DEVELOPMENT AND ASSESSMENT OF A LOCALIZATION ALGORITHM IMPLEMENTED IN BINAURAL HEARING AIDS

Gilles Courtois[*], *Patrick Marmaroli*[*], *Hervé Lissek*[*], *Yves Oesch*[†], *William Balande*[†]

[*] Swiss Federal Institute of Technology (EPFL)
Signal Processing Laboratory (LTS2)
1015 Lausanne
Switzerland

[†] Phonak Communications AG
Laenggasse 17
3280 Murten
Switzerland

ABSTRACT

Digital wireless microphone systems have been developed to enhance speech intelligibility for hearing aid users. The voice of the speaker of interest is picked-up close to the mouth by a body-worn microphone, and rendered in a diotic way (same signal at the two ears) in both hearing aids. This means that binaural spatial cues are missing, whereas they are known to be useful for speech perception, speaker localization and sense of immersion. Some spatial information could be included in the microphone signal if the location of the talker relative to the listener would be known. This paper reports an algorithm that performs the real-time localization and tracking of the speaker's position, taking into account the technical constraints related to hearing aids. It has been assessed with speech signals in a conventional classroom, and has reached convincing performance.

Index Terms— Binaural localization, Tracking, Hearing aids

1. INTRODUCTION

Digital wireless microphone systems for hearing aids have been developed in order to improve speech intelligibility and listening comfort of hearing-impaired subjects. They are generally used in classrooms, conferences or meetings, when a talker is addressing an audience. A typical system consists of a small transmitter microphone, which picks up the voice of the speaker, and sends the speech signal wirelessly into some receivers plugged on the hearing aids of a listener. As the voice is captured very close to the mouth, and thanks to beamforming processing, almost only the direct sound is recorded, with a high signal-to-noise ratio (SNR). Each hearing device then manages two inputs: the demodulated audio signal (radio-transmitted) and the acoustical signal (recorded with the hearing aid microphones), as described in Figure 1. Current systems render the sound from the body-worn microphone in a diotic way, i.e. the same signal is played in the left and right hearing aids. This means that binaural spatial cues are absent. The introduction of binaural audio cues corresponding to the speaker location relative to the aided listener would be of great help for enhancing speech intelligibility and sense of immersion. To this end, a binaural localization algorithm (BLA) is required in order to infer the position of the speaker from the available signals on both hearing aids.

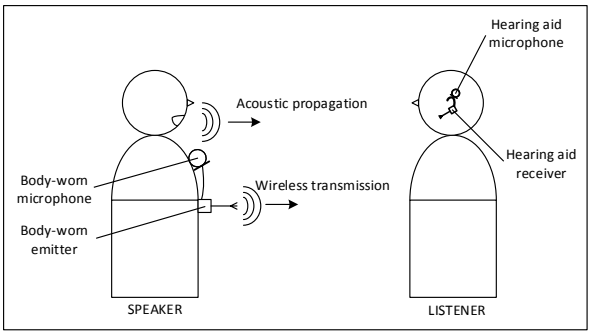

Fig. 1. Wireless microphone systems for hearing aids. In addition to the usual acoustic path captured by the hearing aid microphones, the voice of the speaker is picked up by a body-worn microphone and transmitted to the receivers plugged into the hearing aids of the listeners.

The authors have already reported a review of the state-of-the-art related to BLAs, as well as the constraints demanded by an implementation on hearing aids in [1]. In brief, the limited embedded memory available on conventional hearing instruments cannot store an extensive head-related transfer function (HRTF) database. This is actually needed by several BLAs introduced in the literature so far [2–5]. Another strategy is based on the estimation of the usual interaural cues (i.e. Interaural Time Difference (ITD) and Interaural Level Difference (ILD)), which are compared to a small set of reference values, or fed into a mathematical model [6–10]. For the targeted application, this approach has to restrict the amount of computation, due to the limited processing power of hearing aids, the need for battery saving, and the necessity to ensure real-time processing. Although available in most of current

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

binaural hearing instruments, the wireless communication between both devices has to be limited as much as possible, as it increases the current consumption. Therefore, the exchange of entire audio frames is not desired, i.e. cross-correlation-based procedures should be avoided. Finally, the acoustical environments yield additional constraints in terms of robustness against interfering noises and reverberation.

This paper reports an algorithm that estimates acoustical and radio-frequency (RF) spatial cues. The goal is then to localize a speaker, who is wearing the body-worn microphone of a digital wireless microphone system. The algorithm assumes that the talker is located in the frontal horizontal plane (i.e. half plane) relative to the aided listener, which is a valid hypothesis in the majority of classrooms or meeting rooms. This BLA fulfills all the technical specifications demanded by hearing aids, and performs speaker localization and tracking in real-time.

2. ALGORITHM

2.1. Interaural Phase Difference

In the low-frequency range, the signal captured by the right hearing aid microphone s_R can be modeled as a shifted copy of the signal in the left hearing aid s_L. Both acoustical signals can also be written as a delayed copy of the signal coming from the body-worn microphone s_{RX}. If the potential differences of reverberation and noise between both ears are neglected, the following relation holds:

$$\begin{cases} s_L(t) = s_{\text{RX}}(t - \Delta_t) \\ s_R(t) = s_{\text{RX}}(t - \Delta_t - \delta) \end{cases}, \tag{1}$$

where Δ_t denotes the delay between s_{RX} and the signals from the hearing aid microphone s_L and s_R, and δ is the ITD. The cross-spectra R_L and R_R of the left and right acoustical signals are derived according to:

$$\begin{cases} R_L(f) = |S_{\text{RX}}(f)|^2 e^{2\pi f \Delta_t} \\ R_R(f) = |S_{\text{RX}}(f)|^2 e^{2\pi f \Delta_t} e^{2\pi f \delta} \end{cases}, \tag{2}$$

where S_{RX} is the Fourier Transform of s_{RX}. The Interaural Phase Difference (IPD) is the frequency-domain representation of the ITD, that is:

$$\varphi(f) = 2\pi f \delta. \tag{3}$$

In practice, the IPD can be estimated at certain well-chosen frequencies f_i by exchanging the corresponding complex values of R_L and R_R between the two devices, i.e.:

$$\varphi(f_i) = \angle \left(\frac{R_R(f_i)}{R_L(f_i)} \right). \tag{4}$$

In the reported algorithm, three IPD observations $\tilde{\varphi}(f_1)$, $\tilde{\varphi}(f_2)$ and $\tilde{\varphi}(f_3)$ are derived for a set of three low frequencies f_1, f_2 and f_3 for each analysis frame. These three observed IPD values are compared with a collection of theoretical values for N azimuth angle θ_j in the frontal horizontal plane, and the error is computed as follows:

$$\begin{gathered} \epsilon(\theta_j) = \sum_{i=1}^{3} \sin^2(\varphi(f_i, \theta_j) - \tilde{\varphi}(f_i)), \\ \text{for } j = 1, 2, ..., N. \end{gathered} \tag{5}$$

The theoretical values of the IPD are calculated according to the sine law [11] that associates each location of the source in the frontal horizontal plane with a corresponding IPD at the ears of the listener:

$$\varphi(f_i, \theta_j) = \frac{2\pi f_i a}{c} \sin \theta_j, \tag{6}$$

where c is the speed of sound and a is the distance between the two microphones assimilated to each ear canal entrance.

The spatial resolution of the BLA has been set to five sectors (see Figure 2), which was judged sufficient for the application. The errors, as computed in (5), are gathered into sectors, i.e. the errors for all azimuths θ_j in a sector are summed. Then, the resulting global errors are converted into probabilities for the source to be in each sector. This means that a large difference between the observation and the model for a given sector leads to a low probability of being in the corresponding sector, and conversely.

2.2. Side estimation

The level difference between the left and right acoustical signals (ILD), as well as the strength difference of the RF signals reaching both hearing aids (RSSID, for Received Signal Strength Indication Difference) are used to determine the actual side where the speaker stands relative to the listener. Both ILD and RSSID provide a three-state output that can be either "left", "right", or "unknown", the last stands when it is not possible to estimate the side (due to small level differences between both devices or too noisy information). The side information coming from the estimation of the ILD and the RSSID is taken into account by applying some weightings on the IPD-based computed probabilities $p(\Theta_i)$:

$$\begin{gathered} S_W(\Theta_i) = W_{\text{ILD}}(\Theta_i) W_{\text{RSSID}}(\Theta_i) p(\Theta_i) \\ \text{for } i = 1, 2, ..., 5, \end{gathered} \tag{7}$$

where $S_W(\Theta_i)$ denotes the weighted score associated with the sector Θ_i (as defined in Figure 2), and W_{ILD} and W_{RSSID} are the weightings applied on the IPD-based probabilities. These weightings emphasize the probabilities of being in the two sectors of the current side and lessen the probabilities of being in the two sectors of the opposite side.

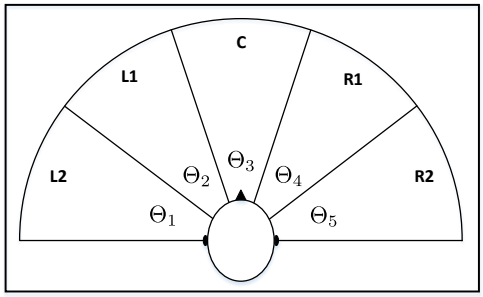

Fig. 2. Spatial resolution as defined for the speaker localization algorithm. The frontal horizontal plane is divided into five spatial sectors.

For example, if the ILD better matches a speaker located on the left (i.e. output is set to "left"), the weights will be:

$$W_{\mathrm{ILD}}(\Theta_i) = \begin{cases} \nu \text{ for i} = 1, 2 \text{ (left sectors)} \\ 1 \text{ for i} = 3 \text{ (central sector)} \\ \dfrac{1}{\nu} \text{ for i} = 4, 5 \text{ (right sectors)} \end{cases}, \qquad (8)$$

with $\nu > 1$. Note that the contributions of the ILD and RSSID are such that they vanish in case of contradictory side indication, which helps to reject erroneous cue estimations. If the output is set to "unknown", all weights are equal to 1.

2.3. Tracking

In order to enhance the system stability, a tracking procedure has been implemented. The tracking model developed in the presented BLA is a probabilistic network, where the five spatial sectors are represented by five nodes connected with arrows. Each arrow corresponds to a probability to go from a sector to another one. Every node is also connected to itself by an arrow representing the probability to stay in the current sector. These transition probabilities act as weights that emphasize or lessen the current score of each sector, depending on the previous estimated location of the speaker.

Let j the index of the estimated sector for the previous frame, $j \in \mathbb{N}_5^*$. The transition probabilities are applied as follows:

$$S_W(\Theta_i|\Theta_j) = p(\Theta_i|\Theta_j) S_W(\Theta_i) \quad \text{for } i = 1, 2, ..., 5, \qquad (9)$$

where $S_W(\Theta_i)$ is computed from (7), $p(\Theta_i|\Theta_j)$ denotes the probability to go to the sector Θ_i knowing that the sound source was previously located in the sector Θ_j, and $S_W(\Theta_i|\Theta_j)$ is the weighted score for the source to be

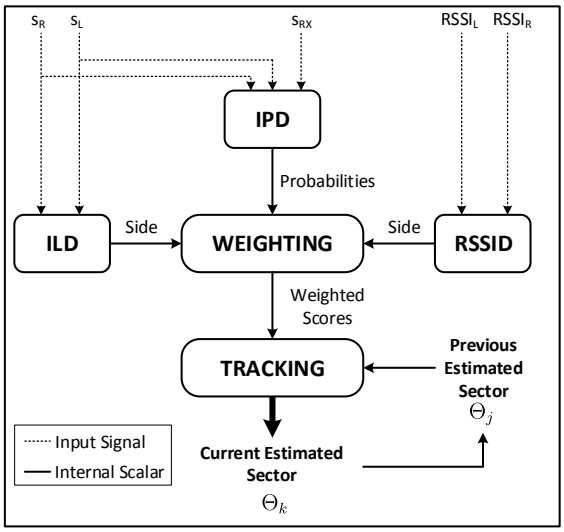

Fig. 3. Overview of the processing performed in the reported localization algorithm.

in the sector Θ_i with the knowledge of the previous sector. Finally, the current location estimation Θ_k (spatial sector) for the observations of IPD, ILD and RSSID, considering the previous location Θ_j of the source is:

$$\Theta_k = \underset{\Theta_i, i \in \mathbb{N}_5^*}{\arg\max}\, S_W(\Theta_i|\Theta_j). \qquad (10)$$

The global processing performed in the reported BLA is summarized in Figure 3.

3. EXPERIMENT

The reported BLA has been assessed with a database of signals recorded in an empty classroom (130 m^3). Two manikins were used and acted as the speaker (HATS, B&K 4128) and the listener (KEMAR, GRAS). The distance between the two manikins was constant and set to 4 m. The microphone of the digital wireless system was worn by the HATS manikin. A RF receiver was plugged into each hearing aid (Phonak Naida IX SP) worn by the KEMAR. The recorded data were the left and right hearing aid microphone signals, the demodulated audio signal from the body-worn microphone, and the left and right RSSIs for each incoming frame.

Two stimuli were used in this study. They consisted in speech signal spoken by a male or a female. The male voice was a 18-second sample of the English-spoken sentence available on the EBU SQAM CD. A 18-second extract from the ISTS V1.0 was chosen as the female voice stimulus. The ISTS V1.0 consists of a mixture of 21 female speakers in 6 different languages (American English, Arabic, Chinese,

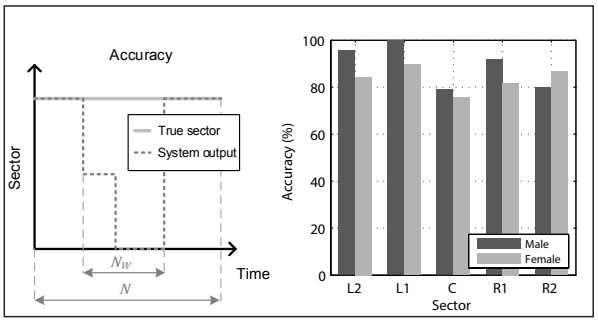

Fig. 4. Left panel: Accuracy as defined in this study, according to (11). Right panel: Accuracy scores of the algorithm for a female and male stimulus in a classroom, with a 4-meter distance between the speaker and listener.

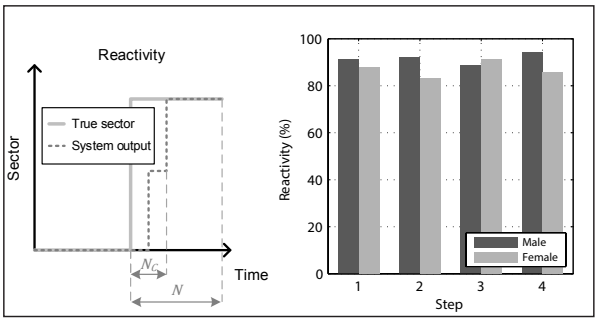

Fig. 5. Left panel: Reactivity as defined in this study, according to (12). Right panel: Reactivity scores of the algorithm for a female and male stimulus in a classroom, with a 4-meter distance between the speaker and listener.

French, German, Spanish), made available by the EHIMA (European Hearing Industry Manufacturers Association). These stimuli were recorded at a 48 kHz sampling frequency, which was reduced to 16 kHz in post-processing. The SNR measured at the center of the KEMAR's head was 10 dB.

The KEMAR was mounted on a turntable, and rotated from -90° to 90° by 10° steps. For each position, the two stimuli were successively emitted by the HATS and the signals were recorded.

4. RESULTS

4.1. Accuracy

The accuracy A describes how well the estimated spatial sector matches the actual angle of the speaker. A is obtained by counting the number of frames leading to an erroneous localization. It is mathematically computed as follows:

$$A(\%) = 100 \times \frac{N - N_W}{N}, \tag{11}$$

where N is the total number of analyzed frames and N_W is the number of wrong frames, as shown on the left panel of Figure 4. The accuracy is derived in several azimuths. It is then averaged in sectors, in order to get an accuracy score in each spatial sector for the male and female stimuli. This is depicted on the right panel of Figure 4. No significant effect of the stimuli was found on the accuracy, according to a paired-sample *t*-test ($t(4) = 1.7169$, p $= 0.1611$). The scores appear to be the lowest in the central sector, presumably because of the absence of contribution from the ILD and RSSID in this area. The asymmetry of the accuracy between the left and right sectors must be due to the presence of a wall on a side, and of large windows on the other, resulting in different acoustical and RF reflections. The accuracy averaged over all sectors for the male and female stimuli is 86.5%.

4.2. Reactivity

The reactivity R describes how fast the algorithm converges to the correct sector. R is obtained by counting the number of frames required to reach the actual sector after a transition. It is mathematically computed as follows:

$$R(\%) = 100 \times \frac{N - N_C}{N}, \tag{12}$$

where N_C is the number of analyzed frames before the correct sector is displayed, as shown on the left panel of Figure 5. The reactivity is derived using the concatenation of two recordings from two different spatial sectors, with various sector steps (one to four). The term N in (12) is an arbitrary time reference and has to be similar for all configurations, in order to allow comparison between cases. The reactivity is defined in such a way that a high value corresponds to a low reaction time, and conversely. The actual convergence time (in seconds) can be recovered knowing the duration corresponding to N frames (half stimuli duration, see Figure 5), equal to 9 seconds in this study. For instance, a reactivity of 80% corresponds to a convergence time of 1.8 seconds ($9 - 0.8 \times 9 = 1.8$).

The reactivity obtained for the male and female stimuli and sector step from 1 to 4 is depicted on the right panel of Figure 5. A paired-sample *t*-test revealed no significant effect of the stimuli on the results ($t(3) = 1.6853$, p $= 0.1905$). The reactivity appears to be similar whatever the sector step size, and its averaged value over all steps for male and female stimuli reaches 89.3% (0.96 second) with a standard deviation of 3.6% (0.32 second) among cases.

5. CONCLUSION

The algorithm reported in this paper is able to localize and track the position of the speaker when using a digital wireless

microphone system for hearing aids. It is based on the estimation of three spatial cues: two acoustical cues (IPD and ILD) and one RF cue (RSSID). The talker is localized in one of the five spatial sectors defined in the frontal horizontal plane relative to the listener. A tracking procedure governed by a probabilistic network improves the stability of the system.

Two metrics have been defined to assess the performance of the algorithm: the accuracy and reactivity. An ideal BLA would reach a score of 100% for both accuracy and reactivity. In reality, a trade-off has to be found between these two components. Indeed, an increase of the accuracy usually results in a lower reactivity, and conversely. In the specific test condition of this study (speech signal, classroom, 4-meter distance between speaker and listener), the algorithm displays the correct spatial sector almost nine times out of ten, after a reaction time of approximately 1 second. This performance was satisfying for the targeted application and shows that the combined use of acoustical and RF cues is a powerful tool to counteract the detrimental effect of acoustic noise and reverberation.

The developed algorithm is fully compatible with the constraints related to hearing aids, and has been implemented in embedded C-code. The localization processing takes less than 6 ms for each 128-sample analysis frame (8 ms length at 16 kHz), and the memory storage requirement is low, since no specific database is needed.

ACKNOWLEDGMENTS

This study was funded by the Commission for Technology and Innovation (CTI) of the Swiss Federal Department of Economic Affairs, Education and Research (EAER), with grant number 14550.1 PFNM-NM, in collaboration with Phonak Communications AG.

REFERENCES

[1] G. Courtois, P. Marmaroli, H. Lissek, Y. Oesch, and W. Balande, "Implementation of a binaural localization algorithm in hearing aids: specifications and achievable solutions," in *The 136th Convention of the Audio Engineering Society*. 2014, Audio Engineering Society.

[2] F. Keyrouz and K. Diepold, "An enhanced binaural 3D sound localization algorithm," in *Proceedings IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2006, pp. 662–665.

[3] F. Keyrouz, Y. Naous, and K. Diepold, "A new method for binaural 3D localization based on HRTFs," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006, vol. 5, pp. V341–V344.

[4] J. A. MacDonald, "A localization algorithm based on head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 123, no. 6, pp. 4290–4296, 2008.

[5] D. S. Talagala, W. Zhang, T. D. Abhayapala, and A. Kamineni, "Binaural sound source localization using the frequency diversity of the head-related transfer function," *J. Acoust. Soc. Am.*, vol. 135, no. 3, pp. 1207–1217, 2014.

[6] C. Lim and R. O. Duda, "Estimating the azimuth and elevation of a sound source from the output of a cochlear model," in *28th Asimolar Conference on Signals, Systems and Computers*, 1994, vol. 1, pp. 399–403.

[7] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2914–2919, 1999.

[8] D. Li and S. E. Levinson, "A bayes-rule based hierarchical system for binaural sound source localization," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003, vol. 5, pp. V–521.

[9] J. Nix and V. Hohmann, "Sound source localization in real sound fields based on empirical statistics of interaural parameters," *J. Acoust. Soc. Am.*, vol. 119, no. 1, pp. 463–479, 2006.

[10] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Trans. Syst. Man Cybern.*, vol. 36, no. 5, pp. 982–994, 2006.

[11] E.M. von Hornbostel and M. Wertheimer, "Über die wahrnehmung der schallrichtung," *Sitzungsberichte der preussischen Akademie der Wissenschaften*, vol. 388, pp. 396, 1920.