in energy is really an onset or just noise. Secondly we made the observation that the cues used for localization are rather unstable at the onsets, stabilize until the maximum and then are affected by the echoes in the part after the maximum. The cues in smaller maxima following a maximum at the signal onset are dominated by the echoes. Therefore we implemented an inhibition of shortly following smaller maxima. For doing so a nonlinear smoothing of the signal envelope was developed. It acts in two modes. In the first mode the smooth envelope $x_s(k)$ rises with the signal envelope $x(k)$. When the signal envelope changes from the rising phase to a falling phase, hence after a maximum, the smoothing changes its mode and now performs a smoothing of the envelope signal with a first order Infinite Impulse Response (IIR) filter. When the smooth signal falls below the envelope signal the smoothing changes again in its rising phase. As a consequence

$$x_s(k) = \begin{cases} 0 & k = 0 \\ x(k) & x_s(k-1) \leq x(k) \wedge k > 0 \\ x(k) \cdot \vartheta & x_s(k-1) > x(k) \wedge x_s(k-1) \leq x(k-1) \wedge k > 0 \\ (1 - 1/\tau) \cdot x_s(k-1) + 1/\tau \cdot x(k) & x_s(k-1) > x(k) \wedge x_s(k-1) > x(k-1) \wedge k > 0 \end{cases} \quad (1)$$

the onsets are conserved and the signal is only smoothed after the onsets. A measurement point for the localization cues is generated one sample before the change from the rising to the falling phase and hence at the maxima of the signal. For the calculation of the envelope we use a rectification and low-pass filtering. In order to also inhibit only slightly stronger maxima following shortly after a maximum we introduced an additional inhibition factor ϑ in the smoothing process. At the measurement point the smooth signal is multiplied with this inhibition factor and therefore raised to a higher value from which it then falls again in the following smoothing phase (compare Eq. 1 where $x(k)$ is the original envelope signal, $x_s(k)$ the resulting smooth envelope, and τ the time constant of the IIR filter)). The result of this smoothing is shown in Fig. 1. Our smoothing process generates maxima at $0.03\,\mathrm{s}$ and

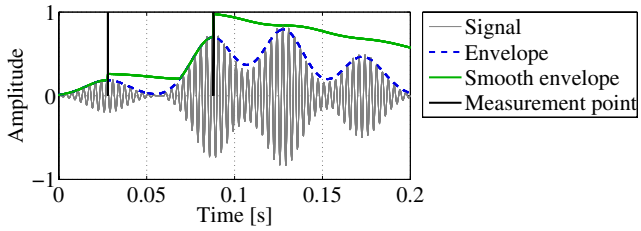

Figure 1: The procedure of the nonlinear smoothing is visualized.

$0.09\,\mathrm{s}$ and suppresses those at $0.13\,\mathrm{s}$ and $0.17\,\mathrm{s}$ which are usually effected by echoes. A single sound event produces a maximum in the left and right channel. Therefore if maxima in the two channels are closer together than $40\,\mathrm{ms}$ only the earlier maximum is kept.

3. Basic System Architecture

Instead of the real Asimo head we used a dummy head for the results presented here. Microphones were attached to the ears of the head. In line with our biology inspired approach we first apply a Gammatone filter bank [10] to the input signals. Stationary noise was estimated in the beginning of the signals and then subtracted from the remaining parts.

3.1. Cue Extraction

We use zero crossings to extract the ITD instead of the autocorrelation. Zero crossings are robust when applied to bandpass signals, significantly faster to calculate than an autocorrelation and biologically more plausible [11, 12, 13]. The ITD is measured at each zero crossing and then kept at this value until the next zero crossing occurs. For the IID values we calculated

$$IID(c,k) = \frac{x_L(c,k) - x_R(c,k)}{\max\left(x_L(c,k), x_R(c,k), x_{Min}\right)}\,, \quad (2)$$

where $x_L(c,k)$ and $x_R(c,k)$ are the envelope signals of the left and right channel after noise reduction at sample k and frequency channel c and x_{Min} the minimal expected signal level which prevents divisions by zero. The cues are evaluated at the time defined by the non-linear envelope smoothing. Based on the found maxima a $10\,\mathrm{ms}$ long measurement window is formed. In the current implementation the measurement window starts $13\,\mathrm{ms}$ before the maximum and ends $3\,\mathrm{ms}$ before the maximum. The final cue value for this channel and instance in time is the mean of the cue value in the window.

3.2. Mapping Matrix Calculation

As the geometry of the artificial head used is rather complex there is no straight forward mapping between the cue values and the corresponding angles possible. We therefore learn this mapping in an offline procedure. Sounds from known directions are presented to the head and localization cue values are extracted. An average cue value for a given location and a given frequency channel can be calculated from the calibration data. With this average value a mapping between the cue value and the angle can be established. We used 25 azimuth positions ranging from $-90°$ to $90°$ with $10°$ increment and a reduced increment around $0°$ in order to increase resolution around $0°$. The localization is limited to $-90°$ to $90°$ azimuth as we currently do not use combinations of cues or spectral characteristics of the signals to perform a front/back decision or elevation estimation.

3.3. Frequency Dependent Cue Confidence

From the data used in the mapping matrix calculation the variances $\sigma_{\mathrm{ITD}}(c,\varphi)^2$ for the three cues at a given channel c and angle φ can be calculated. Based on these variances and the average cue values $\mathrm{ITD}_M(c,\varphi)$ a confidence value for each cue at each channel averaged over all directions M is calculated[1]:

$$\eta_{\mathrm{ITD}}(c) = \sqrt{\frac{1}{M}\sum_{\varphi=-90°}^{90°} \frac{\mathrm{ITD}_M(c,\varphi)^2}{\sigma_{\mathrm{ITD}}(c,\varphi)^2}} \quad (3)$$

To avoid extremely high values due to variances close to zero a limit to the confidence was set. In a final step the confidence was normalized to the maximal confidence for all cues and all frequencies in order to have values in the range 0 to1.

3.4. Integration of the Cues

In a final step the different localization cues are integrated to form a localization estimate. For the integration an approach inspired by neural receptive fields was used [1]. The activation of the Gaussian node centered at angle φ and channel c at time instant k

$$A_{\mathrm{ITD}}(c,\varphi,k) = w_{\mathrm{ITD}}(c,\varphi,k) \cdot \exp\left(-\frac{(\mathrm{ITD}(c,k) - \mathrm{ITD}_{\mathrm{M}}(c,\varphi))^2}{2\sigma(c)^2}\right), \quad (4)$$

represents how close the current measure $\mathrm{ITD}(c,k)$ is to the cue value in the mapping matrix $\mathrm{ITD}_{\mathrm{M}}(c,\varphi)$ for the same channel and at angle φ. The parameter σ determines the width of the Gaussian kernel. The confidence weight $w_{\mathrm{ITD}}(c,\varphi,k) = \tilde{\eta}_{\mathrm{ITD}}(c) \cdot x(c,k)$ combines the previously calculated cue confidence $\tilde{\eta}_{\mathrm{ITD}}(c)$ and the energy of the underlying channel $x(c,k)$ after noise reduction ($x(c,k)$ is either $x_L(c,k)$ or $x_R(c,k)$ depending on which channel produced the maximum). The energy weighting enhances measures from signal parts with high energy as they normally are more reliable due to their better *Signal to Noise Ratio (SNR)*. Furthermore, a noise level dependent threshold $\delta_N(c)$ can be used for $x(c,k)$ so that only measurements where the energy of the underlying channel was above the noise level produce activations. For cases where the mapping is ambiguous, resp. non-injective, multiple activations for the same cue at different angles appear. This is

[1] For the sake of simplicity only the ITD cue is shown, but an identical procedure was used for the IID cue

a desired behavior as despite their ambiguity there is still information about the source location in these cue values. In an integration phase a histogram for the activations is build by summing over all channels K and cues:

$$H(\varphi,k)=\sum_{c=1}^{K} A_{\mathrm{ITD}}(c,\varphi,k)+A_{\mathrm{IID}}(c,\varphi,k)+A_{\mathrm{IED}}(c,\varphi,k) \quad (5)$$

In the histogram peaks form at the source location.

4. Results

The performance of the system is illustrated by means of some results recorded with the dummy head mentioned before in a conference room approximately of the size $7\,\mathrm{m}\times 15\,\mathrm{m}$ and height of $3\,\mathrm{m}$ (reverberation time $\mathrm{RT}_{60}=750\,\mathrm{ms}$). The room had walls with normal wallpaper, a window front partially covered by blinds and carpet on the floor. A running air conditioning and the PC fans caused a constant noise floor. Though the results are only shown for this room we performed also tests with the real-time system in a smaller office room (reverberation time $\mathrm{RT}_{60}=330\,\mathrm{ms}$). Also in this room the system performed good localization. The sampling rate was set to $48\,\mathrm{kHz}$. We used a Gammatone filter bank with 128 channels where center frequencies are increasing logarithmically from $50\,\mathrm{Hz}$ to $5\,\mathrm{kHz}$. Before the envelope smoothing we applied a logarithm to the envelope signal. For the adjustment of $\tau=180\,\mathrm{ms}$ we oriented ourselves at the estimated recovery time for humans from adaptation, the time constants in auditory models used as a front end for speech recognition, and the dynamics of the recorded signals [14]. The inhibition factor $\vartheta=1.07$ was determined empirically. The noise threshold $\delta_N(c)$ was set to the noise floor.

4.1. Comparison to onsets

In Fig. 2 the results of our system are compared to a similar system using onsets and a fixed suppression window for following onsets instead of the maxima and the signal dependent inhibition proposed here. In the upper plot of Fig. 2 the results of the onset based system for a speech signal presented via a loudspeaker at $1.3\,\mathrm{m}$ and $0°$ azimuth and an additional speech signal presented via loudspeaker at $90°$ and a distance of $3\,\mathrm{m}$ are shown. Both signals were presented at the same loudness. The lower plot contains the results for our maxima based system. The left graph shows the activations of the different angles over time. A smoothing along the time was performed with a Gaussian window of $100\,\mathrm{ms}$ width. On the right graph a histogram for all the activations summed up over time is given. As can be seen the activations are much better concentrated on the real location in the case of our system compared to the onsets. The second source has a significant impact on the localization of the first source in the case of the onset based system. In the case of our maxima based system the impairments in the localization of the source at $0°$ due to the additional source at $90°$ are much smaller and in the histogram summed over time the peak is quite sharp and precise. The second source does hardly appear in the graph and the summed histogram but this is largely due to the fact, that it is at $3\,\mathrm{m}$ compared to $1.3\,\mathrm{m}$ for the first source. A similar test was performed with only one source where also our maxima based system yielded much sharper peaks than the onset based system.

4.2. Localization in noisy conditions

In Fig. 3 localization results are given for a person talking at about $2\,\mathrm{m}$ distance and roughly $0°$ azimuth when additionally noise recorded from the fans of Asimo was presented via a loudspeaker

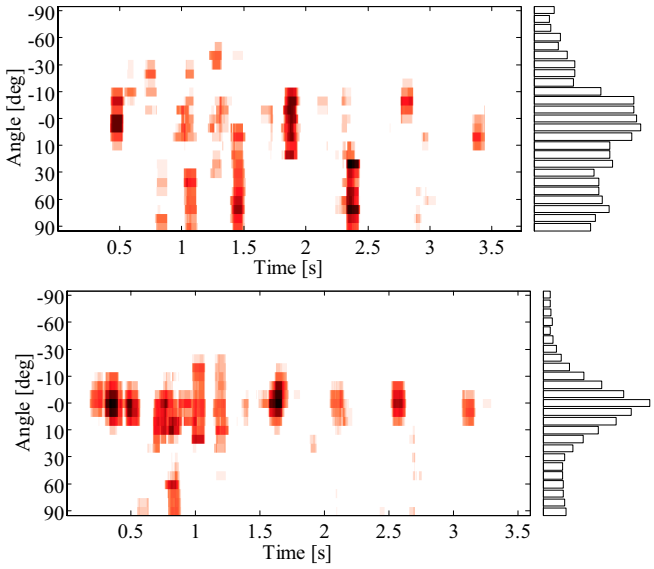

Figure 2: Comparison between the use of onsets and maxima for cue evaluation when two speech signals are presented via loudspeaker. The first is at a distance of $1.3\,\mathrm{m}$ and $0°$ azimuth and the second at $3\,\mathrm{m}$ and $90°$. The upper plot shows the localization results for the onset based system and the lower plot for our maxima based system.

directly from behind the head and piano music from approximately $-80°$(from the left). The values for the distances and angles are

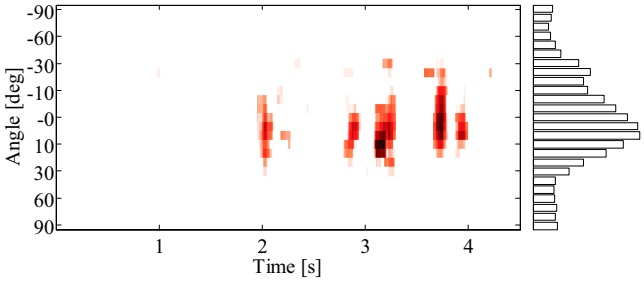

Figure 3: Localization result for our system of a person talking at a distance of about $2\,\mathrm{m}$ and $0°$ when additionally fan noise from the back and music at $-80°$ were present.

only approximative as this was done with a real speaker standing in front of the system. For this reason we are also not able to give a precise SNR for the signals. As the signals started one after the other (first fan noise, then music and finally the speaker) we can give some approximative values though. We calculated them via the mean over the respective segments. The SNR between the music and the fan noise was about $-3\,\mathrm{dB}$ in the left ear and $-5\,\mathrm{dB}$ in the right ear. Higher SNR values in the left ear are due to the fact that the music was on the left side. The SNR of the speech signal to the combined music and fan noise was approximately $1\,\mathrm{dB}$ in the left ear and $2\,\mathrm{dB}$ in the right ear, differences in the ears are due to uncertainty of the true position and measurement errors. In the plot in Fig. 3 the music starts at $0\,\mathrm{s}$ and the speech signal at $2\,\mathrm{s}$. The part with only the fan noise present was cut out for visualization. As can be seen from the plot the fan noise and music are almost completely suppressed in the histogram by the noise reduction. The peak in the histogram on the right side is much wider

than in the previous cases and the main peak is not at $0°$ but at $5°$. In this setup the absolute position of the speaker is not known but in the real-time system the localization has a precision such that the head is facing the speaker after it turned to the speaker [1]. For

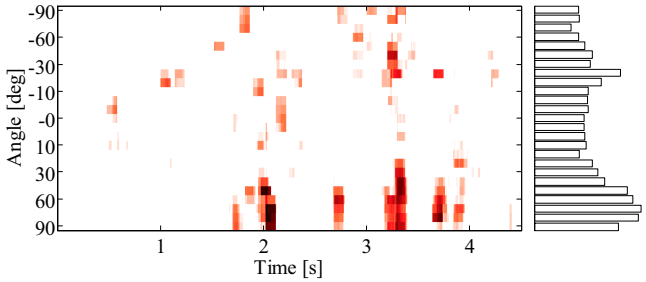

Figure 4: Localization result for our system of a person talking at a distance of about $2\,\mathrm{m}$ and $60°$ when additionally fan noise from the back and music at $-80°$ were present.

the setup in Fig. 4 the music and the fan noise were kept at the same location and level but the speaker was now at roughly $60°$. The SNR between the music and the fan noise was about $0\,\mathrm{dB}$, in the left ear and $-2\,\mathrm{dB}$ in the right ear. Changes in the values compared to the previous setup are due to imprecsions in the measurements. As SNR between speech and combined fan and music we estimated in this scenario $-3\,\mathrm{dB}$ in the left ear and $0\,\mathrm{dB}$ in the right ear. The SNR also varies due to the fact that the speaker could not utter at exactly the same loudness in each trial. The music starts at $0\,\mathrm{s}$ and the speech signal at $1.8\,\mathrm{s}$. As can be seen the main peak forms at $70°$ and some side peaks in the direction of the music are present. In general we see a trend for more precise localization at around $0°$ and decreasing performance at the outer regions. This is due to the fact that the cue sensitivity is highest at $0°$ and decreases to the side. Finally Fig. 5 shows a setup where

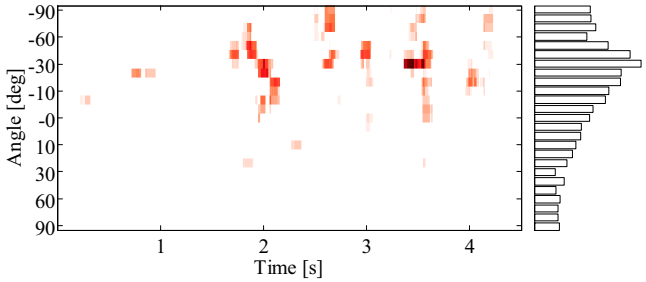

Figure 5: Localization result for our system of a person talking at a distance of about $2\,\mathrm{m}$ and $-30°$ when additionally fan noise from the back and music at $-80°$ were present.

the speaker was at roughly $-30°$, hence at the same side as the music. The remaining setup remained unchanged. The SNR between the music and the fan noise was about $-2\,\mathrm{dB}$, in the left ear and $-4\,\mathrm{dB}$ in the right ear. As SNR between speech and combined fan and music we estimated $-1\,\mathrm{dB}$ in both ears. The music starts at $0\,\mathrm{s}$ and the speech signal at $1.9\,\mathrm{s}$. As can be seen the main peak forms at $-30°$ and some side peaks in the direction of the music are present. The music interferes more with the localization in this case as it is on the same side but the speaker is still correctly localized.

5. Discussion

We developed a system which is able to perform sound source localization with 2 microphones in strongly echoic and noisy conditions. Our system was inspired by the human auditory system which is reflected in the binaural setting with a dummy head, the auditory preprocessing by the Gammatone filter bank, the use of zero crossings, the neural integration of the cues, and the modeling of the precedence effect. Especially for the precedence effect we largely modified and extended previous approaches which relied on onsets. Our system uses the maxima of the envelope signal and performs a signal dependent, not fixed as in previous systems, inhibition of shortly following maxima. The faster the signal falls the shorter the inhibition time. A following maximum with sufficent heigth overwrittes the inhibition in any case. These properties are in line with the findings from psychoacoustics. We compared the results of our system to an onset based system. There we could show that the localization results of our system are more reliable and precise. Furthermore we evaluated the performance of the system in a three source scenario with very bad SNR for the target signal. Here performance degrades in comparison to the single and two source scenario but results are still good enough for the use on a robot. The use of the zero crossings enabled the implementation of the system in real-time[1].

6. Acknowledgments

We thank Mark Dunn for his support with technical problems and Julian Eggert for fruitful discussions.

7. References

[1] Tobias Rodemann, Martin Heckmann, Björn Schölling, Frank Joublin, and Christian Goerick, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proc. Int. Conf. on Intelligent Robots & Systems (IROS)*, 2006, p. submitted.

[2] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 34, no. 3, pp. 1526–1540, 2004.

[3] D. Bechler and K. Kroschel, "Reliability criteria evaluation for TDOA estimates in a variety of real environments," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Proc. (ICASSP)*, Philadelphia, PA, 2005.

[4] E.-E. Jan and J. L. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," in *Proc. Int. Conf. Spoken Language Proc. (ICSLP)*, Philadelphia, PA, 1996, vol. 3, pp. 1321–1324.

[5] K. Nakadai, H. Nakajima, K. Yamada, Y. Hasegawa, T. Nakamura, and H. Tsujino, "Sound source tracking with directivity pattern estimation using a 64 ch microphone array," in *Proc. Int. Conf. on Intelligent Robots & Systems (IROS)*, Edmonton, Canada, 2005, pp. 196–202.

[6] H. Okuno and K. Nakadai, "Active audition for humanoid robots that can listen to three simultaneous talkers," *Journ. of the Acoust. Soc. of America (JASA)*, vol. 113, no. 4, pp. 2230, 2003.

[7] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Körner, "A probabilistic model for binaural sound localization," *to appear in IEEE Trans. on Systems, Man and Cybernetics - Part B*, 2006.

[8] T. Zahn, *Neural architecture for echo supression during sound source localization based on spiking neural cell models*, Phd. thesis, TU Ilmenau, Ilmenau, Germany, 2003.

[9] B. C. J. Moore, *An introduction to the psychology of hearing*, Academic Press, London, 5th edition, 2003.

[10] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Tech. Rep., Apple Computer Co., 1993, Technical report #35.

[11] Y.-I. Kim, S. J. An, R. M. Kil, and H.-M. Park, "Sound segregation based on binaural zero-crossings," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 2325–2328.

[12] D. Kim, S. Lee, and R. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech and Audio Proc.*, vol. 7, no. 1, pp. 55–69, 1999.

[13] C. Kaernbach and L. Demany, "Psychophysical evidence against the autocorrelation theory of auditory temporal processing," *Journ. of the Acoust. Soc. of America (JASA)*, vol. 104, pp. 2298–2306, 1998.

[14] M. Holmberg, D. Gelbart, and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing," *IEEE Trans. Speech and Audio Proc.*, vol. 14, no. 1, pp. 43–49, 2006.