IDENTIFICATION AND TRACKING OF ACTIVE SPEAKER'S POSITION IN NOISY ENVIRONMENTS

Tomasz M. Rutkowski, Masahiro Yokoo,
Keisuke Yagi, Yoshinari Kameda[†],
Koh Kakusho, Michihiko Minoh

Academic Center for Computing
and Media Studies
Kyoto University, Kyoto, Japan
http://www.mm.media.kyoto-u.ac.jp/

Danilo P. Mandic

Department of Electrical and Electronic
Engineering, Imperial College of Science,
Technology and Medicine,
London, United Kingdom

ABSTRACT

Due to a large size of lecture theaters and associated multiple sound sources, the common beamforming techniques cannot be straightforwardly applied in those environments. Other difficulties include significant propagation delays between the microphones and the directionality mismatch between speakers and microphones. These problems are particularly emphasized in the distance lecturing environment equipped with a large scale microphone array (distance between neighboring microphones 1.86m). To reduce some of these effects, we propose a technique based upon a combination of noise reduction and active speaker tracking. Experiments in a real teleconferencing environment are provided to support the analysis.

1. INTRODUCTION

The core of modern multimedia distance learning and virtual presence (teleconferencing) applications rests upon the clear and realistic capture of the sound signal. An ideal solution should preserve the spatial hearing comfort, together with the robust reduction of environmental and other noise signals. In distance learning applications, therefore, it is an imperative that the voice of a lecturer and students in the classroom be captured at a level that provides features for clear understanding at the far end. To that cause, it is desired that every participant in a distance lecture is equipped with a portable microphone placed close to the mouth. On the other hand, for a large number of participants, e.g. the microphone could be passed among the students, but such a solution would seriously hamper the flow of discussion, whereas the recording/transmission system could be disturbing to the participants. To balance the need for the voice capture of an active speaker and the problems associated with a large size of the classroom, a convenient solution would be a ceiling–mounted large–scale microphone array, which covers the area occupied by the audience. Since the microphones mounted above the students inevitably capture and often enhance both the *nonspoken* activity and noise, it is of crucial importance to employ appropriate filtering and sound separation techniques, which is the focus of this paper. To start off with, in the next section we discuss preprocessing modules for localization of the active speaker in the classroom.

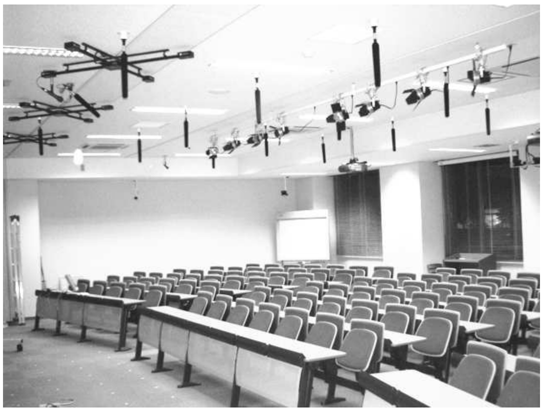

Fig. 1. A multimedia classroom at the Media Center of Kyoto University designed for distance lectures and equipped with a microphone array installed above the students area.

2. NOISE REDUCTION AND SPEAKER LOCALIZATION

The set–up of the recording and active speaker localization situation which we consider is illustrated in Fig.1. and can be described as follows. Consider two or more participants

[†] Dr. Kameda is currently with Institute of Engineering, Mechanics and Systems, University of Tsukuba, Japan.

in a remote multimedia session who would like to have a trouble–free and realistic communication. Positions of the participants in the remote room are arbitrary within the area where the sound can be physically captured. The rooms themselves can be equipped with air conditioners or other devices generating unwanted sound sources and noise. Additionally, the participants may use portable devices (e.g. laptops) that are the sources of specific local noise.

2.1. Signal predictability for speech enhancement

The significant distance between the microphones causes them to capture much of ”local” noise not shared across the microphone array. To remove such local interferences that usually degrade the performance of speaker localization algorithms, we employ linear adaptive predictors whose filter lengths, for simplicity, are limited to several taps. The output of every section of an adaptive predictor $x_i(k),\ i = 1, \ldots, N$, where N is the number of microphones $s_i(k)$ in the array, is associated with a prediction error $e_i(k)$ [1]. The operation of linear adaptive predictors, together with the coefficient update ($\mathbf{w}_i(k)$) is given by

$$e_i(k) = s_i(k) - x_i(k), \tag{1}$$

$$x_i(k) = \sum_{j=1}^{N} s_i(k-i)w_i(k) = \mathbf{s}_i^T(k)\mathbf{w}_i(k), \tag{2}$$

$$\mathbf{w}_i(k+1) = \mathbf{w}_i(k) + \mu(k)e_i(k)\mathbf{s}_i(k), \tag{3}$$

$$\mu(k) = \frac{\lambda_w(k)}{\|\mathbf{s}_i(k)\|_2^2}. \tag{4}$$

The dynamical learning rate adaptation parameter $\lambda_w(k)$ is the critical variable toward catering for the unknown dynamics of the recorded signals. Since our desire is to enhance voices of students from the auditorium area that ask questions during the lecture, the focus of the following analysis is to identify the speaker location and efficiently suppress the additive noise. To that end, there are known approaches in the open literature, from which for an accurate identification of the noise region, a technique called the spectral subtraction technique might be employed [2]. Such techniques, however, although very effective, might introduce the so called *musical noise*, which contributes to a very bad auditory sensation and consequently creates rather artificial auditory effects. To avoid such problems, several adaptive step size normalized least mean square (NLMS) based algorithms have been proposed [3, 4]. In our approach, we employ the adaptive step size for NLMS $\lambda_w(k)$, proposed in Eq.4, which is obtained from the cepstral voice activity detector. Speech and noise are assumed to be mutually statistically independent, therefore the spectrum of the enhanced speech signal $\left|\tilde{S}\left(\omega\right)\right|$ can be obtained from the noisy version $\left|X\left(\omega\right)\right|$, after subtracting the noise spectrum estimate $\left|N\left(\omega\right)\right|$, calculated from regions labelled as noise [2]. As mentioned above, these techniques require a very good voice activity detector and are prone to causing side effects. We therefore use this technique only to detect the possible sound activity which has a different spectral image from the noise spectrum (speech, music, etc.). The spectrum of such a signal with noise subtracted can be expressed as

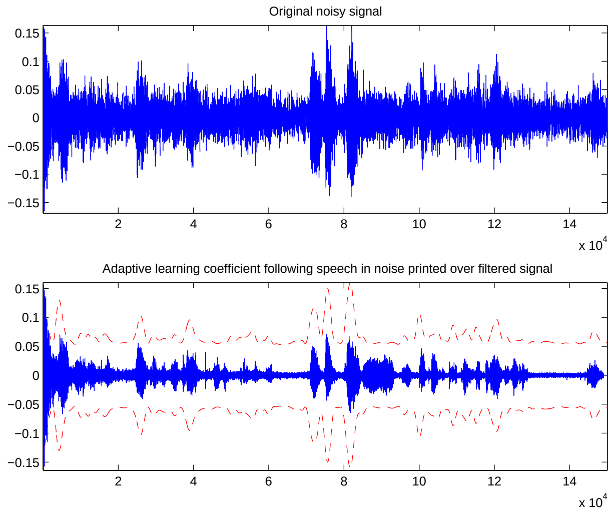

Fig. 2. Top diagram: the noisy original speech signal, cepstrally preprocessed for $\lambda_w(k)$ estimation. Bottom diagram: the cepstrally evaluated speech activity identifier used for NLMS adaptation ($\lambda_w(k)$) plotted over the preprocessed speech.

$$\left|\tilde{S}\left(\omega\right)\right|^2 = \begin{cases} \left|X\left(\omega\right)\right|^2 - \left|N\left(\omega\right)\right|^2, & \text{if } \left|X\left(\omega\right)\right|^2 > \left|N\left(\omega\right)\right|^2, \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

In order to remove the ”stationary–in–time” interference (stable noise) from a time window at time k, the power of the subtracted signal $\tilde{S}(k)$ can now be evaluated over the time window and compared to the total (with noise) power S_{total} to update the the value of $\lambda_w(k)$, as

$$\lambda_w(k) = \begin{cases} \tilde{S}(k)/S_{total}(k), & S_{cs}(k) \geq \nu_{room}, \\ const, & \tilde{S}(k) < \nu_{room}, \end{cases} \tag{6}$$

In the above equation, $\tilde{S}(k)$ is the cepstrally subtracted signal power and $S_{total}(k)$ is the total power in the windows around time sample k [2]. Value ν_{room} reflects the background room noise and is calculated during the system installation and initialization. The value of this threshold may also be evaluated before the distance lecture session starts. To speed up the filter coefficient adaptation for the time windows where voice activity is not detected, we opt for the use of constant $\lambda_w(k)$ from Eq.4. An illustrative example of the effects of speech activity evaluation and $\lambda_w(k)$ estimation is shown in Fig.2.

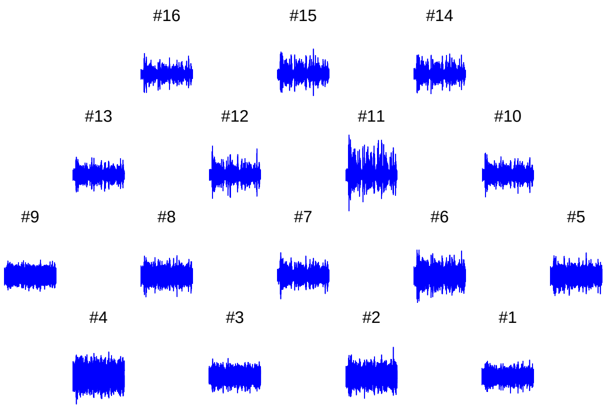

Fig. 3. The original recordings from 16 microphones located on the ceiling of a classroom.

2.2. Active speaker localization

The preprocessed (locally filtered) sound waveforms are now much better suited for the next stage: active speaker localization (compare the plots of the raw data from Fig.3. and their filtered versions from Fig.4). Since the high power noisy interference is removed at in this stage a voice detector with delay and speech loudness estimation facilities can be employed. For the estimation of the time frames containing spoken utterances we employ the technique described in ITU-T P.56 [5] recommendation. Only the time frames with active voice are considered for postprocessing. To speed up the search for such frames, only neighboring microphones are compared in the estimation of time and power differences. After the adaptive preprocessing, the localization unit was able to detect the correct three neighboring microphones. The result is presented on Fig.6. with signal and tracking diagrams before and after adaptive noise reduction. The position localization is accurate and remains stable as can be seen on the lower left side plots of the above figure.

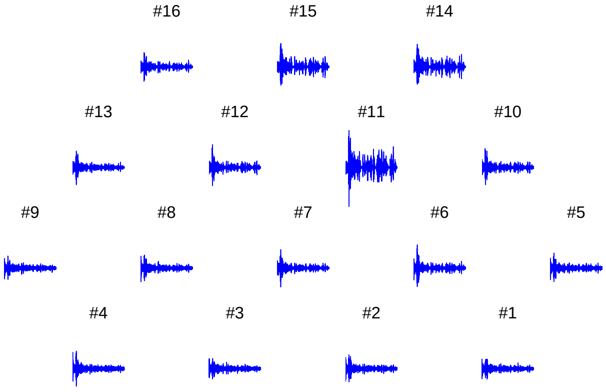

Fig. 4. The filtered (denoised) recordings from 16 microphones as the input to speaker localization.

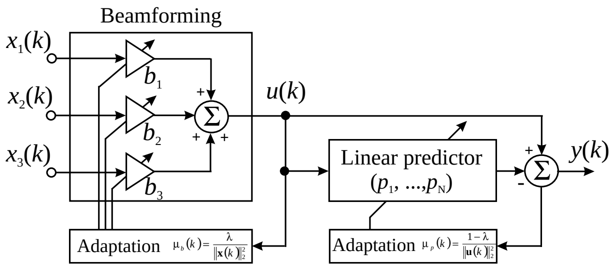
Beamforming
$x_1(k)$
$x_2(k)$
$x_3(k)$
b_1
b_2
b_3
+
+
+
Σ
$u(k)$
Linear predictor
$(p_1, ..., p_N)$
+
-
Σ
$y(k)$
Adaptation $\mu_b(k)=\frac{\lambda}{\|\mathbf{x}(k)\|_2^2}$
Adaptation $\mu_p(k)=\frac{1-\lambda}{\|\mathbf{u}(k)\|_2^2}$

Fig. 5. The overview of the final step of the local beamformer based upon the predictability of the speech signal.

3. LOCAL BEAMFORMING AS A SECOND STEP OF SPEECH ENHANCEMENT

After the localization of the speaker position, we perform the final step of speech enhancement. In the first step of speaker localization, our approach locates just the three neighboring microphones with limited delays, since the longest distance between two opposite microphones is up to eight meters. Once the speaker location is identified, the local beamforming approach can be employed, since the delays between microphones and resulting convolutive mixtures are no longer very significant. A detailed derivation of this approach can be found in [6]. Here, we present only the final stage of this algorithm, which performs blind extraction of signals based upon predictability constraints. Such a blind signal extraction scheme is shown in Fig.5. To suit the approach proposed in this paper, three (triangulation like procedure) microphones, indicating the highest power of spoken utterances, are taken into account. Since the location of the speaker is already identified, the delays of captured speech are not significantly different. To remove the remaining interferences in the second processing step, we employ a combined beamforming and linear prediction form [6], given by

$$y(k) = \sum_{i=1}^{M} b_i(k)x_i(k) - \sum_{j=1}^{N} p_j(k) \sum_{i=1}^{M} b_i(k-j)x_i(k-j), \tag{7}$$

Here we present only the final steps that lead to the update of the adaptive beamformer, given by

$$\mathbf{b}(k+1) = \mathbf{b}(k) + \mu_b(k)y(k)\mathbf{b}(k), \tag{8}$$

where:

$$\mu_b(k) = \frac{\lambda}{\|\mathbf{x}_i(k)\|_2^2}, \tag{9}$$

The update of a postprocessing linear predictor that is combined with the beamformer in order to reduce the possible

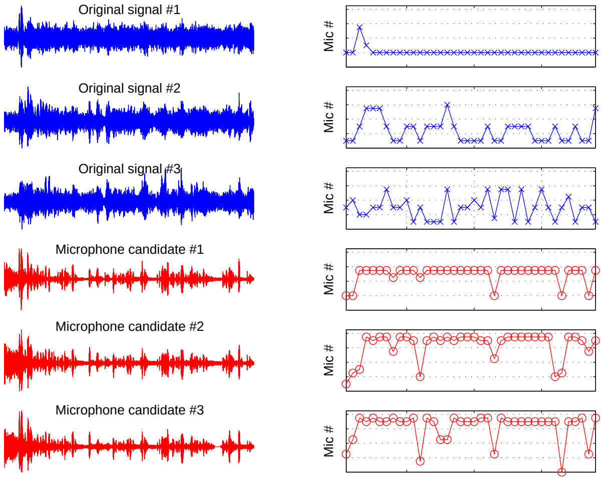

Fig. 6. Three microphones which are candidates for the beamformer input. Since the first step of noise reduction was to localize the speaker position, the situation is still noisy, however the localization procedure was already able to perform accurately. The three top rows of the plots present the candidates for a steady speaker in an ascending order. The three bottom rows of plots present the preprocessed speech (the noise is still present) but the locations presented on the right sides are steady and coherent, suggesting three neighboring microphone all the time.

convolutive effects is given by

$$\mathbf{p}(k+1) = \mathbf{p}(k) + \mu_p(k) \tag{10}$$

where:

$$\mu_p(k) = \frac{1-\lambda}{\|\mathbf{u}(k)\|_2^2}, \tag{11}$$

where for both cases above $0 \leq \lambda \leq 1$.

4. CONCLUSIONS AND FURTHER REMARKS

The proposed methodology has been shown to be suitable and to provide successful speaker localization for large scale microphone arrays, where multiple sound sources are recorded (see result on Fig.7.). We have proposed a two stage approach. Firstly, we locally equalize the captured waveforms to be able to locate the active speaker. In the next step, since we can identify the closest microphones, to finally equalize the signal, local beamforming equipped with the second stage based upon adaptive prediction is utilized. To support the analysis, experimental results in a real teleconferencing environment are provided.

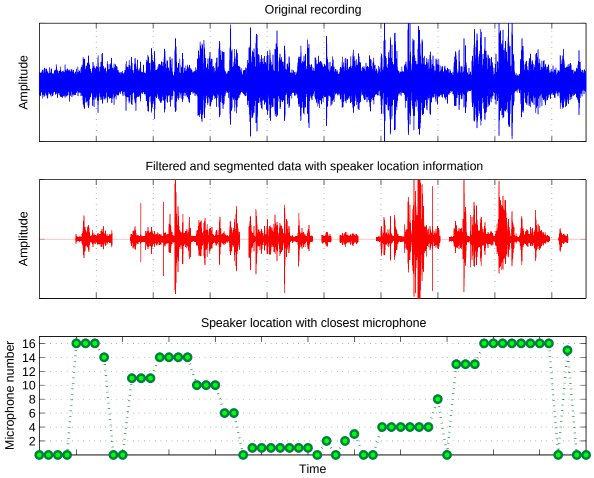

Fig. 7. The example of a single channel speech output from the walking speaker (this is usually the most difficult case). The top diagram presents the composition of original signals according to speaker location presented in the bottom diagram (position was localized after first stage of speech signal preprocessing). The middle diagram shows the speech signal after our approach has been applied, exhibitting effectively suppressed non-speech regions. The path of walking speaker was correctly reconstructed from the microphones locations.

5. REFERENCES

[1] D. Mandic and J. Chambers, *Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability*, John Wiley & Sons, 2001.

[2] H. Gustafsson, S.E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 799–807, November 2001.

[3] S. Gollamudi, S. Nagaraj, S. Kapoor, and Y-H. Huang, "Set-membership filtering and a set-membeship normalized LMS algorithm with an adaptive step size," *IEEE Signal Processing Letters*, vol. 5, no. 5, pp. 111–114, May 1992.

[4] S. Gazor and K. Shahtalebi, "A new NLMS algorithm for slow noise magnitude variation," *IEEE Signal Processing Letters*, vol. 9, no. 11, pp. 348–351, November 2002.

[5] ITU-T Telecommunication Standardization Sector of ITU, "Telephone transmission quality objective measuring apparatus - objective measurement of active speech level," ITU-T Recommendation P.56, International Telecommunication Union, 1994.

[6] D. P. Mandic and A. Cichocki, "An online algorithm for blind extraction of sources with different dynamical structures," in *Proceedings of 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, S. Amari, A. Cichocki, and N. Murata, Eds., Nara, Japan, April 2003, pp. 645–650.