EXTRACTION OF SOUND LOCALIZATION CUE UTILIZING PITCH CUE FOR MODELLING AUDITORY SYSTEM

Takatoshi Okuno, Thomas M. McGinnity, Liam P. Maguire
Intelligent Systems Research Centre, University of Ulster, Derry, BT48 7JL UK.
t.okuno@ulster.ac.uk

ABSTRACT

This paper presents a simple model for the extraction of a sound localization cue utilizing pitch cues in the auditory system. In particular, the extraction of the interaural time difference (ITD) as the azimuth localization cue, rather than the interaural intensity difference (IID), is constructed using a conventional signal processing scheme. The new configuration in this model is motivated by psychoacoustical and physiological findings, suggesting that the ITD can be controlled by the pitch cue in the simultaneous grouping of auditory cues. The localization cues are extracted at the superior olivary complex (SOC) while the pitch cue may be extracted at a higher stage of the auditory pathway. To explore this idea in the extraction of ITD, a system is introduced to feed back information on the pitch cue to control and/or modify the ITD for each frequency channel.

1. INTRODUCTION

Computational modelling of the auditory system has been recently investigated at the neuronal level. In particular, in a model for the extraction of auditory cues related to sound localization, such as ITD and IID, spiking neural networks (SNN) are utilized [1]. Such modelling can be established by defining which part of the auditory pathway functions to process each auditory cue. However, it is unclear where other cues such as pitch are processed in the auditory pathway. Wrigley et al. [2] states that the neurophysiological mechanisms underlying auditory stream formation are poorly understood and it is not fully known how groups of features are coded and communicate within the auditory system.

In physiological studies, it is understood that the auditory cues for sound localization are extracted at the medial superior olive (MSO) and lateral superior olive (LSO) in the SOC, then integrated at the inferior colliculus (IC) to extract representations of positions in space [3]. However, the extraction process of *pitch*, which is recognized as one of the most primitive cues among the auditory cues, is not well identified. According to some recent papers, the extraction of pitch may be processed at the IC by the existence of some neurons responding to the sinusoidally amplitude modulated sound within a restricted range [3]; or it may be processed at the SOC by the existence of Huggins pitch known as a result of the binaural interaction of noise stimuli [4], [5]; or that there may exist an extraction process of pitch at the brainstem and thalamus. The decision process of pitch may occur at lateral Heschl's gyrus in the auditory cortex through the analysis of fMRI [6]. Therefore, it is not currently possible to identify exactly which part of the auditory pathway has a particular role for extracting pitch. Assuming that the decision process of pitch ends at the auditory cortex, it may be possible to have the decision process of pitch performed after the extraction process of the sound localization cues.

In psychoacoustical studies, there have been extensive findings about sound localization, pitch and other cues, that have been summarized in the research framework referred to as auditory scene analysis (ASA) [7]. Treating ASA with a computational approach (computational ASA: CASA) to resolve certain engineering problems such as signal separation issues has enabled many computational models for ASA systems to be undertaken [8].

Recently, sound localization cues were used for sequential organization. This means that auditory objects from the same spatial direction can be organized as one auditory stream, even if those auditory objects are isolated from each other in terms of time, although the sound localization cues have been regarded as one of the primitive cues for simultaneous organization in ASA [9, 10]. Darwin [11] states that ITDs are remarkably ineffective at segregating simultaneous sounds despite the dominance of ITDs in the region around 500 Hz [12]. Culling also mentioned that harmonicity contributes to the grouping of sounds across the frequency integration of ITD, according to experimental results by Hill et al. [13]. Furthermore, the relationship between ITD and pitch indicated that the formation of auditory objects precedes decisions on their location so that a model would allow pooling of location information across frequency channels in order to reduce the variability found in individual channels and so produce a percept with a stable location.

This paper proposes a simple model using the ITD and pitch cues, that considers the interaction between the two cues while they are being extracted. This is not a conventional approach in CASA, that permits the individual extraction of auditory cues independently [8]. Considering the contradiction between the physiological view (the extraction of sound localization cues may precede the decision process of pitch) and psychoacoustical view (the for-

mation of auditory objects including the use of pitch may precede the decision of sound localization), the model is proposed as a feedback system with the extraction of ITD before that of pitch so that the model can be biologically-plausible. The proposed model differs from the frame based method [10] in that it concerns the order of the process as a feedback system and the frequency dependence of ITDs. The model is constructed at the level of conventional signal processing, incorporating the use of an auditory periphery model and a correlation based calculation as this will allow the model to be reconstructed by an SNN in the future in terms of biological plausibility.

2. A PROPOSED MODEL

The proposed model is described by the systematic configuration shown in Fig.1 and each calculation method is explained in turn as follows. The signals presented here are speech signals and white noise for simplicity and quantification. The ratio between the levels of the two signals is controlled by the signal to noise ratio (SNR) at the origin of the signals. By convolving each signal with head-related transfer functions (HRTFs), it can convey the information of sound location. The signals for the left and right ears are added to yield the binaural signal. HRTFs utilized here are produced from the MIT media lab, they are bilaterally symmetric measurement data sets using a dummy head with the same size pinnae for both ears in an anechoic chamber [14, 15]. Since all HRTFs are prepared at 44.1 kHz sampling frequency, computer simulation performed later on is undertaken using the same sampling frequency. For the directions of sound, a range of $\pm 90^\circ$ azimuth with the midline as the centre (5° intervals) is considered.

Each binaural signal is decomposed into frequency channels by applying a Gammatone filter bank which models the filtering at the cochlea [16]. The frequency decomposition by the cochlear filtering is performed in 64 channels covering 50-8000 Hz, which should be a sufficient number of channels in terms of the equivalent rectangular bandwidth (ERB) rate [17]. The frequency range covered by the filter bank should be appropriate even though ITD and pitch frequencies processed at a later stage are taken into account. The output of the filter bank is used as the input to calculate the summary cross correlation function (SCCF) in order to obtain the ITD. From this stage, the calculation is processed on a frame by frame basis. If SCCF is calculated in the range of the lag time between ± 1 ms, it covers the range of the azimuth between $\pm 90^\circ$. However, one frame length is set as 30 ms here because of the stability of the pitch extraction algorithm, which is performed at a later stage.

Different ways to obtain the signal to be used for the pitch extraction from the binaural signal are proposed in the literature. In [9], the signal used for the pitch extraction is named as the *better ear signal* which has a better SNR determined from the signal before adding the signals convolved with the HRTFs. In [10], a signal produced by averaging the left and right ear signals are used as the better ear signal. Considering the head-shadow effect [18] and the diffraction wave, however, averaging the left and right

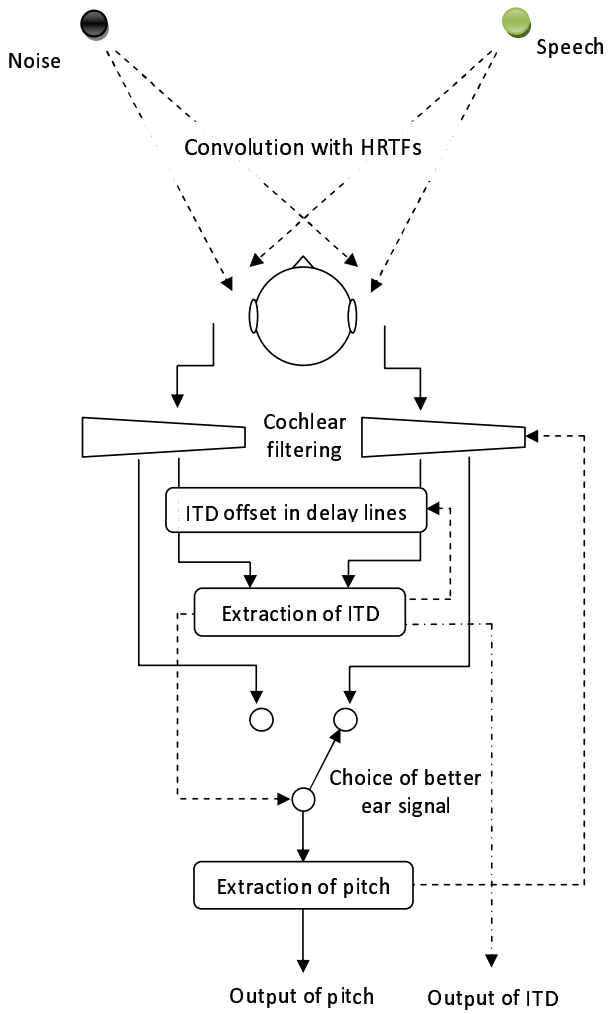

Figure 1. A systematic configuration for a proposed model

ear signals would make the better ear signal complicated. Therefore, in each frame, the better ear signal is obtained by choosing the left or right ear signal, which is based on the value of ITD such as the left for the negative value of ITD or the right for the positive value of ITD, calculated at the earlier stage. This function is depicted by a symbolic switch in Fig.1 and denoted as "Choice of better ear signal". This assumes that the ITD extracted is calculated for a dominant signal in the binaural signal or for the second dominant signal after the dominant signal is removed.

The pitch extraction algorithm is implemented by the minor updated algorithm of enhanced summary auto correlation function (ESACF), based on the summary auto correlation function (SACF) [19]. Practically, the better ear signal obtained in the frequency channels is half-rectified and used to calculate the auto correlation functions, which are then summarized over the frequency channels. Considering pitch extraction under noisy conditions, the minor update is conducted by subtracting the mean value from the SACF, then re-performing half-rectification in order to make the sensitivity of the pitch extraction higher. In ESACF, the minor updated SACF is interpolated on the basis of harmonics, then the signal is subtracted from the

minor updated SACF and half-rectified again. This process is repeated to estimate the fundamental frequency (F0) within a frame. This interpolation is especially performed by considering the second, third and fifth harmonics. The summation over frequencies in SACF is calculated for 25 channels from 99 Hz to 961 Hz, which corresponds to the centre frequencies of the Gammatone filter bank. However, the limitations for the possible frequency range of the F0s are defined from 100 Hz up to 450 Hz because the pitch extraction is basically designed for a speech signal. In addition, to avoid the mis-extraction during silent intervals of speech (when clean speech is the only input), the pitch extraction algorithm can be turned on or off with the estimation of the power of the signal such as comparing the lag-zero values of SACF with a certain threshold.

The accuracy of the estimated F0s is dependent on the capability of the algorithm and the characteristics of the input signals. This means that it is not always possible to extract the true F0s. Therefore, the *harmonic stream* which is composed of a F0 up to third harmonics, is defined by classifying the estimated F0s. In the n-th frame, a candidate of F0 extracted by the ESACF algorithm is defined as $F0(n)$, and three frequencies of the harmonic stream are defined as $f0(n)$, $f1(n)$ and $f2(n)$ respectively. Then the classification is performed by the following three equations:

$$\text{if } F0(n) < 200, \begin{cases} f0(n) = 2 \times F0(n) \\ f1(n) = 3 \times F0(n) \\ f2(n) = 0 \end{cases} \quad (1)$$

$$\text{if } 200 \leq F0(n) \leq 350, \begin{cases} f0(n) = F0(n) \\ f1(n) = 2 \times F0(n) \\ f2(n) = 3 \times F0(n) \end{cases} \quad (2)$$

$$\text{if } F0(n) > 350, \begin{cases} f0(n) = 2 \times 0.5 \times F0(n) \\ f1(n) = 3 \times 0.5 \times F0(n) \\ f2(n) = 0 \end{cases} \quad (3)$$

Owing to the limitation of Eq.(1), the $F0(n)$ from 100 to 200 Hz, which is generally said to be the F0 of speech, is not used for any of the following processes. This is because the ITDs for the frequencies lower than 200 Hz are not reliable due to the phase analysis of HRTFs between left and right ears. In addition, since the frequencies higher than 300 Hz are dominant for the pitch sensation [20], it would be reasonable to construct the harmonic stream higher than 200 Hz. $f2(n)$ in Eq.(1) and (3) is set to zero since the 4th harmonics of the expected F0 is not used to construct the harmonic stream here.

The harmonic stream is then replaced to the nearest centre frequencies of the Gammatone filter bank, and the chosen frequency channels are utilized to calculate the ITD at the next frame. However, due to the effect of the diffraction wave around the dummy head in the measurement of HRTFs, it is known that ITD and IID are changed depending on the frequencies, the direction of sound source and the physical size of the dummy head [21]. This is because there is an object between two microphones (like a dummy head) which is not negligible when the human head is considered. Especially, in the sampling frequency that just covers the audible range such as 44.1 kHz, a small number of sample difference would affect the accuracy in estimation of the sound direction. In [18], the fact that their proposed system does not work due to the diffraction wave in the case where there is an object between two microphones in their 2ch microphone array system, is discussed. They indicate that it can be interpreted as filtering with the transfer function characteristic of the diffraction. This filter can be also incorporated into the delay units of the dual delay line. Here, before the summation over frequencies in SCCF, the offset of ITD is applied to correct the estimated angles or ITDs. It is proposed that a matrix of the offset as a function of the sound direction and frequency is prepared, and the offset selected by the estimated angle in the previous frame is applied. With the sampling frequency f_s, the angular frequency ω, the phase difference of HRTFs between left and right ear $\varphi_\theta(\omega)$ at the measured angle θ of HRTFs, the offset is calculated as the difference between the phase delay (in samples) in Eq.(4) and the sample difference for the whole frequency range obtained by the cross correlation of HRTFs:

$$Phase\ delay\ (\theta) := \frac{-f_s\ \varphi_\theta(\omega)}{\omega} \quad (\text{in samples}) \quad (4)$$

Fig.2 shows the offsets as a function of the centre frequencies of Gammatone filter bank, for $\theta = 30°, 60°, 90°$. These values are rounded to the nearest integer for the ITD offset axis. It is noted that the offset is applied even if the estimated angle is wrong or changed significantly from the previous frame since this offset is utilized based on the results of the estimation in the previous frame. Therefore, for example, if the sound direction changes from $-60°$ to $+60°$ drastically, the offset gives the error to the estimates of $+60°$ since the offset is applied based on the the estimates for $-60°$.

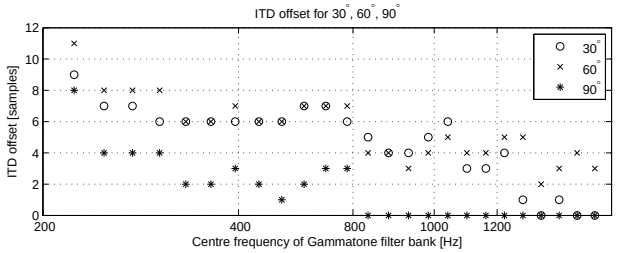

Figure 2. Examples of the offset in case of $\theta = 30°, 60°, 90°$

Using the geometric approximation often used for a 2ch microphone array system, the estimated ITDs in the samples can be converted to azimuthal angles. However, this conversion cannot keep the linearity for angles close to $\pm 90°$ because of the diffraction of the dummy head. Therefore, by calculating the cross correlation function of HRTFs between both ears for all azimuth, and then interpolating the obtained ITDs linearly, the angles when HRTFs are measured are linked with estimated ITDs in samples. Since the azimuth between $-90°$ and $0°$ is symmetric to the azimuth between $0°$ and $90°$, the relationship between the estimated ITDs and the angles of HRTFs is shown in Fig.3.

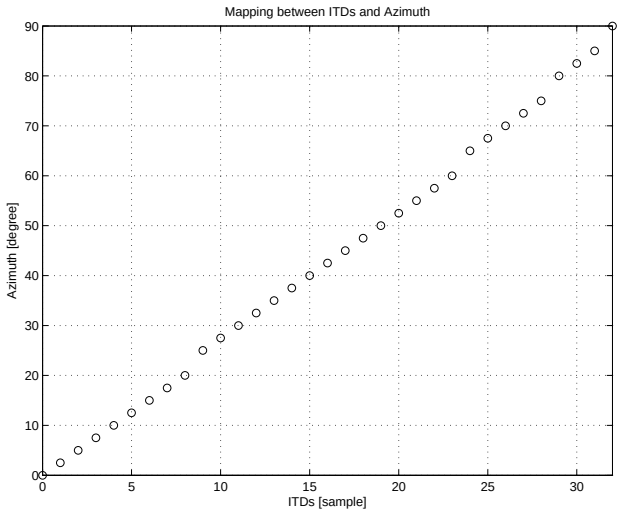

Figure 3. The relationship between the estimated ITDs and the azimuth

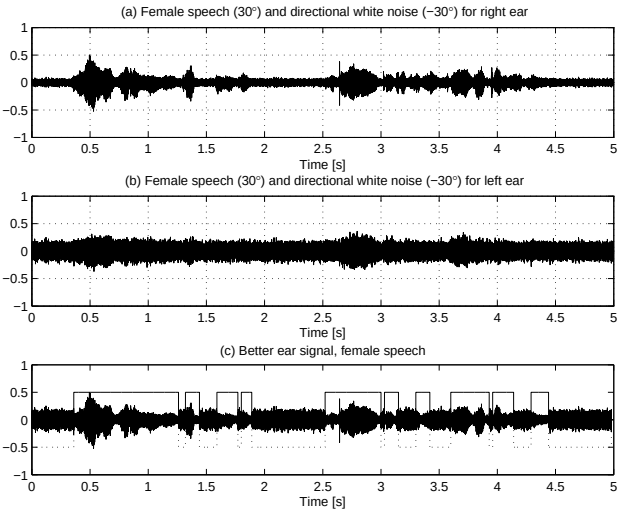

Figure 4. (a) Right ear signal, (b) left ear signal, (c) better ear signal and the results of "Choice of better ear signal" (SNR:10 dB at the location of sound sources)

3. SIMULATION RESULTS

To examine the performance of the proposed system, a computer simulation was conducted for a speech signal and a directional white noise. Five seconds of female speech sample was utilized and convolved with the HRTF for the 30° angle. Similarly, white noise was generated and convolved with the HRTF for -30°, which becomes a directional white noise. Both signals were added in order to generate the binaural signal. The right ear signal, left ear signal and better ear signal are shown in Fig.4(a), (b) and (c) respectively. At the location of sound sources, the SNR is controlled as 10 dB. Comparing (a) with (b), it appears there is a difference in SNR because of the head-shadow effect. The solid and dotted lines in Fig.4(c) are the results of "Choice of better ear signal"from both (a) and (b). Namely, when there is a solid line with a value $+0.5$, which means the intervals of the positive value of ITD, which indicates that the right ear signal is used. Conversely, when there is a dotted line with a value -0.5, this means the intervals of the negative value of ITD, which indicates that the left ear signal is used. Therefore, (c) is a combination of the signals from both (a) and (b).

As mentioned before, a modified ESACF method is employed for pitch extraction. The extraction results are shown in Fig.5 with circles on the spectrogram which uses a log frequency axis (100-2000 Hz) for the vertical axis. Fig.5(a) shows the extracted $F0(n)$ on the spectrogram. Although the detection of the silent intervals for speech is performed at the same time, it does not work since there is a directional white noise in the intervals. However, it can be confirmed that the pitch extraction algorithm does not pick up the wrong $F0(n)$ for the most of the intervals. Even though the pitch extraction algorithm is not evaluated here, it could be concluded that the lowest components of speech in the spectrogram, namely the F0s are followed by the algorithm. Applying the extracted $F0(n)$ for Eq.(1)-(3), the harmonic stream is constructed as shown in Fig.5(b). $f0(n)$, $f1(n)$ and $f2(n)$ are indicated by the symbols shown in Fig.5(b). It can be seen that there are some $F0(n)$ discarded under 200 Hz intentionally, based on Eq.(1)-(3).

The frequency channels are selected by feeding back

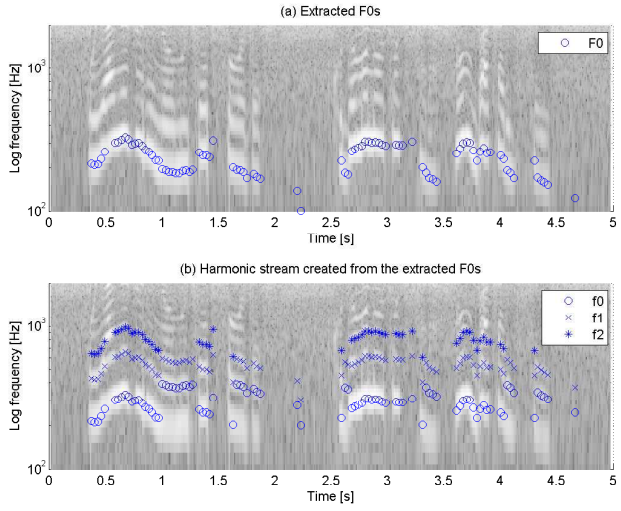

Figure 5. The results of pitch extraction and the harmonic stream based on Eq.(1)-(3)

the harmonic stream in Fig.5(b) at the next frame in order to calculate the ITD. Similarly, the offset is applied at the next frame according to the estimated angle. In Fig.6, the estimated angles are shown in the case of four different conditions, "none", "pitch", "offset" and "both"; "none" has no feedback of pitch and no offset, "pitch" has feedback of pitch but no offset, "offset" has no feedback of pitch but has the offset, and "both" has the feedback of

both pitch and the offset. There are no differences for the four conditions between the speech intervals and the silent intervals in terms of time. The most important point of the evaluation is the accuracy of the estimated angles for *speech*, such as the estimated angles for $30°$. "none" and "pitch", to which the offset is not applied, are far from $30°$ and change around $40°$-$50°$. However, "offset" and "both", to which the offset is applied, the $30°$ angle is achieved, as expected. Therefore, applying the offset is very important and inevitable to building this system. It is noted that the estimated angles change to $30°$ after the estimation overshot when the estimated angles changed from $-30°$ to $+30°$ rapidly since the offset is applied at one frame later. As for the estimated angles for the directional white noise, the estimated angles stay around $-30°$. This seems to be because the power of the white noise is constant over the frequency range, and ITDs in the higher frequency channels dominate the estimation of ITD by SCCF.

To evaluate the results more quantitatively, a correct answer for the estimated angles, $\theta(n)$, is prepared for $30°$ when the SNR is 20 dB, since the whole intervals of the better ear signal become $30°$ if a clean speech is used. To prepare the correct answer, the intervals when the right ear signal is chosen like the solid lines in Fig.4(c) are calculated, and then it is assumed as if the obtained intervals were all $30°$. Here, accuracy (Acc) in [10] is quoted as the evaluation function. If the estimated angle is defined as $\hat{\theta}(n)$, Acc is defined as

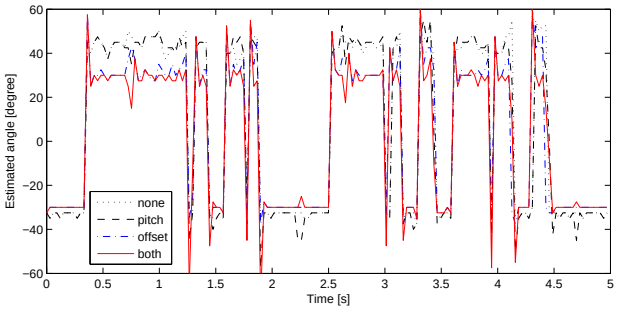

Figure 6. The temporal change of the estimated angles under four conditions)

$$Acc = \frac{1}{N}\sum_{n=1}^{N}\delta(\theta(n),\hat{\theta}(n)) \tag{5}$$

where N is the number of frames. $\delta(a,b)$ is defined as

$$\delta(a,b) = \begin{cases} 1, & \text{if } |a-b| < \beta \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where β is the tolerance and is fixed as $3°$ as [10] did. In Fig.7, Acc is shown under four conditions when SNR is set up as 10 and 0 dB. Acc of "none" and "pitch", in which the offset is not applied is very low as expected from Fig.6, and that means that the system is basically not working. In the case that the SNR is 10 dB, it is remarkable that Acc for "both" has 10% better accuracy than that for "offset" despite using only a few frequency channels. In the case

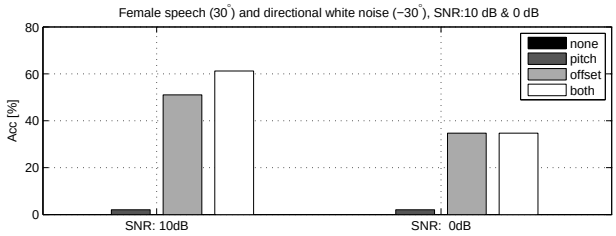

Figure 7. Acc under four conditions when SNR is set up as 10 and 0 dB

of SNR of 0 dB, Acc is wholly decreased. It is known that white noise as a masker affects the accuracy in the estimation of sound direction [22]. The reason why Acc of "offset" and "both" became the same accuracy might be the reduction of the accuracy in SCCF itself rather than the the reduction of the accuracy in the pitch extraction. It would be necessary to have a further investigation between the accuracy in the pitch extraction and SNR to determine the reason.

4. DISCUSSION AND FURTHER WORK

When the estimated angle moved from the angle of the directional white noise to that of the speech, overshoots were observed at "offset" and "both" in Fig.6. This might be understandable intuitively, in that it might take a certain time for the auditory system to estimate the accurate sound location if the location moved rapidly. In the proposed system, the frame size to calculate ITD was fixed at 30 ms, which is the same length as the frame size to calculate F0s. However, ± 1 ms is good enough to obtain the angles for the azimuth. Therefore, if the frame size is fixed at 1 ms for ITD and 30 ms for F0s, it might be possible to reduce the overshoots assuming a smoothing function. In addition, the frame size could be crucial to the temporal boundary between simultaneous and sequential organization. It is necessary to consider the compatibility with the findings of psychoacoustics.

In this paper, although the number of streams for the pitch extraction is defined as 1 such as speech or white noise, the multi-pitch algorithm is utilized in [9]. It is important for the model to investigate how many streams should be extracted at the same time in the simultaneous and sequential organization.

Although the importance of biological plausibility has been discussed since then, the pitch extraction algorithm is still too complicated to be implemented at neuronal level. Also, the method to create the better ear signal is indefinite in terms of the physiological view. Compatibility between the physiological and psychological view with the engineering tool requires more investigation.

5. CONCLUSIONS

In this paper, a simple model integrating ITD as a sound localization cue with the pitch cue is proposed. Considering the order that the decision process of pitch could be per-

formed after the extraction process of sound localization cues, a feedback system is employed. Although only one speech sample was utilized in the simulation, the integration shows 10% improvement of accuracy in the extraction of ITD. Integrating other auditory cues including IID for elevation is planned future work.

Acknowledgments

This research is supported under the Centre of Excellence in Intelligent Systems (CoEIS) project, funded by the Northern Ireland Integrated Development Fund and InvestNI.

6. REFERENCES

[1] B. Glackin, J. A. Wall, T. M. McGinnity, L. P. Maguire, and L. J. McDaid, "A spiking neural network model of the medial superior olive using spike timing dependent plasticity for sound localization," *Front. Comput. Neurosci*, vol. 4, no. 18, pp. 1–16, 2010.

[2] S. N. Wrigley and G. J. Brown, "A computational model of auditory selective attention," *IEEE Trans. Neural Network*, vol. 15, no. 5, pp. 1151–1163, 2004.

[3] J. K. Bizley and K. M. M. Walker, "Sensitivity and selectivity of neurons in auditory cortex to the pitch, timbre, and location of sounds," *The Neuroscientist*, vol. 16, no. 4, pp. 453–469, 2010.

[4] E. M. Cramer and W. H. Huggins, "Creation of pitch through binaural interaction," *J. Acoust. Soc. Am*, vol. 30, no. 5, pp. 413–417, 1958.

[5] K. M. M. Walker, J. K. Bizley, A. J. King, and J. W. H. Schnupp, "Cortical encoding of pitch: Recent results and open questions," *Hear Res.*, vol. 271, pp. 74–87, 2011.

[6] R. D. Patterson, S. Uppenkamp, I. S. Johnsrude, and T. D. Griffiths, "The processing of temporal pitch and melody information in auditory cortex," *Neuron*, vol. 36, pp. 767–776, 2002.

[7] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[8] E. D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis*. Wiley/IEEE Press., 2006.

[9] J. Woodruff and D. L. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1856–1866, 2010.

[10] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "Integrating pitch and localisation cues at a speech fragment level," *in Proc. of INTERSPEECH 2007*, pp. 2769–2772, 2007.

[11] C. J. Darwin, E. W. A. Yost, A. N. Popper, and R. R. Fay, *Spatial hearing and perceiving sources in* Auditory Perception of Sound Sources. Springer, 2007, ch. 8, pp. 215–232.

[12] J. F. Culling and Q. Summerfield, "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," *J. Acoust. Soc. Am*, vol. 98, pp. 785–797, 1995.

[13] N. I. Hill and C. J. Darwin, "Effects of onset asynchrony and of mistuning on the lateralization of a pure tone embedded in a harmonic complex," *J. Acoust. Soc. Am*, vol. 93, no. 4, pp. 2307–2308, 1993.

[14] W. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab Perceptual Computing, Tech. Rep. 280, 1994.

[15] W. G. Gardner, *3-D Audio Using Loudspeakers*. Boston: Kluwer Academic, 1998.

[16] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," *Cambridge, UK., Tech. Rep., MRC Applied Psychology Unit*, 1988.

[17] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear Res.*, vol. 47, pp. 103–138, 1990.

[18] C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Am*, vol. 108, no. 4, pp. 1888–1905, 2000.

[19] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process*, vol. 8, no. 6, pp. 708–716, 2000.

[20] R. J. Ritsma, "Frequencies dominant in the perception of the pitch of complex sounds," *J. Acoust. Soc. Am*, vol. 42, pp. 191–198, 1967.

[21] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *J. Acoust. Soc. Am*, vol. 62, pp. 157–167, 1977.

[22] C. Lorenzi, S. Gatehouse, and C. Lever, "Sound localization in noise in normal-hearing listeners," *J. Acoust. Soc. Am*, vol. 105, no. 3, pp. 1810–1820, 1999.