Adaptive Noise Suppression for Wake-Word Detection by Temporal-Difference Generalized Eigenvalue Beamformer

Takehiko Kagoshima*, Ning Ding* and Hiroshi Fujimura*
* Toshiba Corporation R&D Center, Kawasaki, Japan
E-mail: {takehiko.kagoshima, ning.ding, hiroshi4.fujimura}@toshiba.co.jp

***Abstract*—This paper proposes an adaptive noise suppression method for wake-word detection by a temporal-difference generalized eigenvalue (TDGEV) beamformer. To emphasize wake-word utterances, which are leading phrases with short duration, the proposed method is based on a generalized eigenvalue beamformer regarding current and past spatial covariance matrices for speech and noise, respectively. It can emphasize wake-words with small distortion and suppress any noises regardless of directions of arrival (DoAs) and noise sources. We perform experiments of wake-word detection with and without beamformers using test data including wake-word utterances from various DoAs. The results show that the proposed TDGEV method reduces false rejects with 32.9% relative error rate reduction.**

Index Terms: keyword spotting, wake-word detection, microphone array, adaptive beamforming, generalized eigenvalue beamformer

I. Introduction

Speech-controlled devices, such as smart speakers and car navigation systems, are gaining popularity in daily life. These devices use keyword spotting (KWS) methods to initiate speech recognition or directly execute commands. Because speech-controlled devices are often distant from users and can be located in noisy environments, noise robustness is crucial for KWS methods.

Introducing a beamformer with a microphone array for front-end KWS processing is an effective approach to improving noise robustness. Many adaptive beamforming methods have been proposed for noise suppression [1], with the generalized sidelobe canceller (GSC) being one well-known method [2], [3]. GSC passes signals from predetermined target directions of arrival (DoAs) and suppresses those from other DoAs. Its algorithm is based on an adaptive filter whose coefficients are updated to minimize filter output under a constraint to pass target DoAs. GSC achieves better signal-to-noise ratio (SNR) improvement than does fixed beamformers, but it is not suited to applications such as smart speakers, where keyword utterance DoAs are unknown and thus impossible to set as a target DoA beforehand.

Mask-based beamformers supported by neural networks (NNs) [4], [5], [6] have been proposed to emphasize speech signals under conditions in which DoAs are unknown. These methods use max-SNR beamformers such as a generalized eigenvalue (GEV) beamformer [7] or a minimum variance distortionless response beamformer [8]. These beamformers are calculated based on speech and noise spectra estimated from input signals. The NN is trained to separate speech and noise from input spectra, so speech signals can be emphasized under blind conditions. However, these methods cannot suppress undesired speech signals from televisions in home environments and other common noise sources, making them unsuited to applications such as smart speakers.

Hotword cleaner (HC) [9], [10] overcomes such problems by focusing on wake-word detection applications. In contrast to KWS applications, which attempt keyword detection in continuous speech, HC limits its detection to wake-words, which are leading phrases with short durations (typically less than 1 s). In the HC method, an adaptive filter is updated to minimize power output. Filter coefficients are stored in a buffer for a short term, corresponding to wake-word lengths. Beamforming is performed using the filter coefficients read from the buffer. When a wake-word is observed, the filter can suppress only noise signals, because the wake-word is not used to adapt the filter. As a result, HC emphasizes a wake-word against any noises, including television speech noise. However, wake-word signal distortion can be problematic, because the adaptive filter is optimized using only noise signals, and thus does not guarantee a response from wake-word signals.

To improve accuracy of wake-word detection by blind beamforming with low distortion against various noise sources, including speech noise, we propose an adaptive noise suppression method for wake-word detection by a temporal-difference generalized eigenvalue (TDGEV) beamformer. Because the proposed method is based on the assumption of a wake-word, its target is wake-word detection as a special application of KWS. We introduce a GEV beamformer whose spatial covariance matrices for speech and noise are respectively calculated using current observation signals and those from a short preceding term. When a wake-word is observed, the beamformer is adapted to form a beam pattern that maximizes output SNR. Thus, the proposed method effectively suppresses all noise types under blind conditions, preventing wake-word distortion.

A related work [11] proposed a multi-channel KWS method using multiple fixed beamformers. In this method, the input signal for KWS is the weighted sum of output signals from four fixed beamformers using a 4-ch microphone array. Weights are predicted by a NN trained with another NN for KWS in an end-to-end manner. The fixed beamformers need more microphones to form a sufficient beam pattern and the weights for the beamformers depend on the following KWS module. In this paper, we evaluate 2-ch versions of the proposed TDGEV method and the HC method to compare the beamforming methods independent of KWS methods.

We use a DNN-based KWS method [12] to evaluate both beamforming methods. Experimental results show that the

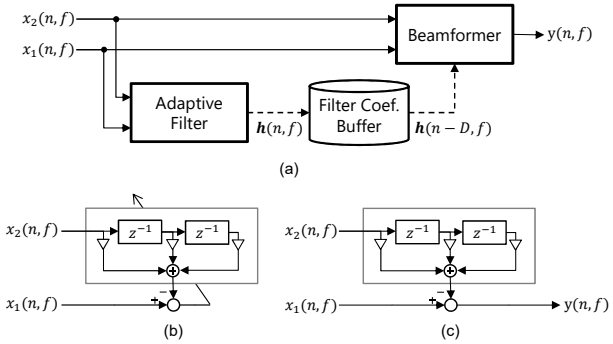

Fig. 1. Hotword Cleaner process flow: (a) overview, (b) adaptive filter, (c) beamformer.

proposed method suppresses noise, including television noise, and reduces false rejects of wake-word detection by 35.7% relative to the HC method.

The reminder of this paper is organized as follows. Section II reviews the conventional HC method. Section III proposes the TDGEV method. Section IV introduces our keyword detection method. Section V describes experiments for wake-word detection to evaluate the beamforming methods, and Section VI concludes this paper.

II. Hotword cleaner

This section briefly reviews the 2-ch version of the conventional HC method [9], shown in Fig. 1. The inputs are short-time Fourier transform (STFT) domain signals represented by $x_1(n,f)$ and $x_2(n,f)$, where n and f are respectively time-frame and frequency-bin indexes. Beamforming is independently applied to each frequency bin. Figs. 1(b) and (c) respectively show structures of the adaptive filter and the beamformer where the number of filter taps $L = 3$. The adaptive filter updates coefficients $\boldsymbol{h}(n,f) = [h_0(n,f), h_1(n,f), \cdots, h_{L-1}(n,f)]^T$ to minimize output power using a fast recursive least squares (RLS) algorithm. The filter coefficient buffer delays the filter coefficients by D frames. The beamformer with coefficients $\boldsymbol{h}(n-D,f)$ generates output $y(n,f)$ from the input signals. Because the parameter D is set to be larger than the wake-word duration, when microphones observe a wake-word, the beamformer has been adapted using noise signals from up to the previous D frames. Therefore, assuming a stationary DoA for noise during those D frames, the beamformer suppresses noise but not wake-words. As a result, HC emphasizes wake-words and suppresses any noise with stationary DoA. However, beamformer gain for wake-words depends on wake-word DoAs and frequency bins, because it is adapted using only noise signals. This can suppress or degrade wake-words by gain dispersion for each frequency bin.

III. TDGEV

This section presents the proposed TDGEV method. Fig. 2 shows the 2-ch version of TDGEV method. Like HC, this is a STFT-domain beamformer. Unlike HC, however, the beamformer in the proposed method is adapted using both past noise signals and currently observed wake-word signals to maximize output SNR.

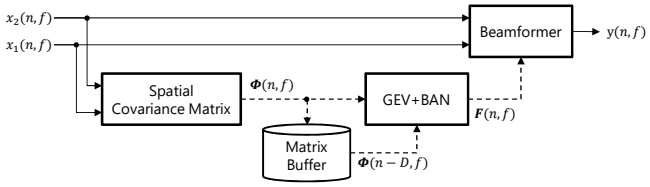

Fig. 2. Process flow of the proposed TDGEV method.

First, a spatial covariance matrix $\boldsymbol{\Phi}(n,f)$ is calculated using the exponential moving average as

$$\boldsymbol{\Phi}(n,f) = \alpha\boldsymbol{\Phi}(n-1,f) + (1-\alpha)\boldsymbol{x}(n,f)\boldsymbol{x}^H(n,f), \quad (1)$$

where $\boldsymbol{x}(n,f) = [x_1(n,f), x_2(n,f)]^T$ is a vector of input STFT spectra and α is a forgetting factor. Then, a matrix buffer delays $\boldsymbol{\Phi}(n,f)$ for D frames. The delay parameter D is set to be larger than the wake-word duration. Coefficients of the beamformer $\boldsymbol{F}(n,f)$ are obtained using the current and past spatial covariance matrices $\boldsymbol{\Phi}(n,f)$ and $\boldsymbol{\Phi}(n-D,f)$. The beamformer $\boldsymbol{F}(n,f)$ is composed of a GEV beamformer $\boldsymbol{F}_{GEV}(n,f)$ and a post filter $g(n,f)$ based on blind analytical normalization (BAN) as

$$\boldsymbol{F}(n,f) = g(n,f)\boldsymbol{F}_{GEV}(n,f) \quad (2)$$

$$g(n,f) = \frac{\sqrt{\boldsymbol{F}^H_{GEV}(n,f)\boldsymbol{\Phi}(n-D,f)\boldsymbol{\Phi}(n-D,f)\boldsymbol{F}_{GEV}(n,f)/2}}{\boldsymbol{F}^H_{GEV}(n,f)\boldsymbol{\Phi}(n-D,f)\boldsymbol{F}_{GEV}(n,f)}, \quad (3)$$

where the GEV beamformer $\boldsymbol{F}_{GEV}(n,f)$ is the eigenvector corresponding to the largest eigenvalue of a matrix $\boldsymbol{\Phi}^{-1}(n-D,f)\boldsymbol{\Phi}(n,f)$. Finally, the output $y(n,f)$ is obtained by the beamforming

$$y(n,f) = \boldsymbol{F}^H(n,f)\boldsymbol{x}(n,f). \quad (4)$$

In the proposed method, the current and past spatial covariance matrices $\boldsymbol{\Phi}(n,f)$ and $\boldsymbol{\Phi}(n-D,f)$ are for speech and noise, respectively. When a wake-word is observed, $\boldsymbol{\Phi}(n,f)$ indicates the wake-word DoA and $\boldsymbol{\Phi}(n-D,f)$ indicates the noise DoA. Thus, supposing $\boldsymbol{\Phi}(n-D,f)$ approximates the current spatial covariance matrix for noise, the proposed method maximize the output SNR regardless of the wake-word DoAs. Moreover, the BAN post-filter normalizes the gain of each frequency bin, thus reducing distortion of emphasized wake-words. However, the proposed method does not suppress stationary noise when no wake-word is observed. The method as described above is a 2-ch version, but it can be easily expanded to more channels.

IV. Keyword detection

This section describes our keyword detection method [12] as used for evaluation of the beamforming methods. This keyword detection method is designed for both wake-word detection tasks and general keyword-spotting tasks. This method comprises a NN to calculate phoneme-state probability and a keyword detector based on a customized Viterbi algorithm. The method can detect arbitrary keywords by specifying their phoneme sequences, because the NN supports a full set of phonemes. The NN, which comprises feed-forward layers, is

trained with a large speech corpus, including various utterances independent of keywords. In the following experiments, we use a NN having 4 hidden layers with 256 nodes each.

Keywords are represented as a left-to-right Hidden Markov model state sequence $\{s_1, s_2, \cdots, s_N\}$, where N is the total number of states. The NN predicts $score(\boldsymbol{x}_\tau, s_j)$ between an input speech feature vector $\boldsymbol{x}_\tau$ and a phoneme state s_j in the keyword, where τ and j denote time frame and state indexes, respectively. The keyword score $S(b, e)$ corresponding to a segment from a start frame b to an end frame e is

$$S(b,e) = \frac{1}{e-b+1} \max_{Q} \sum_{\tau=b}^{e} score(\boldsymbol{x}_\tau, s_{q_\tau}), \quad (5)$$

where q_τ indicates a state number mapped to $\boldsymbol{x}_\tau$, and Q is a set of possible state number sequences $\{q_b, q_{b+1}, \cdots, q_e\}$. A keyword is detected when a speech segment with $S(b, e) > \theta$ exists, where θ is a predefined threshold. The calculation for finding the segment $[b, e]$ in all observed speech segments with length T is

$$\max_{1 \le b < e \le T} \frac{1}{e-b+1} \max_{Q} \sum_{\tau=b}^{e} score(\boldsymbol{x}_\tau, s_{q_\tau}) > \theta. \quad (6)$$

The computational cost when using the conventional Viterbi algorithm to search for the segment is $O(NT^3)$. To reduce computations, we restate (6) as

$$\max_{1 \le b < e \le T} \max_{Q} \sum_{\tau=b}^{e} \{score(\boldsymbol{x}_\tau, s_{q_\tau}) - \theta\} > 0. \quad (7)$$

This equation does not have a normalization term $(e-b+1)$, so intermediate scores can be determined independent of e and b. Therefore, (7) can be efficiently solved by a Viterbi-like algorithm, and the computational cost is drastically reduced to $O(NT)$ without any accuracy degradation.

V. Experiments

To evaluate the performance of the proposed TDGEV method versus the conventional HC method, we conducted experiments using test data including wake-word utterances from various DoAs. The test data are made by adding wake-word data over noise data. False accepts and false rejects are measured by applying the keyword detection method to the test data with and without beamformers. Performance of a wake-word detection method is illustrated by a receiver operating characteristic (ROC) curve, which is drawn by plotting false rejects and false accepts for various values of the threshold parameter θ. Beamformer performance can be evaluated by comparing ROC curves. Because some beamformer parameters depend on the test data, we performed parameter tuning for both HC and TDGEV using the test data before comparing the two methods.

A. Test data

Wake-word data were recorded in a soundproof room using a 2-ch microphone array of omni-directional microelectromechanical-systems microphones with 2 cm distance. Fig. 3 shows a layout of the recording. Source data for wake-words were played back from one of the 7 DoAs $(-90°, -60°, -30°, 0°, +30°, +60°$, and $+90°)$ in a mouth simulator. Wake-word source data comprise 15 Japanese full

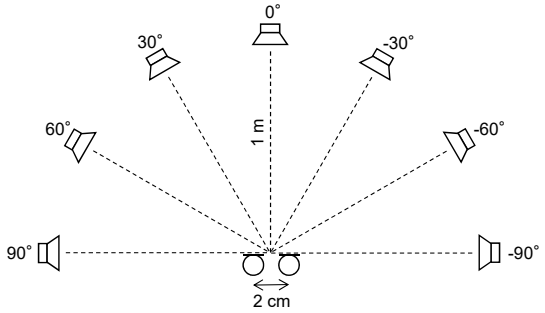

Fig. 3. Recording layout for wake-word data.

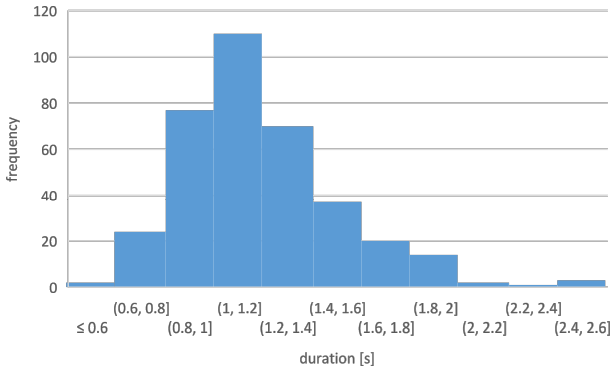

Fig. 4. Histogram of wake-word utterance durations.

names uttered twice by each of 12 speakers (5 female, 7 male). Fig. 4 shows a histogram of durations for all 360 wake-word utterances. Numbers of syllables in the Japanese names ranged from 5 to 10, so duration variance exceeded that for a single wake-word task.

Noise data were recorded in actual environments. Each 60 min noise data segment was recorded in the five environments listed in Table I using the microphone array described above.

Test data were generated by adding a set of the 360 wake-word data over each noise dataset. Wake-word directions were selected at random from among the seven DoAs. Timings at which wake-words were added to noise data were also randomized. SNR was controlled by adjusting the power of wake-word data versus that of the overlapping noise data segment. We generated five test datasets with SNRs of 20, 15, 10, 5, and 0 dB. Test data were saved as 16 kHz samplings of 16-bit linear PCM.

TABLE I
Recording conditions for noise data.

Place	Predominant noise
Living room	Ambient noise
Living room	Speech from a television
Living room	Music from stereo speakers
Kitchen	Cooking and splashing noises
Office	Noise from keyboards, printers, and speech

TABLE II
FALSE REJECTS IN WAKE-WORD DETECTION USING HC AND TDGEV FOR COMBINATIONS OF STFT PARAMETERS WITH $FAh = 0.1$.

STFT parameters		forgetting factor		False Rejects [%]	
L_{FFT}	L_{shift}	λ(HC)	α(TDGEV)	HC	TDGEV
2048	1024	0.993	0.923	11.1	7.3
2048	512	0.997	0.961	10.7	6.8
1024	512	0.997	0.961	10.9	7.4
1024	256	0.998	0.980	10.3	6.9
512	256	0.998	0.980	10.9	6.4
512	128	0.999	0.990	10.9	7.0

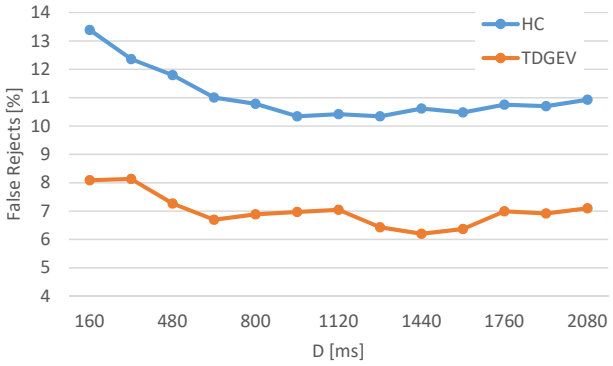

Fig. 5. False rejects of wake-word detection using HC and TDGEV for D values with $FAh = 0.1$.

B. Experimental results

We optimized the STFT parameters (FFT window length L_{FFT} and frame shift L_{shift}) and the delay parameter D, because they are common parameters for TDGEV and HC. Other HC parameters were set following Ref. [9] as follows: filter length $L = 3$, initial parameter $\delta = 0.1$, and forgetting factor $\lambda = 0.993$ for $L_{\text{shift}} = 1024$. Based on a preliminary experiment, the forgetting factor for TDGEV was set as $\alpha = 0.990$ for $L_{\text{shift}} = 128$.

We first evaluated both methods using the L_{FFT} and L_{shift} combinations listed in Table II with $D = 1280$ ms. Because forgetting factors depend on L_{shift}, these were set as $\alpha(2L_{\text{shift}}) = \alpha(L_{\text{shift}})^2$ and $\lambda(L_{\text{shift}}/2) = \sqrt{\lambda(L_{\text{shift}})}$ using the initial values $\alpha(128) = 0.990$ and $\lambda(1024) = 0.993$. False rejects in Table II show the values for the following $FAh = 0.1$:

$$FAh = \frac{\#FA}{T_{\text{test}} * \#words}, \tag{8}$$

where $\#FA$ is the number of false accepts, T_{test} is the duration of test data in hours, and $\#words$ is the number of wake-word types. As Table II shows, the best combinations of $(L_{\text{FFT}}, L_{\text{shift}})$ for HC and TDGEV are $(1024, 256)$ and $(512, 256)$, respectively.

We next searched for the best D parameter for both methods using the best $(L_{\text{FFT}}, L_{\text{shift}})$ combinations derived above. Fig. 5 shows false reject values for $FAh = 0.1$ corresponding to D values ranging from 160 to 2080 ms. The best values for D were 960 for HC and 1440 for TDGEV.

Finally, we compared the ROC curves in Fig. 6 for HC, TDGEV, and the baseline, which is the average of the results for each of the left and right channels without beamformers. The value of false rejects for the proposed TDGEV method is remarkably reduced from the baseline. A relative error rate reduction (RERR) for TDGEV at $FAh = 0.1$ was 32.9% versus the baseline. False rejects for the conventional HC method is not reduced from the baseline. Fig. 7 shows the value of false rejects at $FAh = 0.1$ for DoAs of wake-word data. HC reduces false rejects from the baseline for the DoAs of $-90°$, $-60°$, $+60°$, and $+90°$ but not for other DoAs. TDGEV reduces false rejects for all the DoAs in contrast to HC. Fig. 8 shows examples of the final results.

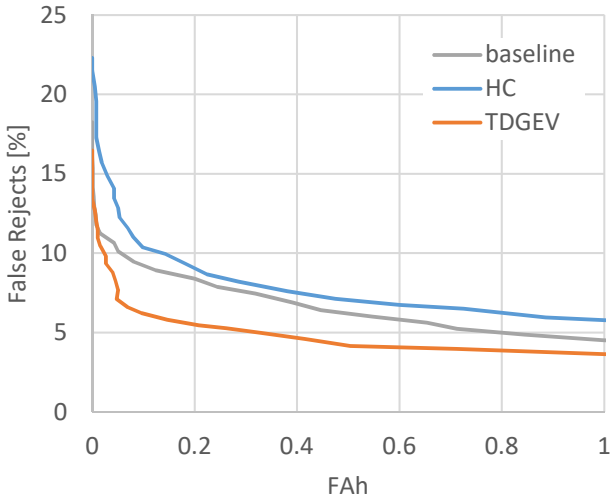

Fig. 6. ROC curves of wake-word detection for HC, TDGEV, and without beamformers.

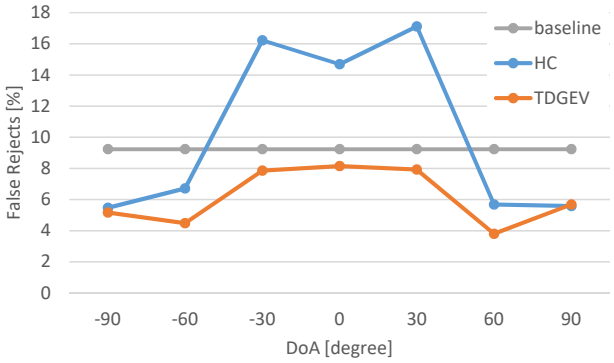

Fig. 7. False rejects of wake-word detection with $FAh = 0.1$ for DoAs of wake-word utterances.

C. Discussion

Figs. 8(c) and (d) indicate that HC is better in terms of noise suppression, and TDGEV is better in terms of low distortion of wake-words. This result supports the theoretical advantages and disadvantages of the two methods. Fig. 6 and 7 show that the proposed TDGEV method is superior to the conventional HC method in terms of wake-word detection performance because TDGEV is robust for wake-words' DoAs.

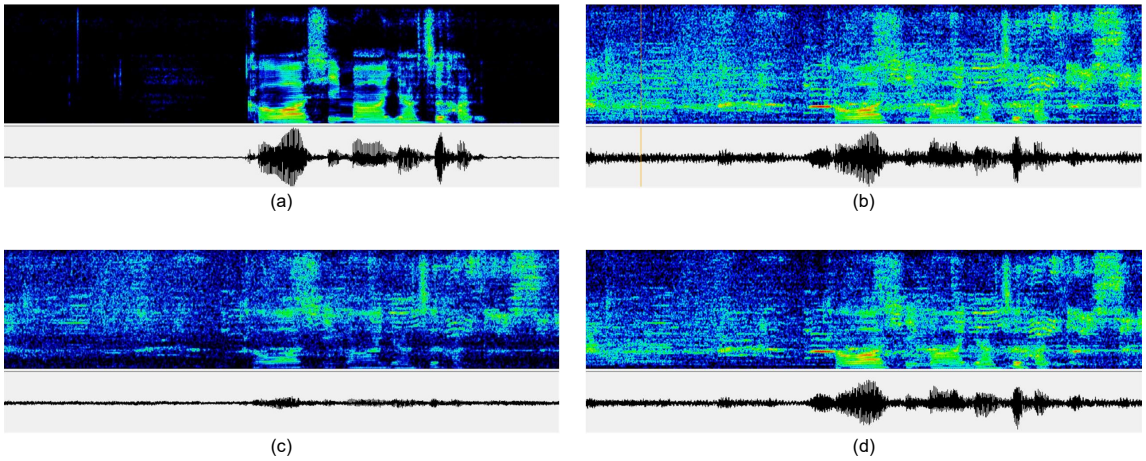

Fig. 8. Examples of waveforms and spectrograms: (a) wake-word source data, (b) test data generated from the wake-word data with DoA=$-30°$ and television noise under 5 dB SNR, (c) test data processed by HC, and (d) test data processed by TDGEV.

This result suggests that low-distortion beamforming by the proposed GEV-based method improves wake-word detection performance.

Table II shows that TDGEV can perform with smaller FFT window lengths. Thus, additional latency by TDGEV is 32 ms with the best parameter and that of HC is 64 ms. This insensitivity to FFT window length is another advantage. A keyword detection module can be connected to the TDGEV beamformer in the STFT domain without converting to the time domain by adopting a FFT window length common between the two modules, thereby reducing computation times and latency.

VI. Conclusion

We proposed the TDGEV method for adaptive noise suppression in wake-word detection. The proposed method is based on a GEV beamformer that regards current and previous spatial covariance matrices as those for speech and noise, respectively. It emphasizes a wake-word utterance, which is a leading phrase with short duration, and suppresses any noises with stable DoA regardless of DoA and noise source. Experimental results showed that the proposed TDGEV method improved wake-word detection performance with 32.9% RERR for test data including wake-word utterances from various DoAs versus the baseline without beamformers.

References

[1] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov. A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 25:602–730, April 2017.

[2] L. J. Griffiths and C. W. Jim. An alternative approach to linear constrained adaptive beamforming. *IEEE Trans. AP*, AP-30:27–34, Jan 1982.

[3] O. Hoshuyama, A. Sugiyama, , and A. Hirano. A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. *IEEE Trans. SP*, 47:2677–2684, 1999.

[4] J. Heymann, L. Drude, and R. Haeb-Umbach. Wide residual blstm network with discriminative speaker adaptation for robust speech recognition. In *The 4th International Workshop on Speech Processing in Everyday Environments (CHiME2016)*, page 12–17, 2016.

[5] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani. Robust mvdr beamforming using time-frequency masks for online/offline asr in noise. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5210–5214, 2016.

[6] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach. Exploring practical aspects of neural mask-based beamforming for far-field speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[7] Reinhold Haeb-Umbach Ernst Warsitz. Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1529–1539, 2007.

[8] S. Araki, H. Sawada, and S. Makino. Blind speech separation in a meeting situation with maximum snr beamformers. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 1, pages I–41–I–44, 2007.

[9] Y. A. Huang, T. Z. Shabestary, and A. Gruenstein. Hotword cleaner: Dual-microphone adaptive noise cancellation with deferred filter coefficients for robust keyword spotting. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6346–6350, 2019.

[10] Y. A. Huang, T. Z. Shabestary, A. Gruenstein, and L. Wan. Multi-microphone adaptive noise cancellation for robust hotword detection. In *INTERSPEECH 2019, 20th Annual Conference of the International Speech Communication Association*, pages 1233–1237, 2019.

[11] X. Ji, M. Yu, J. Chen, J. Zheng, D. Su, and D. Yu. Integration of multi-look beamformers for multi-channel keyword spotting. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7464–7468, 2020.

[12] H. Fujimura, N. Ding, D. Hayakawa, and T. Kagoshima. Simultaneous flexible keyword detection and text-dependent speaker recognition for low-resource devices. In *10th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2020.