Dynamics and Periodicity Based Multirate Fast Transient-Sound Detection

Jun Yang (IEEE Senior Member) and Philip Hilmes

Amazon Lab126, 1100 Enterprise Way, Sunnyvale, CA 94089, USA

***Abstract*—This paper proposes an efficient real-time multirate fast transient-sound detection algorithm on the basis of emerging microphone array configuration intended for multimedia signal processing application systems such as digital smart home. The proposed detection algorithm first extracts the dynamics and periodicity features, then trains the model parameters of these features on Amazon machine learning platform. The real-time testing results have shown that the proposed algorithm with the trained model parameters can not only achieve the optimum detection performance in all various noisy conditions but also reject all kinds of interferences including undesired voice and other unrelated transient-sounds. In comparison with the existing algorithms, the proposed detection algorithm significantly improves the false negative and false positive performance. In addition, the proposed multirate strategy dramatically reduces the computational complexity and processing latency so that the proposed algorithm can serve as a much more practical solution for the digital smart home related applications.**

Keywords—feature extraction; fast transient-sound detection; sound source localization; digital-positioning system; smart home

I. INTRODUCTION

A home-based digital device combines projection, vision, and audio technologies to redefine the home digital experience by connecting to the home network and providing access to information, entertainment, and communication.

Audio gestures produce the sounds, or fast transient-sounds by hand clap, finger snap, tap using knuckle tips at wall or table, tap flats at wall or table, finger flats at wall or table, palm to wall or table, fingertip with nails at wall or table, fingertip without nails at wall or table, etc.. The home-based digital device is equipped with multiple microphones (say, 4 or 8 microphones). By using the corresponding audio gestures which are the user's events of interest, the user tells the digital device where the user wants to display. For the above events of interest where the fast transient-sounds or impulse signals are produced, the digital device needs to detect, locate these events, and position the display towards the desired location of the events, so as to present a menu of options on user's hand. Therefore, accurate detection of the desired fast transient-sound is the key component in such a kind of digital devices.

In order to prevent the digital device from positioning the built-in projection misleading by single accident undesired audio gesture, the double-audio gesture is proposed for the control mechanism. In other words, the double-audio gesture is used to define the user's events of interest, such as double knuckle to wall or table, double palm to wall or table, double fingertip to wall or table, double hand clap, etc..

In practice, various environmental noises and interferences can greatly degrade the fast transient-sound detection (FTSD) accuracy and sound source localization performance. These noises and interferences include but not limited to fan noise, speakers' voice, laughing sounds, and the noise of the built-in motor incurred when the digital device moves or rotates. In addition, other impulse interference signals, such as door shutting and cup dropping on the floor can misguide the digital device to position the display.

Although some FTSD algorithms including double fast transient-sounds detection (DFTSD) have been proposed [1-3], these existing algorithms have significant drawbacks mainly because of the very expensive computational complexity, long latency, not robust to noise, voice, and laughing sounds. Moreover, as shown in Section 5, the design of these existing FTSD algorithms is independent from the training for the false positive and false negative performance, which in turn cannot maximize the detection performance and sound source localization performance. The above problems prevent these existing FTSD and DFTSD algorithms from practical use and being accepted by the users. It is the goal for this paper to propose a new FTSD algorithm that overcomes the above drawbacks so as to achieve the optimum processing performance. The proposed FTSD algorithm has been trained and verified by a large database on Amazon machine learning platform.

More specifically, the proposed multirate FTSD scheme firstly performs the detection processing by extracting the key features including dynamics and periodicity on the basis of the emerging microphone array configuration. Secondly, an effective and robust positioning system is established by making full use of the proposed FTSD algorithm and its generalization to DFTSD. The given theoretical analyses and objective test results show that the proposed system can offer a significant improvement for FTSD, DFTSD, sound source localization, and positioning performance in smart home devices.

It is worth mentioning that the artificial intelligent speakers which include the built-in screens can benefit from the proposed FTSD algorithm as well. The screens of these artificial intelligent devices can be automatically adjusted towards the location of the fast transient-sound events of interest.

The rest of this paper is organized into the following five sections. Section 2 mainly presents the proposed signal processing architecture of the digital-positioning system in smart home. Section 3 is devoted to the details of the proposed algorithm of robust multirate FTSD. In Section 4, a sound source positioning system is provided by employing the proposed FTSD and sound source localization. By conducting various testing, Section 5 mainly presents various test results to show that the smart home device implemented with the proposed FTSD algorithm can have significant improvements in terms of FTSD, DFTSD, sound source localization, and positioning performance. Section 6 will make some conclusions and further discussions.

II. THE ARCHIECTURE OF THE PROPOSED SYSTEM

Fig. 1 shows the signal processing architecture of a digital-positioning system in smart home by using the proposed scheme.

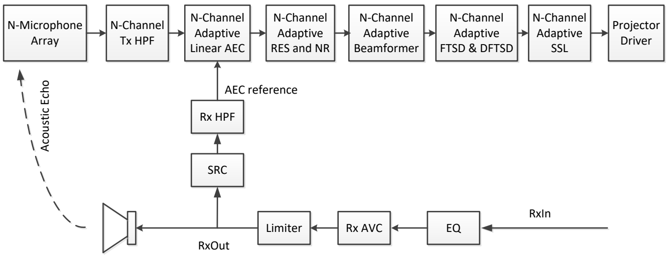


Fig. 1. Architecture of the Proposed Digital-Positioning System

In Fig. 1, HPF, AEC, RES, NR, SSL, SRC, AVC, and EQ denote for high-pass filter, acoustic echo cancellation, residual echo suppression, noise reduction, sound source localization, sampling rate convertor, automatic volume control, and speaker equalizer, respectively. The algorithms of HPF, AEC, RES, NR, and SRC have been described in [4]. The algorithms of Limiter and EQ have been described in [5].

The proposed system generates the *(x, y, z)* location of transient-sound events of interest, which means that the block "Projector Driver" in Fig. 1 can guide and position the display towards the location of the events.

III. THE DETAILS OF THE PROPOSED FTSD ALGORITHM

A. The Basis of the Proposed Detection Algorithm

The fast transient-sound is typically instantaneous sharp (i.e., large dynamics, short-duration) and non-periodic. Through extensive analyses of the characteristics of the fast transient-sound, a multirate detection scheme is proposed and shown in Fig. 2.

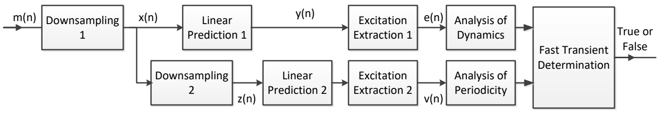


Fig. 2. The Processing Blcoks of the Proposed Detection Algorithm

This is a frame-by-frame time domain processing. For the sake of the fast processing, the frame length could be short, such as 16 ms. Since this is feature extraction-based approach, the processing is not sensitive to the signal level. The output of this detection processing is a decision of a value of either True or False.

In Fig. 2, the input signal *m(n)* of sampling rate fs_1 (ranging from 96 kHz to 32 kHz) is down-sampled to *x(n)* of sampling rate fs_2 (say, 16 kHz). The down-sampled signal *x(n)* is processed by the block "Linear Prediction 1". The linear prediction representation is described as

$$y(n) = \sum_{i=1}^{p} a_i x(n-i) \qquad (1)$$

where *y(n)* is the predicted signal, a_i are predictor coefficients. Levinson-Durbin recursion algorithm can be used to obtain the coefficients a_i. The parameter *p* could be any number around 10. The variable *e(n)* is a linear prediction error generated by the following equation

$$e(n) = x(n) - y(n) \qquad (2)$$

The block "Analysis of Dynamics" in Fig. 2 is a very important part of the proposed solution in this paper and is shown in more details by Fig. 3.

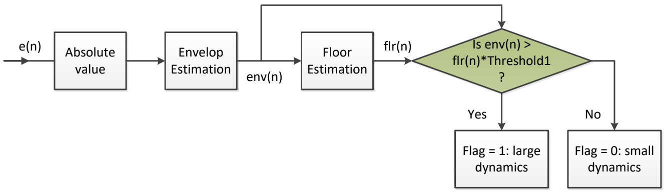


Fig. 3. The Proposed Dynamics Analyzer

In Fig. 3, the "Envelope Estimation" block is implemented as follows.

$$env(n) = env(n-1) + \beta\big(|e(n)| - env(n-1)\big) \qquad (3)$$

where β is a smoothing factor between 0.0 and 1.0. The "Floor Estimation" block is implemented as follows.

$$flr(n) = flr(n-1) + \gamma\big(env(n) - flr(n-1)\big) \qquad (4)$$

where γ is a smoothing factor between 0.0 and 1.0.

The block "Downsampling 2" converting from *x(n)* to *z(n)* in Fig. 2 could be of the 8:1 ratio so that the *z(n)* is of sampling rate $fs_3 = fs_2/8 = 2$ kHz. Therefore, the computational complexity can be greatly reduced. The block "Linear Prediction 2" can have small order *p* (say, $p = 4$).

The block "Analysis of Periodicity" in Fig. 2 is implemented by the autocorrelation approach. The autocorrelation function of a discrete-time signal is defined as follows.

$$R(k) = \sum_{m=-\infty}^{\infty} v(m)v(m+k) \qquad (5)$$

If digital signal *{v(m)}* is of period *P* with zero mean, then its autocorrelation is periodic as well, i.e., $R(k) = R(k+P)$. A short-time autocorrelation function of a sequence is more useful and can be defined as Eq. (6) shows, where the final subscript is understood to be taken modulo *M*.

$$\rho_i = \sum_{j=0}^{M-1} v(j)v(j+i) \qquad (6)$$

The output of the "Analysis of Periodicity" block would be True if *v(n)* is periodic, or False if *v(n)* is non-periodic. Please note that signals *m(n)*, *x(n)*, *z(n)*, and *v(n)* have the same periodicity.

The block "Fast Transient Determination" in Fig. 2 includes the following implementation steps:

(1). If *e(n)* is of large dynamics and *v(n)* is not periodic, then *m(n)* is determined as the transient sound in the current frame which means that this FTSD block outputs True.

(2). If *e(n)* is of low dynamics or *v(n)* is periodic, then *m(n)* is not the transient sound in the current frame, and this FTSD block outputs False.

In addition, the microphone array can provide us with spatial information of the fast transient-sound. Any of the N-channel signals can be selected as the *m(n)* signal in Fig. 2. If the computational complexity budget allows, each channel signal can be applied by the proposed FTSD processing, so that *N* decisions can be obtained. The final decision of FTSD can be voted, as shown in Fig. 4, on the basis of the obtained *N* decisions.

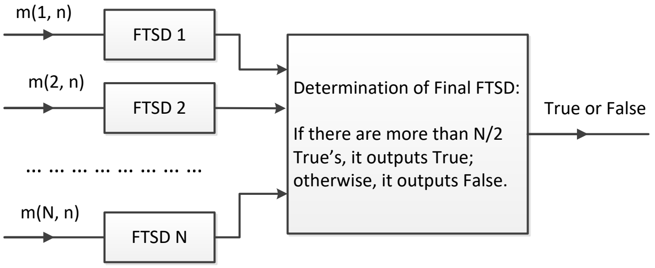

m(1, n)
FTSD 1
m(2, n)
FTSD 2
… … … … … … … …
m(N, n)
FTSD N
Determination of Final FTSD:
If there are more than N/2 True's, it outputs True; otherwise, it outputs False.
True or False

Fig. 4. The Final Determination of FTSD with Microphone Array

The number of microphones, *N*, should be no less than 2. There is no limitation for the configuration of the *N* microphones in the proposed FTSD algorithm. All the N-channel signals *m(1, n)* through *m(N, n)* have the same sampling rate fs_1 which can range from 96 kHz to 32 kHz. It should be noted that the smaller is the fs_1, the computational complexity is the less.

B. The Generalization to DFTSD Algorithm

Through extensive analyses of their spatial properties of N-channel signals, the above algorithm can be generalized to perform a DFTSD with adding the block "Spatial Analysis and DFTS Determination" as shown in Fig. 5.

In Fig. 5, the block "Spatial Analysis" is to calculate the time delay estimation (TDE) among the N-channel signals. To achieve a more accurate detection, the linear prediction error *e(i, n)* (*i = 1, …, N*) signals are chosen to be the inputs of block "Spatial Analysis". If Channel-1 signal is used as reference, there are *(N-1)* TDE values. The block "DFTS Determination" outputs True if all the *(N-1)* TDE values are between Threshold-1 and Threshold-2. Otherwise, it outputs False, because the desired double fast transient-sounds should happen closely in location and closely over the time.

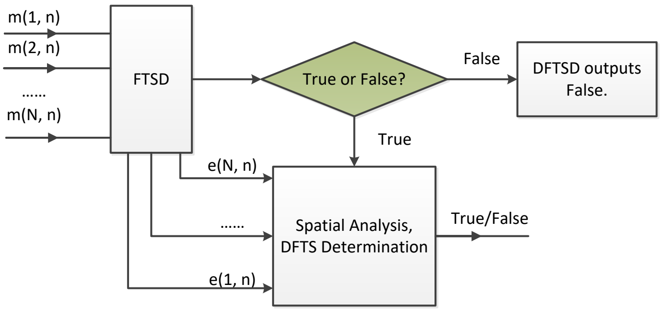


Fig. 5. The Proposed DFTSD with Microphone Array

The TDE calculation is based on the cross-correlation between Channel-1 signal and Channel-i signal (*i = 2, 3, ..., N*) or phase transform approaches. The peak in the cross-correlation indicates the TDE. This means that the time index corresponding to the maximum value of the cross-correlation sequence denotes for the Time Delay between Channel-1 and Channel-i (*i = 2, 3, ..., N*) signals. A more robust-to-noise approach is to use least-mean-square adaptive finite impulse response algorithm to adaptively estimate the time delay. The peak in the finite impulse response taps indicates the TDE.

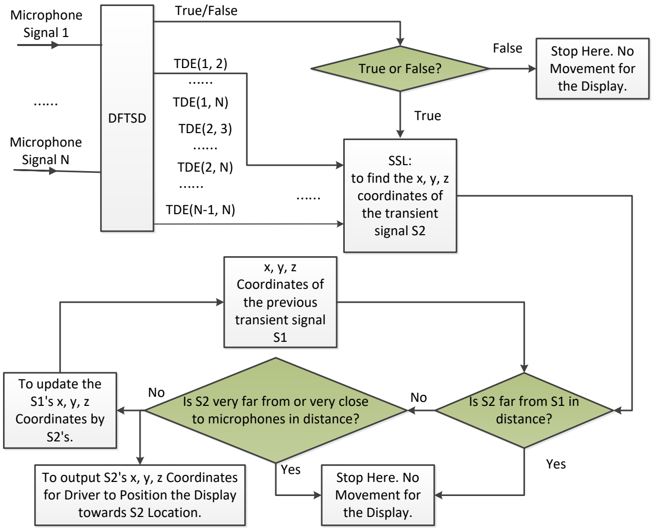
If the computational complexity budget is enough, each channel signal can be used as reference, so, there are *(N-1) + (N-2) + ... + 1 = N(N-1)/2* TDE values. The block "DFTS Determination" outputs True if all the *N(N-1)/2* TDE values are between Threshold-1 and Threshold-2. Otherwise, it outputs False.

It should be noted that the related feature parameters, such as *β*, *γ*, the related thresholds for dynamics and periodicity features, Threshold-1, and Threshold-2, are determined in an offline mode by tuning and training with a large database on a cloud-based Amazon machine learning platform.

IV. The Sound Source Positioning System with the Proposed DFTSD

By using the proposed DFTSD, a sound source positioning system is proposed and shown in Fig. 6.

In Fig. 6, the number of microphones N should be no less than 4 in order to find the x, y, and z coordinates of the desired transient-sound. The major processing steps in Fig. 6 include transient-sound detection, delay estimation, localization of the detected transient-sound, and determination of the desired transient-sound. By using the linear prediction excitation signals *e(i, n)* (1 ≤ i ≤ N) extracted by FTSDs as the inputs of the delay estimation and inputs of localization of the detected transient-sound, the positioning system is very robust to noise. The sound source localization can be performed by the spherical interpolation or Brandstein-Adcock-Silverman algorithms.



Microphone Signal 1
......
Microphone Signal N
DFTSD
True/False
TDE(1, 2)
......
TDE(1, N)
TDE(2, 3)
......
TDE(2, N)
......
TDE(N-1, N)
......
True or False?
False
Stop Here. No Movement for the Display.
True
SSL: to find the x, y, z coordinates of the transient signal S2
x, y, z Coordinates of the previous transient signal S1
To update the S1's x, y, z Coordinates by S2's.
No
Is S2 very far from or very close to microphones in distance?
No
Is S2 far from S1 in distance?
Yes
Yes
To output S2's x, y, z Coordinates for Driver to Position the Display towards S2 Location.
Stop Here. No Movement for the Display.

Fig. 6. The Proposed Sound Source Positioning System

V. Evaluations of the Proposed System

In this section, the evaluation results and test analyses of the proposed system in Fig. 6 (including the proposed DFTSD in Fig. 5 and FTSD in Fig. 4) will be presented in terms of the FTSD and DFTSD performance as well as their false positive and false negative values.

As mentioned in Section III, all the key factor parameters in Figs. 4, 5, and 6 are first obtained by a cloud-based Amazon machine learning platform.

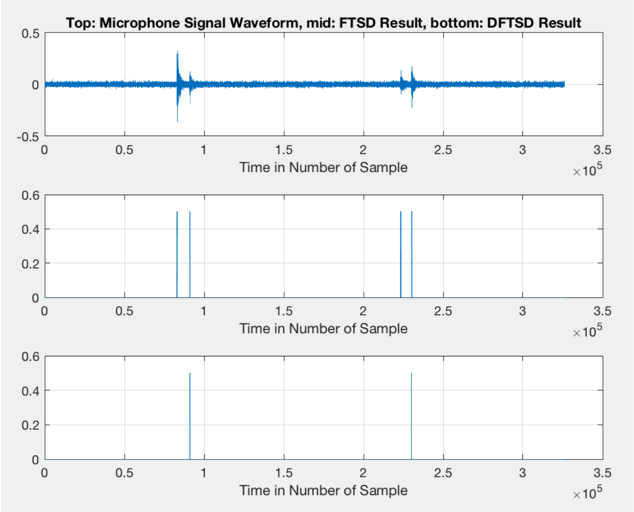


Fig. 7. An Example of Applying the Proposed FTSD and DFTSD to the Noisy Transients, input SNR = 3 dB

The top plot of Fig. 7 shows the waveform consisting of two pairs of clicker's clicking sounds in noisy environment with signal-to-noise ratio (SNR) of 3 dB. The middle and bottom plots have shown the outputs of the proposed FTSD and DFTSD, respectively. Obviously, the proposed FTSD and DFTSD algorithms work perfectly in this example.

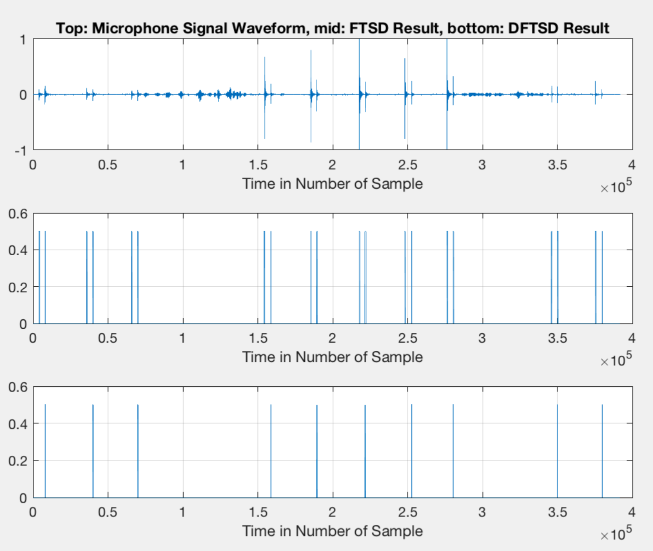


Fig. 8. An Example of Applying the Proposed FTSD and DFTSD to various Transients and Voice

The top plot of Fig. 8 shows the waveform consisting of voice and ten pairs of fast transient-sounds. The middle and bottom plots have shown the outputs of the proposed FTSD and DFTSD schemes, respectively, which demonstrates that the proposed algorithms are robust to the undesired voice and achieve 100% detection accuracy in this example.

Fig.9 and Fig. 10 compare the false negative and false positive results of the proposed DFTSD algorithm with those of the existing algorithm for 100 pairs of fast transient-sounds, respectively. No bar means 0%. The lower is the bar, the better is the detection performance. Test case of "Laughing S." includes one pair of true events and seven pairs of laughing sounds and voice signals.

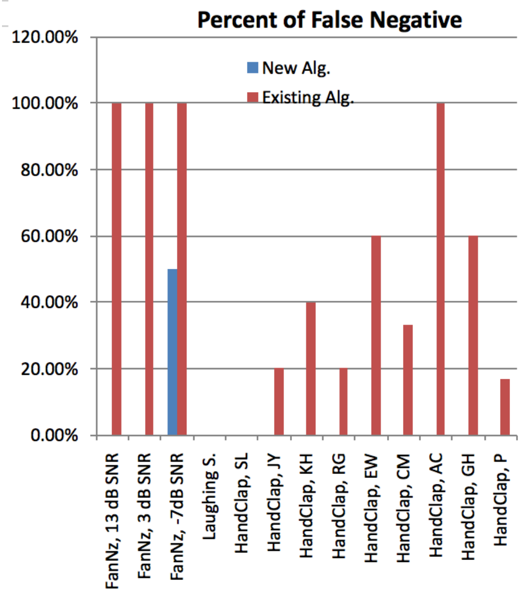


Fig. 9. False Negative of the Proposed DFTSD versus the Existing DFTSD

The test cases of Fig. 7 and Fig. 8 are also included in the bar of “FanNz, 3 dB SNR” and in the bar of “Handclap, AC” in Fig. 9, respectively.

It can be seen from the first three bars in Fig.9 that the existing DFTSD algorithm fails to detect all the double fast transient-sounds when environment has fan noise, even when the SNR is as high as 13 dB. Instead, the proposed DFTSD works perfectly for SNR of 13 dB and 3 dB, and even detects a pair of fast transient-sounds in the case of SNR = -7 dB. Therefore, the proposed DFTSD algorithm outperforms the existing algorithm in different SNR conditions.

Moreover, the last eight bars in Fig.9 shows that the existing algorithm fails to detect some double fast transient-sounds while the proposed algorithm correctly detects all the double fast transient-sounds.

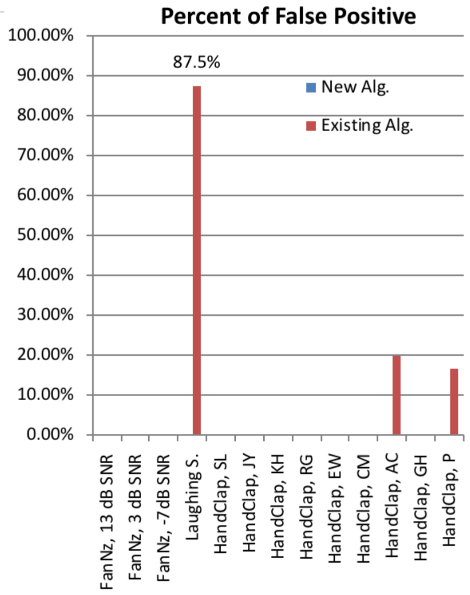


Fig. 10. False Positive of the Proposed DFTSD versus the Existing DFTSD

Fig. 10 shows that the existing algorithm falsely determines 7 pairs of laughing sounds and voice signals as 7 pairs of fast transient-sounds in the test case of “Laughing S.”, and hence results in a false positive rate as 7/8 = 87.5%, while the proposed algorithm has zero false detection under exactly the same conditions.

A comprehensive experiment on the proposed system is conducted with 47 pairs of hand-clap, 6 pairs of tap, and 1 pair of clicker’s clicking, i.e., total 54 pairs of fast transient-sounds. Table I provides the real-time test results.

TABLE I. PERCENT OF FALSE POSITIVE AND FALSE NEGATIVE

	False Positive	False Negative
The Proposed DFTD Alg.	1.85%	1.85%
The Existing DFTD Alg.	16.67%	44.44%

This suggests that the proposed DFTSD algorithm outperforms the existing DFTSD algorithm through having a much lower false positive rate and false negative rate.

VI. SUMMARY

As pointed out in Section 1, the existing FTSD algorithm only focuses on the basic features, such as energy, width of the pulse, energy ratio between two consecutive fast transient-sounds, etc.. Therefore, a major lacking is the robustness to various noises, the unrelated voices and transient-sounds. In addition, the existing FTSD algorithm is very expensive in terms of latency and computational complexity. Its design is independent from the training for the false positive and false negative performance, which in turn cannot maximize the detection performance.

This paper presents a new multirate FTSD algorithm by extracting the key features including dynamics and periodicity on the basis of the emerging microphone array configuration. The proposed FTSD algorithm has been trained and verified by a large database on Amazon machine learning platform. By using multirate approach, the proposed FTSD algorithm and its generalized version for DFTSD not only deliver a better detection performance but also consume much less MIPS and latency than the existing algorithms. Moreover, an effective and robust positioning system is established by making use of the proposed FTSD and DFTSD algorithms. The given theoretical analyses and objective test results show that the proposed system can offer a significant improvement for FTSD, DFTSD, sound source localization, and positioning performance in smart home devices. All of the above shows that the proposed algorithm can serve as a much more practical solution for the digital smart home related applications.

REFERENCES

[1] Daniel J. Esman, Vahid Ataie, et al., “Detection of Fast Transient Events in a Noisy Background,” Journal of Lightwave Technology, Vol. 34, No. 24, pp. 5669-5674, December 15, 2016

[2] Lutshayzar Gueorguieff, “Fast Transient Detection and Processing Algorithms for Time Scaling of Audio Files via a Rigid Phase-Locked Vocoder,” Proceedings of the 2012 International Conference on Information Technology and Software Engineering, Information Technology & Computering Intelligence, ITSE 2012, Beijing, China, pp.22-30, December 8-10, 2012

[3] Volker Gnann and Martin Spiertz, “Transient Detection with Absolute Discrete Group Delay,” 2009 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2009, Kanazawa, Japan, pp.311-314, December 7-9, 2009

[4] Jun Yang, “Multilayer Adaptation Based Complex Echo Cancellation and Voice Enhancement,” ICASSP 2018, Calgary, Albert, Canada, pp. 2131 - 2135, April 15-20, 2018

[5] Jun Yang, Philip Hilmes, Brian Adair, and David W. Krueger, "Deep Learning Based Automatic Volume Control and Limiter System," ICASSP 2017, New Orleans, USA, pp. 2177 - 2181, March 5 - 9, 2017