PASSIVE SOUND SOURCE LOCALIZATION SYSTEM

Jarosław Baszun

Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: Acoustic source localization system for speech signals based on five microphone array was developed. Three dimensional position computation is based on time delay estimation between pairs of microphones. The psyhoacoustically motivated voice activity detector was used to robustly determine activity of speaker in presence of background noise. The detector was based on modulation properties of human speech. Good performance was obtain by selecting frames with speech and nulling frequency bands without speech components. As the result more precisely computation of the time delay was possible. Real experiments shown good immunity of the proposed algorithm to noise and reverberation.

Keywords: phase transform, source localization, microphone arrays.

1. Introduction

Location of sources of waves using array of sensors is the important field of research in radar, seismology and sonar systems. Also similar techniques were developed over for four decades in acoustics. The knowledge about spatial position of sound source can be useful in many audio applications such as automatic camera tracking for video conferencing, suppressing noise and reverberation in voice control for robots hearing systems and audio surveillance. This work concerns the tracking of voice source.

Localization methods in acoustics can be divided into three categories: steered beamformer, high-resolution spectral estimation and time delay estimation based techniques. The most widely used localization techniques are based on time delay estimation in which localization systems computes the location of source in two step process. In the first step a set of time delay of arrivals (TDOA) among different microphone pairs is calculated. The relative time delay for each pair of microphones is determined. In the second step this set is used to estimate the acoustic source location based on knowledge of used microphone array geometry. To perform this different methods can be used to source position calculation: e.g. the triangulation, the maximum likehood method, the spherical intersection method, the spherical interpolation

Zeszyty Naukowe Politechniki Białostockiej. Informatyka, vol. 7, pp. 5-16, 2011.

method [9]. In time delay estimation approach an important role plays a parametric model for an acoustical environment. Usually two models are used: free-field model and reverberation based model. The time delay estimation algorithm estimates TDOA based on the model.

In this paper passive voice localization system was proposed based on computation TDOA using generalized cross-correlation method with modifications which allow to distinguish between speech and non speech signals in time-frequency domain. The speech to noise estimate is computed in modulation frequency domain for each band separately and used as a feature for speech-pause detection. This psychoacoustically motivated voice activity detector was integrated with computation of weighting function for cross-correlation. Detecting of speech content for each frequency band separately allows for nulling non speech components to avoid influence of disturbing sources on position computation.

2. Time delay estimation

The problem of finding the distance between the sound source and the microphone array is usually defined for near-field case as shown in Fig. 1, but is also possible under some conditions for far-field case. The radius of near-field for array of microphones is defined

$$R_{nf} = \frac{2R_a^2}{\lambda}, \tag{1}$$

where R_a is the is the size of the array and λ is the wavelength of the operating frequency.

In such situation it is always possible to estimate angle of arrival for wave and the distance between the source and microphones. The time difference of arrival (TDOA) for the pair of microphones is

$$\tau_{1,2} = \frac{r_2 - r1}{c}, \tag{2}$$

Where c is a speed of sound calculated based on air temperature t_{air} in deg. Celsius, from formula [9]:

$$c = 331 + 0.61 t_{air}, \tag{3}$$

If the distance between microphones is know it is possible to calculate unknown parameters r_1, r_2,..., θ_1, θ_2,.... When information about TDOA is available it is

possible to calculate position of source in relation to the array using for example triangulation rule.

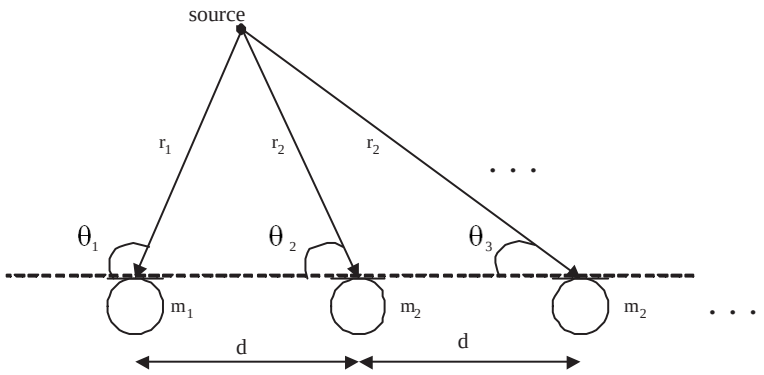

Fig. 1. Linear array

Many approaches can be applied to locate source of signals. In case of multiple narrow band sources, methods based on the eigenvalue analysis of the spatial covariance matrix of the signals from matrix of sensors are commonly used. Such methods were successfully applied especially for radiolocation technology. For wideband signals where we cannot assume the hypothesis about statistical properties of the source of signal and interferences other solutions must be used.

In acoustics we can distinguish two main methods. The first approach is based tuned beamformer. The algorithm scans a set of positions to find the place where the maximum acoustic power is received by the array of sensors. This method have some disadvantages: the computational complexity and the poor resolution in space and in time for moving objects. The advantage of this method is possibility of creation of acoustics maps for objects search [5].

Second approach is based on computation of Time Difference of Arrival (TDOA) between pairs of sensors - microphones. There is a lot of literature on this approach e.g. [2]. One of the basic method of computing of time delay between two signals is to compute maximum of cross-correlation function. But this simple approach gives bad results in case of narrowband signals and in presence of strong reverberation. To overcome this shortage methods of pre-filtering of signals can be applied in case were statistics of source and noise is known or in case when statistics is unknown an effective method is based on whitening the input signals, so only the

phase information in cross-power spectrum of the two signals is used. Such methods are known as Generalized Cross-Correlation (GCC) [10].

Different models can be employed to describe an acoustic environment in the TDOA problem [9]. The ideal model it is assumed that the signal acquired by each sensor is a delayed and attenuated version of the original source signal plus additive noise. This model takes into account the direct signal path only and do not consider multipath signal propagation encountered in many real environments due to reflections. Much more realistic is the multipath model in which received signal is described as a sum of direct signal and weighted sum of delayed reflections [12]. This multipath effect is widely used in the oceanic propagation environment. The drawback of the multipath propagation model is the difficulty to estimate time difference of arrival for pairs of sensors in case of many different paths. So more realistic model for room acoustic environment seems the reverberation model in which for the source signal $s(t)$, the signal received at the two sensors can be described as follows:

$$\begin{aligned} x_1(t) &= h_1 * s(t) + n_1(t), \\ x_2(t) &= h_2 * s(t) + n_2(t) \end{aligned} \tag{4}$$

Where $x_1(t)$, $x_2(t)$ - received signals, h_1, h_2 represent reverberations and n_1, n_2 are noise signals received at two sensors. It is assumed additive noise conditions. This model in case of week reverberation can be simplified to the following model:

$$\begin{aligned} x_1(t) &= k_1 s(t) + n_1(t), \\ x_2(t) &= k_2 s(t+D) + n_2(t), \end{aligned} \tag{5}$$

where $s(t)$ is a source signal , $x_1(t)$, $x_2(t)$ - received signals, k_1, k_2 are certain weights, D - the delay of the signal arrival between the two sensors and $n_2(t)$, $n_2(t)$ are additive noise. It can be shown that for slowly changed environment parameters the cross-correlation function of signals $x_1(t)$ and $x_2(t)$ can be used to determine the time delay D:

$$R_{x_1x_2}(\tau) = E[x_1(t)x_2(t-\tau)], \tag{6}$$

where E denotes expectation. For the model from Eq. 5 assuming that noise is not correlated with the signal $s(t)$ the cross-correlation is:

$$R_{x_1x_2}(\tau) = k_1k_2R_{ss}(\tau - D) + R_{n_1n_2}(\tau), \tag{7}$$

The cross-power spectrum of (the Fourier transform of cross-correlation) is:

$$G_{x_1x_2}(\omega) = k_1k_2G_{ss}(\omega)e^{-j\omega D} + G_{n_1n_2}(\omega) \tag{8}$$

Background noise can be correlated due to the fact that it is produced by single source e.g. computer fan and can be estimated using energy detector and next subtracted. The cross power spectrum with subtracting noise becomes

$$\hat{G}_{x_1x_2}(\omega) = G_{x_1x_2}(\omega) - G_{n_1n_2}(\omega) = k_1k_2G_{ss}e^{-j\omega D} \tag{9}$$

This leads to normalized cross-correlation called Phase Transform (PHAT) [4], [10]:

$$\hat{R}_{x_1x_2}(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\hat{G}_{x_1x_2}(\omega)}{|G_{x_1x_2}(\omega)|}e^{j\omega\tau}d\omega = (\tau - D)\,. \tag{10}$$

Weighting factor:

$$\Psi(\omega) = \frac{1}{|G_{x_1x_2}(\omega)|}, \tag{11}$$

cause the whitening of the input signals. In effect only the phase information in the cross-power spectrum of the two signals is used. Because this operation weights $\hat{G}_{x_1x_2}(\omega)$ as the inverse of $G_{s_1s_2}(\omega)$ so errors arise for frequencies where signal power is small in compare to interferences. As a result in case of lack of signal source $s(t)$ knowledge appropriate weighting in spectrum domain of signals from sensors is required to avoid influence of this errors.

The TDOA between two microphones can be found by selecting the maximum location of the Eq. 10:

$$D = argmax_\tau \hat{R}_{x_1x_2}(\tau), \tag{12}$$

In [14], [3] was shown that phase correlation approach can be used also in case of reverberation using some additional processing.

3. Voice activity detector

In this system psyhoacoustically motivated voice activity detector (VAD) was used for two proposes: to select frames with speech and to select frequency bands where speech signal is dominant to minimalize noise influence on time delay calculation. This voice detector is an expansion of the idea developed in previous work in this area [1]. The detector exploits properties of modulation spectrum of human speech [6], [11]. It is known that modulations of sound are the carrier of information in speech. The background noise encountered in real environments is usually stationary

or changing differently in compare to the rate of change of speech. Modulation components of speech are mainly concentrated in range between 1 and 16 Hz with higher energies around $3-5$ Hz what corresponding to the number of syllables pronounced per second [8]. Slowly-varying or fast-varying noises will have components outside the speech range. Further, steady tones will only have constant component in modulation domain. Additive noise reduces the modulation peak in speech modulation spectrum. System capable of tracking speech components in modulation domain allows to distinguish between frequency bands with dominant speech signal and band with dominant background noise. This operation is the key element of effective computation of GCC-PHAT algorithm because it is possible to set to zero signal in bands classified as noise.

The block diagram of the voice activity detector was shown in Fig. 2. Signal from microphone with sampling frequency 16 kHz is split into $M = 512$ frequency bands using Short Time Fourier Transform (STFT) with Hamming window and 25 % overlapping. Next amplitude envelope is calculated for first 256 bands:

$$y_k(nM) = \sqrt{Re^2[x_k(nM)] + Im^2[x_k(nM)]} \tag{13}$$

Then amplitude envelope is filtered by passband IIR filters with center frequency 3.5 Hz and frequency response shown in Fig. 3. The output of the filters is half-wave filtered to remove negative values from output of the filters. The following computation is carried out on the filtered and not filtered envelopes:

$$S(nM) = \frac{Y'}{Y - mean(Y) - Y' - mean(Y')} \tag{14}$$

Above parameter is an estimate of speech to noise ratio for each of analyzed channels. Mean value of filtered and nonfiltered envelope is computed based on exponential averaging with time constant approximately 1 s. Then all channels are summed and the square of this estimate is used as a classification parameter for voice activity detector. Speech decision is based on comparison between classification parameter and the threshold computed based on the following statistics [13]:

$$Thr = mean(d) + \alpha \cdot std(d) \tag{15}$$

where d is a classification parameter and α controls confidence limits and is usually in the range 1 to 2, here was set to be equal 2. Both mean value and standard deviation is estimated by exponential averaging in pauses. Frame is considered to be active if value of the classifier is greater than the threshold. Speech to noise computed

parameter for each channel in combination with the threshold is used to select channels with speech signal and nulling channels with noise in time delay computation algorithm.

To avoid isolated errors on output of VAD caused by short silence periods in speech or short interferences correction mechanism described in [7] was implementing. If current state generating by the VAD algorithm does not differ from n previous states then current decision is passed to detector output otherwise the state is treated as a accidental error and output stays unchanged.

4. Implementation of time delay estimation algorithm

In Fig. 4 block diagram of time delay estimation algorithm was shown for two channels x_1 and x_2. Signals from both sensors are grouped into frames. One of the channels from microphone array is used by voice activity detector to calculate which frame contain speech and in what channels the speech signal is present. When speech signal is detected power spectra of signals for pair of channels are calculated using Fast Fourier Transform (FFT). Then cross-spectrum is calculated. The cross-spectrum of signal is averaged over several frames. The averaged cross-spectrum is then normalized according to equation:

$$G'_{x_1x_2}(\omega) = \frac{\hat{G}_{x_1x_2}(\omega)}{|G_{x_1x_2}(\omega)|}. \tag{16}$$

The normalized cross-spectrum $G'_{x_1x_2}(\omega)$ of the frequencies that were classified as non speech components are set to zero. Then inverse FFT is calculated on the averaged and normalized cross-spectrum. In classical GCC-PHAT algorithm the time delay is chosen as the lag that corresponds to the maximum of the normalized cross-correlation function. To increase resolution of time delay estimation three sample interpolator was implemented. The maximum value from the normalized cross-correlation function is selected and its both sides neighbours, as shown in Fig. 5. Therefore, time delay D becomes:

$$D = \frac{A_{i-1}t_{i-1} + A_i t_i + A_{i+1}t_{i+1}}{A_{i-1} + A_i + A_{i+1}} \quad (2 \leq i \leq N-1), \tag{17}$$

where A_i is the largest value of the normalized cross-correlation function and t_i corresponding time delay. This method makes it possible to calculate time delay more accurately without a large number of FFT samples.

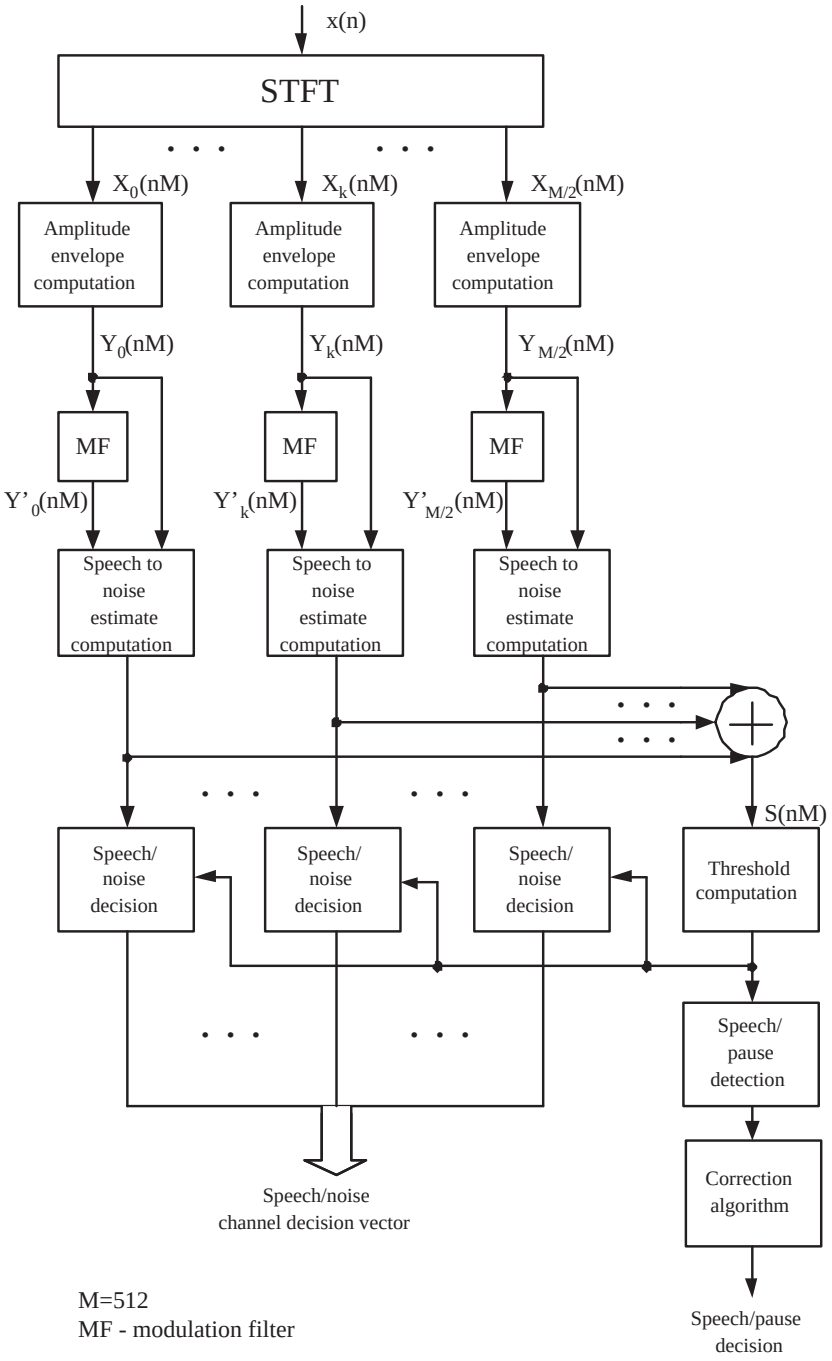

Fig. 2. Block diagram of the voice activity detector (VAD) based on modulation properties of speech

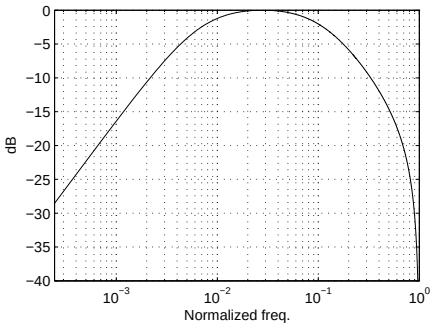

Fig. 3. Magnitude frequency response of modulation filter (MF)

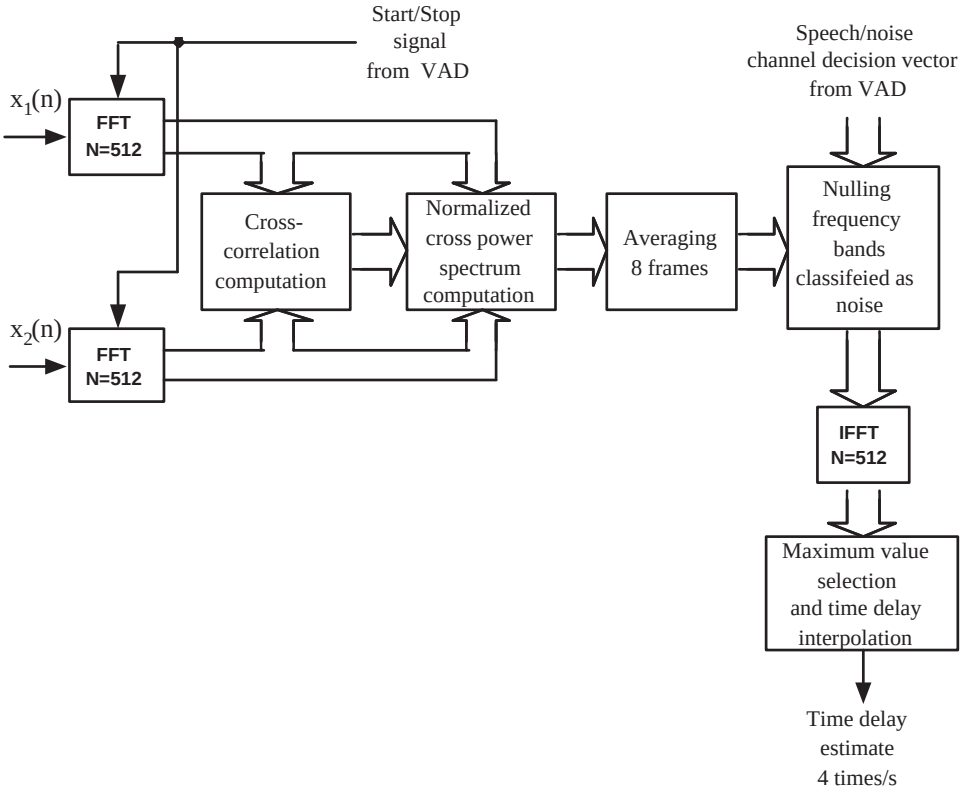

Fig. 4. The time delay estimation algorithm block diagram for two channels

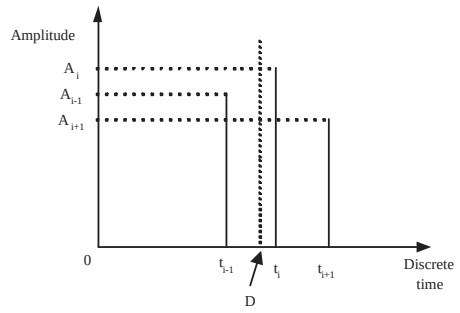

Fig. 5. The time delay interplation method

5. Position calculation and results

The microphone array used for voice source localization was shown in Fig. 6. Two edge pair of microphones 1 and 3 are used for azimuth calculation by triangulation. Pairs 1,2, 2,3 and 2,5 are used for the depth calculation. Using the microphone pair 2,5 it is possible to distinguish between sources localized in front and behind the array. Experiment was carried out in room 7m x 5.5m x 2.8m. Sampling rate was 16 kHz, frame 512 samples, 8 frames ware averaged to calculate time delay sample. In Table 1 measured distances for three positions was shown.

Table 1. Distance measurements and standard deviation for averaged 20 measurements

Distance (m)	Std. deviation	Azimuth (deg)	Elevation (deg)
2.21	0.05	-24.7	-15.4
3.08	0.2	20.3	-10.2
5.12	0.32	15.2	2.5

For elevation angle calculation pair 2,4 was used. For this pair some problems with strong reflections of the signal from the floor were observed. This situation can happen when a floor surface is made of terracotta tiles. In such situation time delay corresponding to the reflection of the source is longer than time delay of direct signal. To overcome this, in situation when two highest peaks of the cross-correlation function differ only on less than ten per cent, the peak closer to zero lag is chosen as the true lag.

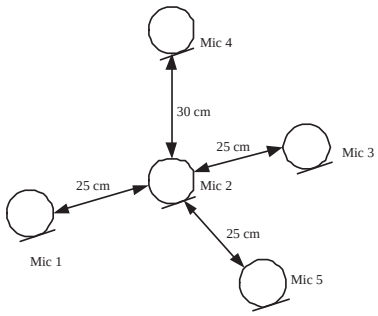

Fig. 6. The microphone array for sound source localization system

6. Conclusions

Passive sound source localization system for speech signals was developed and tested. Combination of time delay estimation algorithm with voice activity detector based on modulation properties of human speech gave the significant improvement in performance allowing to select precisely frames with speech but also to eliminate frequency bands without speech components and more accurately compute time delay. These future make possible to build reliable three dimensional speaker localization system using a small microphone array.

References

[1] Baszun J., Voice Activity Detection for Speaker Verification Systems. Joint Rougth Set Symposium, Toronto, Canada, (14-16 May, 2007), 181–186.

[2] Benesty J., Chen J., Huang Y., Microphone Array Signal Processing, Springer Topics in Signal Processing Series, Vol. 1, Springer-Verlag, 2010.

[3] Brutti A.B., Omologo M., Svaizer P., Comparision Between Different Sound Localization Techniques Based on a Real Data Collection IEEE HSCMA, (2008), 69–72.

[4] Carter C.G., Nuttal A.H., Cable P.G., The Smoothed Coherence Transform, Proc. IEEE (Letter), Vol. 61, (Oct. 1973), 1497–1498.

[5] Dmochowski J.P., Benesty J., Affes S., A Generalized Steered Response Power Method for Computationally Viable Source Localization, Audio, Speech and Language Processing, IEEE Trans. on, Vol. 15, I. 8, (Nov. 2007), 2510–2526.

[6] Elhilali M., Chi T., Shamma S., A Spectro-temporal Modulation Index (STMI) for Assesment of Speech Intelligibility. Speech Communication, Vol. 41. (2003), 331–348.

[7] El-Maleh K., Kabal P., Comparision of Voice Activity Detection Algorithms for Wireless Personal Communications Systems. Proc. IEEE Canadian Conference Electrical and Computer Engineering, (May 1997), 470–473.
[8] Houtgast T., Steeneken H.J.M., A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria. J. Acoust. Soc. Am., Vol. 77, No. 3 (Mar. 1985), 1069–1077.
[9] Huang Y.A., Benesty J., (Eds.), Audio Signal Processing for Next Generation Multimedia Communication Systems, Kluwer Academic Publishers, 2004.
[10] Knapp C.H., Carter C., The Generalized Correlation Method for Estimation of Time Delay, IEEE Transaction on Acoustics, Speech, And Signal Processing, Vol. ASSP-24, No. 4 (Aug. 1976), 320–327.
[11] Mesgarani N., Shamma S., Slaney M., Speech Discrimination Based on Multiscale Spectro-Temporal Modulations. ICASSP, (2004), 601–604.
[12] Moghaddam P.P., Amindavar H., Kirlin R.L., A New Time-Delay Estimation in Multipath, IEEE Transaction on Signal Processing, Vol. 51, (May 2003), 1129–1142.
[13] Sovka P., Pollak P., The Study of Speech/Pause Detectors for Speech Enhancement Methods. Proc. of the 4th European Conference on Speech Communication and Technology, Madrid, Spain (Sep. 1994), 1575–1578.
[14] Wang H., Chu P., Voice Source Localization for Automatic Camera Pointing System in Videoconferencing Proc. IEEE ASSP Workshop Applications on Signal Processing Audio Acoustics, (Oct. 1997), 1497–1498.

PASYWNY SYSTEM LOKALIZACJI ŹRÓDEŁ DŹWIĘKU

Streszczenie Opracowano metodę lokalizacji akustycznych źródeł dźwięku zorientowaną na sygnału mowy. System zbudowano w oparciu o macierz pięciu mikrofonów. Obliczenia pozycji źródła w trzech wymiarach dokonano na podstawie estymacji różnicy czasu przybycia dla par mikrofonów. Zastosowany psychoakustycznie motywowany detektor mowy umożliwia ocenę aktywności mówcy w obecności zakłóceń. Dobrą efektywność uzyskano poprzez selekcję ramek z mową oraz zerowanie zakresów częstotliwości w których sygnał zakłócający maskuje sygnał mowy. Jego zaletą jest możliwość precyzyjnego obliczania czasu opóźnienia. Eksperymenty w warunkach rzeczywistych pokazują dobrą odporność zaproponowanego algorytmu na szum i pogłos.

Słowa kluczowe: transformacja fazy, lokalizacja źródła, macierze mikrofonów.

Artykuł zrealizowano w ramach pracy badawczej S/WI/4/08.