International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:8, No:2, 2014

Time Delay Estimation Using Signal Envelopes for Synchronisation of Recordings

Sergei Aleinik, Mikhail Stolbov

***Abstract*—**In this work, a method of time delay estimation for dual-channel acoustic signals (speech, music, etc.) recorded under reverberant conditions is investigated. Standard methods based on cross-correlation of the signals show poor results in cases involving strong reverberation, large distances between microphones and asynchronous recordings. Under similar conditions, a method based on cross-correlation of temporal envelopes of the signals delivers a delay estimation of acceptable quality. This method and its properties are described and investigated in detail, including its limits of applicability. The method's optimal parameter estimation and a comparison with other known methods of time delay estimation are also provided.

***Keywords*—**Cross-correlation, delay estimation, signal envelope, signal processing.

I. Introduction

METHODS for time delay estimation (TDE, TD estimation) between two signals are widely used in acoustic data processing, etc. [1]-[5]. Most of the time these methods are based on some evaluation of the "similarity" of the signals to each other: on cross-correlation (CC), generalized cross-correlation, Euclidean distance, etc. Reviews of different TDE methods can be found, for example, in [6]-[9]. If the signals were recorded in a closed room, one of the main factors that decrease TDE accuracy is reverberation [6]. The reverberation problem becomes even more difficult when there is a large distance between the main and reference microphones. The worst case is with "asynchronous" recordings, when the main signal is recorded by microphone, but the reference signal is obtained from an external storage, for example, is read directly from a computer hard drive or CD, in different conditions and at different times [10]. In either case, coherence between the signals degrades significantly which leads to low (or even unacceptably low) TDE quality.

For example, Fig. 1 shows three cross-correlation functions (CCF) of speech signals recorded in a room when the distance between the main and reference microphones was 100, 200 and 300 cm. Reverberation leads to a notable decrease in CCF maximums, i.e. the correlation between the signals drops significantly. This paper focuses on a TDE method based not on the CC of the signals themselves, but on the CC of their temporal envelopes (hereafter simply "envelopes").

TDE methods using signal envelopes are reported in the literature, but these are in relation to either short radar signals or ultrasonic pulses [7]-[9], or envelopes of CC functions [6], [11], which, in both instances, are trivial.

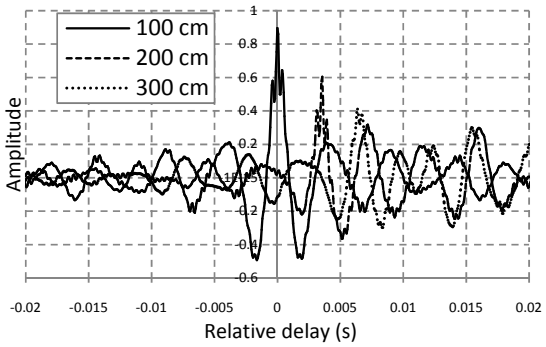

Fig. 1 Cross-correlation functions of speech signals for different distances between main and reference microphones

The method is also briefly mentioned in [12], but does not involve a detailed study.

Our aim was to undertake a detailed study of the method and to evaluate its optimal parameters.

This paper is organized as follows. Section II describes the basics of the method. Section III presents a description of the parameters. Section IV describes the experimental results. Comparison with the others TDE algorithms, Discussion and Conclusions are provided in Sections V, VI, VII and VIII, respectively.

II. Principles of Method

A. Asynchronous Recordings

The presented TDE method shows good performance in the "asynchronous case," i.e. when asynchronous recordings are processed [10]. Assume the standard situation: in a room, a human speech signal is recorded using a single-microphone mono recorder. At the same time, the same microphone receives a music signal played, e.g., on a TV or CD player. So we have a mixture: "speech + interfering music" as the main signal. In this case, the music is noise, which can significantly reduce speech intelligibility. The well-known method to increase the intelligibility of speech in such mixtures is to use a second recording channel with another microphone placed

S. V. Aleinik is with the Scientific department in Speech Technology Center, 4 Krasutskogo street, St. Petersburg, 196084, Russia (phone: 8-921-5830168; e-mail: aleinik@speechpro.com).

M. B. Stolbov is with the Department of Speech Information Systems in National Research University of Information Technologies, Mechanics and Optics, St. Petersburg, Russia (e-mail: stolbov@mail.ifmo.ru).

This work was partially financially supported by Government of Russian Federation, Grant 074-U01.

near the music source. The signal from this channel (recorded at the same time as the main signal and often called "reference" signal) can later be subtracted from the main signal using some well-known adaptive algorithm.

But what can be done if the main signal is already recorded and we have no reference signal?

The problem can be solved if we know which music we have to subtract. In this case "there is a possibility to obtain "artificial" or "asynchronous" reference signal, e.g. from the relevant music CD" [11]. It is clear that for correct adaptive subtraction, it is necessary to provide signals alignment; we therefore have to provide correct TDE. In the described case, however, the coherence between the asynchronous reference signal and the music in the main signal is low because of its different recording condition and equipment, etc. The delay between the main signal and the reference signal during the initial alignment task can be extremely high: several seconds or more. Furthermore, in our experiments [10] we frequently encountered a slight difference in the sampling frequency of the main signal and the reference signal. The result was "delay drift" – delay increases as function of time. Thus, standard TDE methods, based on signal coherence (CCF, Generalized CC-Phase Transform (GCC-PHAT), etc. [2]-[3]) perform badly. In our previous study [10], we achieved good results using CC of signal envelopes in asynchronous filtration of speech signals. These results motivated the present investigation.

B. Basics

In the method presented, TDE is provided (as it is in "standard" CC-method) as follows [2]:

$$\hat{\tau}_{CC} = \arg\max(R(\tau)), \quad (1)$$

where $\hat{\tau}_{CC}$ is estimated delay and $R(\tau)$ is the cross-correlation function. But in contrast to standard methods, for $R(\tau)$ estimation, signal envelopes are used, rather than the signals themselves. We therefore called the described method CC-ENV.

CC-ENV method can be carried out as follows:

- Calculate envelopes of the main and the reference signals.
- Calculate cross-correlation of the envelopes.
- Calculate time delay according to (1).

The core of the method is envelope calculation. We do not use the Hilbert transform [9] because it is optimal for short pulses and requires considerable computing resources on long signals. Instead of the Hilbert transform, a known procedure called "rectification and lowpass filtering" [13] (with some modification) was used.

C. Envelope Calculation

Temporal envelope $En(i)$ of discrete-time signal $x(i)$ is calculated as follows:

$$En(i) = HPF(LPF(|x(i)|)), \quad (2)$$

where: *HPF* is highpass filter; *LPF* is lowpass filter and '$|\ |$' denotes the absolute value. So, after rectification and lowpass filtering, highpass filtering was carried out. The physical effect of lowpass filtering is smoothing. After smoothing of the rectified speech signal, its envelope becomes a slowly oscillating process with a high level of constant and low-frequency components. To suppress these unwanted components, highpass filtering is used.

III. Algorithm Parameters

A. Lowpass Filter

As a lowpass filter, a first-order filter was used [14]:

$$y(i) = \beta(x(i) + x(i-1)) + \alpha y(i-1) \ , \quad (3)$$

where i is time index; $0 \le \alpha < 1$ and $\beta = (1-\alpha)/2$ are filter coefficients; $x(i)$ and $y(i)$ are input and output signals at the time i, respectively. Coefficient α is calculated using the well-known simplified formula:

$$\alpha = 1 - \frac{2}{1 + TF_s}, \quad (4)$$

where F_s is sampling frequency in Hz and T is equivalent window length in seconds. It is easy to see that T must match the speech rate. Indeed, if T is small, the envelope fluctuates greatly. On the other hand, abnormally high T leads to excessive envelope smoothing. In both cases, the correlation of envelopes decreases and, correspondingly, TDE quality degrades. Therefore, it is possible to assume that there is an optimal T_{lp}^{opt}.

B. Highpass Filter

A first-order highpass filter was used [14]:

$$y(i) = \beta(x(i) - x(i-1)) + \alpha y(i-1) \ , \quad (5)$$

where α is also calculated according to (4) (for another T, of course), but $\beta = (1+\alpha)/2$. Highpass filtering causes a decrease in envelope correlation (negative influence) (see Fig. 2). On the other hand, it narrows the main lobe of the envelopes' CCF (positive influence). It is therefore possible to expect that there is also an optimal T_{hp}^{opt}.

As an example, Fig. 2 shows 1.8 s segments of human speech signal, it's temporal envelope and filtered (with highpass filter) temporal envelope for optimal T_{lp} and T_{hp} (optimal parameters will be discussed and evaluated below).

C. Cross-Correlation Calculation

The main parameter in CCF calculation is the length of the segment used: T_{cc}. It is clear that T_{cc} must be consistent with

the envelopes' oscillations. For example, if T_{cc} is small (0.1 seconds or less); the time envelope of the speech signal on such a short segment can be a uniformly increasing (or decreasing) time series. In this case, it is impossible to estimate CCF correctly and, accordingly, TDE is impossible. Our experiments [10] showed that relatively adequate TDE in the case of CC-ENV method can be obtained using time segments of 1-2 seconds or longer, at a large computational cost. Decimation (possible because envelopes are slowly varying functions) increases the discreteness of TDE. In this paper, time-domain calculations with "steps" are used. This significantly accelerates CCF calculation with almost no loss of accuracy [10]. Let i and j are time indexes; $x_1(i)$ and $x_2(i)$ are discrete-time signals; N is the number of signals' samples; m is the time shift and $j = i - m$. The CCF $R(m)$ of x_1 and x_2 signals can therefore be obtained as:

$$R(m) = \frac{\sum_i x_1(i)x_2(j) - \frac{1}{M}\left(\sum_i x_1(i)\sum_j x_2(j)\right)}{\sqrt{d}}, \qquad (6)$$

where

$$d = \left(\sum_i x_1^2(i) - \frac{\left(\sum_i x_1(i)\right)^2}{M}\right)\left(\sum_j x_2^2(j) - \frac{\left(\sum_j x_2(j)\right)^2}{M}\right) \qquad (7)$$

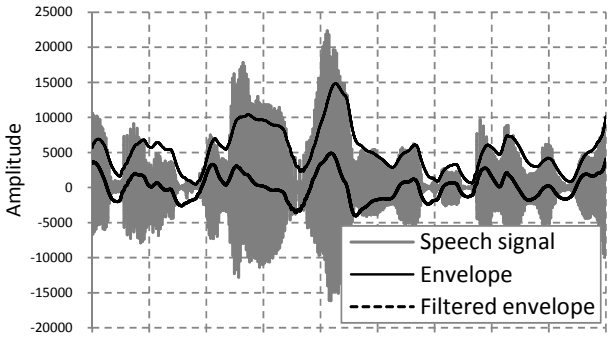

Fig. 2 Speech signal, its envelope and filtered envelope examples

We point out here that in (6) and (7):

- Sums are calculated in a single loop.
- $m = 0, \pm 1, \pm 2...$
- The loop variable i is changed as follows: $i = i + step$, until $i < N$, where $step \geq 1$.
- $M = \left\lfloor (N - m)/step \right\rfloor$

where $'\lfloor\ \rfloor'$ denotes the integer part. One can see that when $step = 1$, (6) and (7) are well-known equations for standard time-domain calculations of CCF in a single loop. But since the speech signal envelope is a slowly oscillating process, it is possible to use $step$ which is significantly more than 1. In our experiments, $step = 10, 100, 200$ were tested without a notable loss in CCF quality, but with a large increase in rate. We note here that in all the simulations described below $step = 100$.

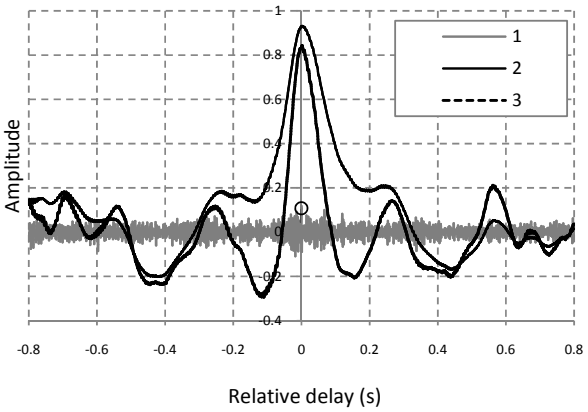

Fig. 3 Cross-correlation functions of speech signals (curve 1), their envelopes and filtered envelopes (curves 2 and 3, respectively)

Our experiments [10] showed that in certain conditions, signal envelopes are more robust than the signals themselves. As an example, three CCFs of the main and the reference signals and their envelopes, calculated using (6) with a segment length of 2 seconds and $step = 1$ are depicted in Fig. 3. The main and the reference signals are speech signals, recorded in a room with a reverberation time equal to 650 ms and a distance of 400 cm between the emitting speakerphone and the microphone; the main signal was the microphone recorded signal and the reference signals were read directly from computer memory (i.e. there were "asynchronous" recordings). In this case the main signal was dramatically corrupted by reverberation and distortion in the playback chain. In order to get any significant results, we therefore had to calculate CCFs using long signal segments: 1, 2 and more seconds (despite the fact that theoretical TD for 400 cm is about 0.01166 s, i.e. 186 samples for 16 kHz). The grey curve in Fig. 3 is a CCF estimated using signals and the black solid and black dashed curves are CCFs estimated using signals envelopes and filtered (with highpass filter) signal envelopes, respectively. There is almost no correlation when the CCF is estimated using signals (the grey curve, maximum is near $\tau = 0$, marked with circle). At the same time, the correlation is significant when signal envelopes are used (both signal envelopes and filtered envelopes).

IV. Experiments and Results

A number of experiments were conducted with different signals and with both model and real transforms.

A. Signals and Envelopes – Model Transforms

It is clear that the following model transforms are somewhat artificial, but they help demonstrate the benefits of the CC-ENV method. Let $x_1(i)$ and $x_2(i)$ be the discrete-time speech signals. Let $R_s(m)$ and $R_e(m)$ be CC functions calculated using the signals and their envelopes, respectively. It is clear that if $x_1(i) = x_2(i+k)$, then $\overline{R_s(k)} = \overline{R_e(k)} = 1$ (where $\overline{R}$ denotes estimated R). At the same time, envelopes are an advantage in cases of simple non-linear transform, e.g. $x_2(i) => |x_2(i)|$, or $x_2(i) => (x_2(i))^2$, where '$|$ $|$' denotes the absolute value. In such cases, R_s decreases sharply, while R_e does not change (when calculating the absolute value), or varies only slightly.

It is also trivial that random ($\pm\pi$) phase change, for example:

$$x^{rnd}(i) = \begin{cases} -x(i) & \text{if } \xi > \mathrm{Tr} \\ x(i) & \text{if } \xi \le \mathrm{Tr} \end{cases}, \tag{8}$$

(where ξ is uniform distributed random value $U(0,1)$ and $\mathrm{Tr} \in [0,1]$ is threshold level)does not change signal envelope and therefore does not affect the R_e. At the same time, $\mathrm{Tr} = 0.5$ leads to completely uncorrelated $x^{rnd}(i)$ and $x(i)$.

B. Signals and Envelopes – Real Transform: Speech+Noise

Let $x_1(i)$ and $x_2(i)$ be calculated as follows:

$$\begin{aligned} x_1(i) &= (1-\mu)s(i) + \mu n_1(i) \\ x_2(i) &= (1-\mu)s(i) + \mu n_2(i) \end{aligned}, \tag{9}$$

where $s(i)$ is the speech signal; $n_1(i)$ and $n_2(i)$ are independent random values and $0 \le \mu \le 1$ is a coefficient. It is clear that when $\mu = 0$,then $x_1(i) = x_2(i) = s(i)$ and $\overline{R_s(0)} = \overline{R_e(0)} = 1$. On the other hand, if $\mu = 1$, $x_1(i)$ and $x_2(i)$ are independent random values, so $\overline{R_s(0)}$ and $\overline{R_e(0)}$ fluctuate near 0. If the powers of $s(i)$, $n_1(i)$ and $n_2(i)$ are equal, it is easy to obtain a theoretical expression for $R_s(0)$ as a function of μ:

$$R_s^t(0,\mu) = \frac{(1-\mu)^2}{\mu^2 + (1-\mu)^2}, \tag{10}$$

Fig. 4 shows the mean values of $\overline{R_s(0)}$, $\overline{R_e(0)}$, their 95% confidence intervals, and $R_s^t(0,\mu)$, calculated using signals (8) as a function of μ. As speech signals, we used different phrases taken from the TIMIT database (16 kHz, WAV PCM 16-bits, mono). As noise, we used non-overlapping noise segments (for purposes of independence) recorded in the cabin of a Buccaneer aircraft (the buccaneer2.wav file from the well known NOISEX-92 database). The noise and speech powers (as well as sampling frequencies) were converted to the same values before calculating the transform (9). Parameters of envelope calculation: $T_{cc} = 2$ s, $T_{lp} = 0.05$ s, $T_{hp} = 0$ (means no highpass filter). The number of trials to estimate mean values is $L = 1024$.

Fig. 4 demonstrates that when μ increases, $\overline{R_s(0)}$ decreases and is almost identical to the theoretical curve calculated by (10). At the same time, $\overline{R_e(0)}$ maintains a high value long enough and only after $\mu = 0.6$ drops sharply. In our opinion, this effect is caused by the fact that a small addition of noise to the speech signal is simply an increase in constant component of the speech signal envelope, which is affected very little in R_e.

C. Signals and Envelopes – Real Transform: Filtered Speech

Let $x_1(i)$ and $x_2(i)$ be obtained as

$$\begin{aligned} x_1(i) &= s(i) \\ x_2(i) &= Lowpass\ \ filter(x_1(i)) \end{aligned}, \tag{11}$$

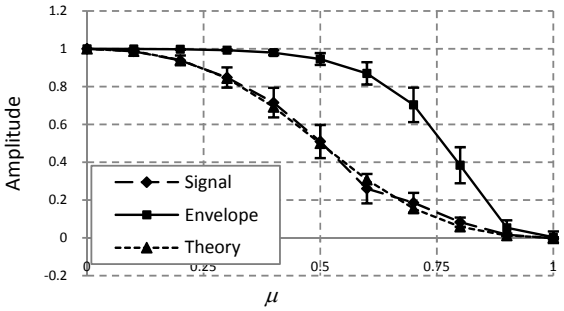

Fig. 4 $\overline{R_s(0)}$, $\overline{R_e(0)}$ and $\overline{R_s^t(0)}$ as functions of μ

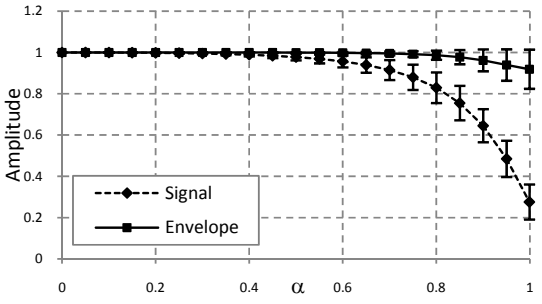

Fig. 5 $\max_k(\overline{R_s(k)})$ and $\max_k(\overline{R_e(k)})$ as functions of α

Let the lowpass filter be a simple one-pole filter:

$$y(i) = \alpha y(i-1) + (1-\alpha)x(i), \quad (12)$$

where $0 \le \alpha < 1$ is the filter parameter, and $x(i)$ and $y(i)$ are input and output signals, respectively. It is known that the closer α is to 1, the closer the cutoff frequency is to 0.Fig. 5 shows the mean values of $\max_k(\overline{R_s(k)})$ and $\max_k(\overline{R_e(k)})$, as well as their 95%confidence intervals, calculated using signals (11) as a function of α. The operation of finding the maximum is applied because the filter (12) causes a delay in the signal, which depends on α. Accordingly, the CCF maximums for both the signals and their envelopes will not be in the zero sample in this case. Speech signal were also taken from TIMIT database and the parameters of calculating envelopes are the same as in the previous ("*speech+noise*") section. It can be seen that when $\alpha > 0.6$, $\max_k(\overline{R_s(k)})$ drops significantly faster than $\max_k(\overline{R_e(k)})$.

D.Optimal Parameters for Envelope Calculation

Coordinate-wise random search optimization (CRSO) was used to find optimal parameters T_{lp}^{opt} and T_{hp}^{opt} for different T_{cc}. We used WAV PCM 16 kHz, 16 bits, dual-channel data: "human song," "human speech," "music," "pink noise" and "modulated pink noise," obtained in an experiment fully described in [10]. The data was recorded in a 6m × 5m × 3m room with 650 ms reverberation time. The distance between the main and reference microphones was 400cm. While space limitations prevent us from presenting all our results obtained, we shall describe in detail the results corresponding to the "human song" signal (optimal parameters search task) and to the "human speech" signal (comparison with the other TDE methods task). (Results for the remaining signals will be briefly described in Section VI). In our opinion, the song data were the most revealing for the search task, as they included both speech and music. We choose data in which the distance between the main and reference microphone was 400 cm. The CRSO was conducted as follows. Initial settings were: $T_{lp}^{opt} = T_{hp}^{opt} = 0$. Random search was then alternately optimized: (a) only T_{lp} with fixed T_{hp}; (b) only T_{hp} with just fond fixed T_{lp}; (c) both T_{lp} and T_{hp}; then again (a), etc. Optimized parameter(s) switching was performed by achieving $K_0 = 128$ steps without improvement. The algorithm terminated if $K_1 = 512$ steps was reached without improvement. As a target value to be minimized, mean squared error (MSE) of TD was selected:

$$\mathrm{MSE} = \frac{1}{L}\sum_{i=0}^{L-1}(\tau(i) - \tau_{teor})^2,$$

where L is number of trials undertaken to estimate MSE; $\tau(i)$ is TD value, estimated in i-th trial; τ_{teor} is calculated theoretical TD value. The number of trials at each single optimization step was $L = 2048$. T_{cc} was varied from 1.5 to 8 s.

Fig. 6 demonstrates the values of MSE obtained at the first stage (T_{lp} search with fixed T_{hp}) of CRSO for $T_{cc} = 4$ s.

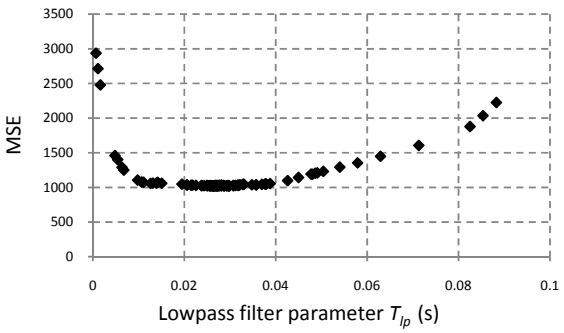

Fig. 6 MSE as a function of T_{lp} for $T_{cc} = 4$ s

It can therefore be concluded that our hypothesis regarding the existence of an optimal T_{lp} is borne out by the experiments. Also, we note here that the data in Fig. 6 (as well as in Figs. 7 and 8) are point estimates; it is therefore impossible to calculate confidence intervals.

An example of parameters and MSE change during the CRSO procedure (only iterations which lead to MSE decrease), for $T_{cc} = 7$ s is presented in Fig. 7.

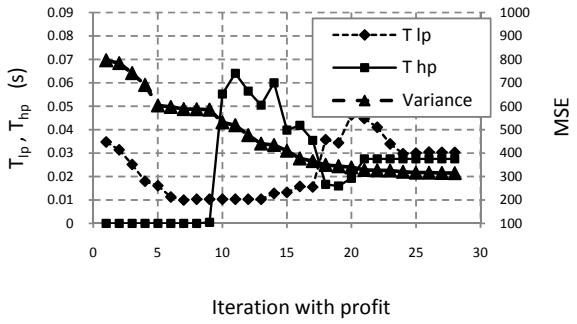

Fig. 7 T_{lp}, T_{hp} and Var_{td} evolution during CRSO

The first stage reaches a plateau (MSE and T_{lp} are almost constant), about 8-9 positive iteration, and then in the second stage (T_{hp} search with fixed T_{lp}), MSE drops again, etc., until all three curves become constant and the algorithm stops.

Fig. 8 shows values of MSE for estimated optimal T_{lp} and T_{hp} as a function of T_{cc} at the end of the CRSO algorithm. It is seen that MSE decreases when T_{cc} increases. Note that using lowpass and highpass filters together gives better results than using only a lowpass filter.

We would like to call attention to one phenomenon which

may explain the behaviour of the curves. When T_{cc} is equal to several seconds or more, the speech signals include not only sounds, but entire words. In this case, the signal has enough pauses between sounds.

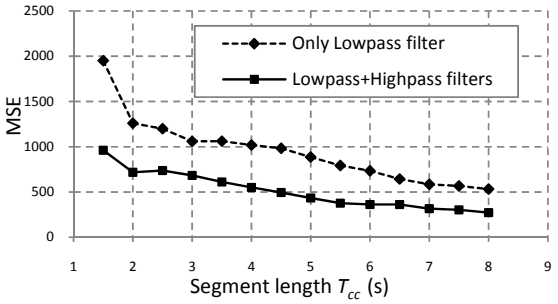

Fig. 8 MSE when T_{lp} and T_{hp} are optimal

In reality, there is only an interfering noise in these pauses, which affects the quality of TDE. Using envelopes, we first suppressed this noise by filtration and, second, these pauses became not parasitic, but actual "working" parts of the envelopes. Roughly speaking, "sound-pause-sound" and "pause-sound-pause" in the signal transforms to "∪" and "∩" - shaped curves in its envelope. The longer the T_{cc}, the more curves of this type in the envelopes; and correspondingly, the higher the accuracy of TDE. Optimal parameters estimated during random search, along with their mean values, are presented in Table I.

TABLE I
OPTIMAL T_{lp} AND T_{hp} FOR "HUMAN SONG DATA"

T_{cc} (s)	Only lowpass filter	Lowpass+highpass filters	
	T_{lp}^{opt} (s)	T_{lp}^{opt} (s)	T_{hp}^{opt} (s)
2.0	0.0212	0.0396	0.0319
3.0	0.0219	0.0311	0.0441
4.0	0.0241	0.0313	0.0394
5.0	0.0107	0.0275	0.0332
6.0	0.0119	0.0315	0.0327
7.0	0.0102	0.0303	0.0275
8.0	0.0137	0.0225	0.0374
Mean	0.0164	0.0321	0.0340

V. COMPARISON WITH CCF AND GCC-PHAT METHODS

We compared the CC-ENV method, presented here, with standard CCF and GCC-PHAT methods. Our experiments show that if signal distortion is low or medium (and coherence is high), best results are obtained using GCC-PHAT and CC-ENV algorithms provide the poorest results. Furthermore, if the parameter T_{cc} is small (less than 0.5 s), CC-ENV performs poorly. But if signals are strongly corrupted, especially in nonlinear transform or asynchronous cases, and if T_{cc} is sufficiently long, the method presented here does offer an advantage. A comparison of the aforementioned TDE methods is provided in Figs. 9 and 10, depicting cases when MSE of estimated delays area function of T_{cc}. Fig. 9depicts the "no reverberation" case. Here, both the reference and the main signals are speech signals (16 bit, 16 kHz, mono PCM).The reference signal is clear, with a signal-to-noise ratio (SNR) of 30 dB, and the main signal is the same as the reference signal, but corrupted with a slightly non-stationary noise (factory1.wav file from the NOISEX-92 database), with SNR of 6 dB.We carried out 1000trials for MSE estimation. When T_{cc} is small, MSE is sufficiently high for all three methods. However, MSE drop significantly as T_{cc} increases. In our experiments, zero MSE for GCC-PHAT was achieved when T_{cc} is equal or more 1.75 s. It is also notable that GCC-PHAT provides the best result (i.e. the lowest MSE) and CC-ENV, the worst.

The results presented in Fig. 10 are entirely different. In this experiment, we used the same clear speech reference signal as above, but used the "asynchronous" signal for the main signal, as described in Section I, corrupted with NOISEX-92 factory1.wav noise with SNR of 6 dB. In this case, the main signal is corrupted by distortion in the playback chain, reverberation and additive noise. It is notable, that CC-ENV provides better results than both CCF and GCC-PHAT methods. And optimal results are achieved when T_{cc} is equal to 2s.We should point out that the slight increase in TD MSE obtained using the CC-ENV method for $T_{cc} > 2$ sec remains unexplained, especially if compared with the curves depicted in Fig. 8.A possible cause is the difference in processed signals("song" and "noised speech") or the fact that not optimal $T_{lp} = 0{,}02$ was used.

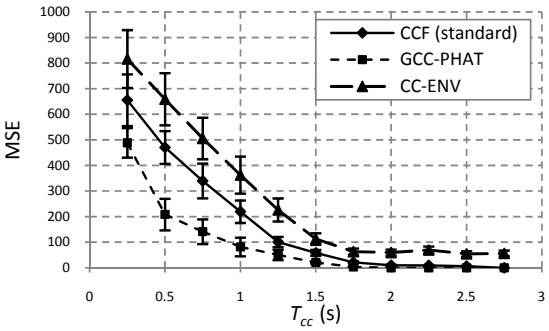

Fig. 9 MSE as a function of T_{cc} for different TDE methods. No reverberation. T_{lp} for CC-ENV is 0.02, T_{hp} is 0

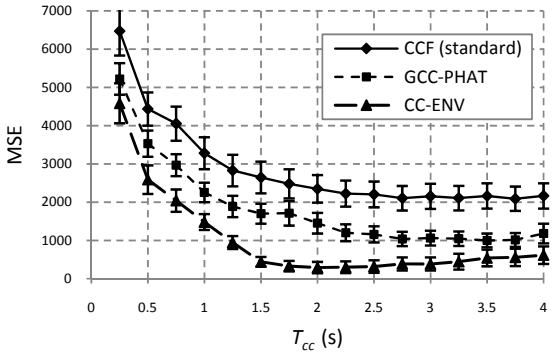

Fig. 10 MSE as a function of T_{cc} for different TDE methods. Asynchronous case T_{lp} for CC-ENV is 0.02, T_{hp} is 0

VI. Comparison with Speech-Envelope Spectrum

Sequentially connected high- and lowpass filters (with overlapping bandwidths) represent a bandpass filter. The frequency response (FR) of this filter is equal to the product of corresponding FRs of its components [14]. The FR of filters (3) and (4) for known α is known (p. 11.2 [14]).Normalized magnitude FRof this bandpass filter (actually only its low-frequency part), calculated for mean T_{lp}^{opt} and T_{hp}^{opt} from Table I, is depicted in Fig. 11.

It is interesting to compare this FR to the Speech-Envelope spectrum (SES), presented in [15].Although SES has a maximum of about 4-5 Hz and falls more sharply when the frequency is increasing, it can be said that the curves correspond to each other.

VII. Discussion

Closer inspection of the simulation results suggests that using temporal envelopes in TDE has a positive effect when some signal transform greatly distorts the signal itself, leaving its envelope unchanged (or distorts it to a lesser degree).

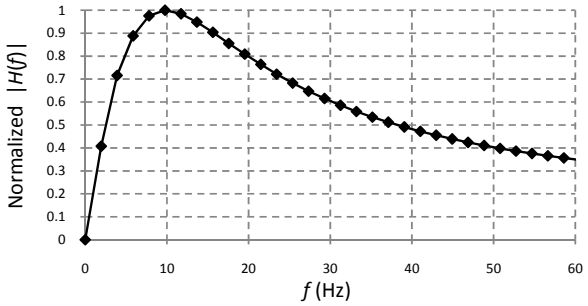

Fig. 11 Low-frequency part of magnitude response for mean optimal bandpass envelope filter (normalized)

For example, the CC-ENV method is highly useful in asynchronous recordings.

Envelopes for TDE are ineffective in cases of weak signal distortion or when signals have envelopes with strong constant periodicity. For example, we obtained poor results on rhythmic music and on harmonically modulated pink noise.

In contrast, it is known that time envelopes of typical human speech or speech+music signals have strong non-periodic fluctuation, so (as confirmed by our experiments); it makes them suitable for TDE.

In our view, the main problem of the proposed method is the dependence of the T_{lp} and T_{hp} parameters on the type of signals and their distortion levels. It is understandable that data presented in Table I correspond to only one type of signal and distortion, etc. However, it can be assumed that the mean values at the bottom of the table might suggest a first approximation of actual values.

VIII. Conclusion

The paper has presented a method for time delay estimation (TDE) of audio signals using a cross-correlation function of their temporal envelopes. The major advantage of this method is that it achieves better performance when signals are strongly corrupted by different transforms, for example, phase randomization, reverberation, etc. On the other hand, the method requires an analysis of long signals, which leads to a large computational cost. Optimal parameters for the method have been evaluated and a comparison with other TDE methods was provided in a simulation using speech and music signals recorded under real conditions.

References

[1] J. Chen, J. Benesty and Y.A. Huang, "Time Delay Estimation in Room Acoustic Environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, 2006, pp. 1-20.

[2] A. Sandmair, M. Lietz, J. Stefan, and F.P. Leon, "Time delay estimation in the time-frequency domain based on a line detection approach", in *Proc. ICASSP- International Conference on Acoustics, Speech, and Signal Processing*, 2011, pp. 2716-2719.

[3] K. Gedalyahu and Y.C. Eldar. "Time-delay estimation from low-rate samples: A union of subspaces approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, 2010, pp. 3017 –3031.

[4] B. Kirkwood, "Acoustic Source Localization Using Time-Delay Estimation", M.S. Thesis, 2003, http://brentkirkwood.com/science/project-ms.html

[5] A. Kozlov, O. Kudashev, Yu. Matveev, T. Pekhovsky, K. Simonchik, A. Shulipa."SVID Speaker Recognition System for NIST SRE 2012,"in. *Proc. of 15th International Conference "Speech and Computer" (SPECOM 2013)*. Springer Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence, 2013, Vol. 8113, pp. 278-285.

[6] S. Bédard, B. Champagne and A. Stéphenne, "Effects of Room Reverberation on Time-Delay Estimation Performance," *IEEE Transactions Acoustics, Speech, and Signal Processing*, vol.2, 1994, pp. 261-264.

[7] R. Raya, A. Frizera, R. Ceres, L. Calderón, E. Rocon, "Design and evaluation of a fast model-based algorithm for ultrasonic range measurements," *Sensors and Actuators A: Physical*, vol. 148, No. 1, 2008, pp. 335–341.

[8] L. Yang, A.V. Lavrinenko, J.M. Hvam, and O. Sigmund, "Design of one-dimensional optical pulse-shaping filters by time-domain topology optimization," *Appl. Phys. Lett*. 95, 2009, 261101.

[9] B.S. Lazarov, R. Matzen, and Y. Elesin, "Topology optimization of pulse shaping filters using the Hilbert transform envelope extraction," *Structural and Multidisciplinary Optimization*, vol. 44, no. 3, pp. 409–419, 2011.

[10] P. Ignatov, M. Stolbov, S. Aleinik, "Semi-automated technique for noisy recording enhancement using an independent reference recording," in

Proc. 46th International Conference of the Audio Engineering Society, 2012, pp. 57-65
[11] N. Thrane, J. Wismer, H. Konstantin-Hansen, and S. Gade, "Practical use of the Hilbert transform," Application Note, Brüel&Kjær, Denmark. Available: http://www.bksv.com/doc/bo0437.pdf
[12] C. Faller, C Tournery, "Estimating the delay and coloration effect of the acoustic echo path for low-complexity echo suppression," in*Proc. Intl. Works. OnAcoust. Echo and Noise Control (IWAENC)*, The Netherlands, 2005, pp. 53-56.
[13] O.M.Bouzid, G. Y. Tian, J.Neasham, and B. Sharif, "Envelope and Wavelet Transform for Sound Localisation at Low Sampling Rates in Wireless Sensor Networks," *Journal of Sensors*, vol. 2012, Article ID 680383, 9 pages.
[14] S. J. Orfanidis, *Introduction to Signal Processing*. Available: http://www.ece.rutgers.edu/~orfanidi/intro2sp/orfanidis-i2sp.pdf
[15] T. Hougast, H. J. M. Steeneken, "A review of the MTF concept in room acoustics and it's use for estimating speech intelligibility in auditoria", *J. Acoust. Soc. Am*. 67, 1985, pp. 1060-1077.