INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Independent Echo Path Modeling for Stereophonic Acoustic Echo Cancellation

Yi Gao[1], *Ian Liu*[2], *J. Zheng*[3] *Cheng Luo*[1], *Bin Li*[1]

[1]Tencent WeChat Work, Chengdu, China
[2]Amazon Lab126, Shenzhen, China
[3]Tencent AI Lab, Shenzhen, China

jackyyigao@tencent.com, ian0302@gmail.com,
{jimzzheng, plusluo, wisonlee}@tencent.com

Abstract

As stereophonic audio devices, such as smart speakers and cellphones, evolve to be daily essentials, stereophonic acoustic echo cancellation becomes more important for voice and audio applications. The cross-correlation between the far-end channels and the associated ambiguity in the estimated echo path transfer functions lead the misalignment and instability issues with conventional stereophonic acoustic echo cancellers (SAEC). In this paper, we propose a novel SAEC algorithm, which can better model the acoustic echo path between each loudspeaker and microphone. Specifically, filter adaptations are modeled independently by applying pre-whitening in solving the misalignment problem. Improvement in echo suppression capability is evaluated in terms of echo return loss enhancement(ERLE) and wakeup word detection accuracy.

Index Terms: SAEC, misalignment, pre-whitening, wakeup word detection.

1. Introduction

As devices with stereo loudspeakers, e.g. smart speakers and cellphones, being widely deployed recently for both voice communication and human-machine interaction, stereophonic acoustic echo cancellation (SAEC) becomes a critical component of the voice enhancement system. Its echo cancellation performance has significant impact on the listening experience, wakeup word detection, and speech recognition.

In the past decades, SAEC as well as its underlying misalignment problem have been investigated, e.g.[1-4], and various solutions have been proposed. The fundamental difficulty with SAEC is the high cross-correlation between the two playback channels, which is very common in practical usages. Since the adaptive filters (AF) employed by SAEC is adapted under the guidance of the final cancellation residual, such cross-correlation causes the AF not converging to the true echo path transfer functions even though the cancellation residual is temporarily small. As a result, when the cross-correlation between the two far-end channels abruptly changes, significant echo leakage will be observed and it costs extra time for the AF algorithm to achieve another temporary convergence. Thus, in practice, SAEC algorithms usually suffer from the problem of slow convergence and/or repeated divergence-to-reconvergence behavior. To tackle this problem, conventional SAEC solutions generally take two types of approaches: 1) modifying the far-end signals so that they are less correlated, e.g. [5][6]; 2) increasing the contribution of the decorrelated components in the preprocessed transmitted signals[7]. 3) exploiting the short-term variation of the cross-correlation, e.g. [2], which is available in most applications. Further, recent advances in deep learning literature has motivated the reformulation of SAEC as a supervised speech separation problem along with a DNN based solution[8][9].

Modifying far-end signal to be less correlated is usually achieved by either adding small amount of random noise or nonlinear processing on the far-end signals before sending them to the loudspeaker. Although the user perception degradation introduced by such modification is roughly controllable, this type of approach is still not preferred by many high-fidelity applications, including the music playback scenario of smart-speakers. On the other hand, deep-learning based solutions showed promising echo suppression performance but their computational complexity is usually prohibitive for mobile and IOT devices. Further, such solutions are vulnerable to low signal-to-echo-ratio(SER) and its robustness highly relies on the availability of recorded data.

In this paper, we will focus on exploiting the short-term variation of the correlation and a novel AF-based algorithm is proposed, which models the separate echo paths by applying pre-whitening process and independently updating each adaptive filter. Such improvement is achieved with a moderate cost of computational complexity. The algorithm's echo cancellation performance is evaluated through real life data in terms of ERLE and wakeup word detection rate.

2. Problem Formulation

The problem is formulated in short time Fourier transform (STFT) domain, where l and k denote the time-frame index and the frequency-bin index, respectively.

$$\boldsymbol{d}(l,k) = \boldsymbol{s}(l,k) + \sum_{n=1}^{N} \boldsymbol{z}_n(l,k) \quad (1)$$

Where $\boldsymbol{d}(l,k)$ denotes the microphone input signal that includes near-end signal $\boldsymbol{s}(l,k)$ and echo signal $z_n(l,k)$ from the n-th loudspeaker. Our frequency domain adaptive filtering（FDAF）uses multidelay block NLMS[10-13], B indicates the block number:

$$\widehat{\boldsymbol{U}}_n(l,k) = [\boldsymbol{U}_n(l,k), \boldsymbol{U}_n(l-1,k), \dots, \boldsymbol{U}_n(l-B,k)]^T \quad (2)$$

$$\widehat{\boldsymbol{W}}_n(l,k) = [\boldsymbol{W}_n(1,k), \dots, \boldsymbol{W}_n(B,k)]^T \quad (3)$$

Where $\widehat{\boldsymbol{U}}_n(l,k)$ denotes far-end signal and $\widehat{\boldsymbol{W}}_n(l,k)$ denotes acoustic echo path transfer function. The general error-signal for SAEC system can be expressed by

$$\boldsymbol{e}(l,k) = \boldsymbol{d}(l,k) - \sum_{n=1}^{2} \widehat{\boldsymbol{W}}_n^H(l,k)\widehat{\boldsymbol{U}}_n(l,k) \quad (4)$$

The NLMS solution for this problem is:

$$\widehat{\boldsymbol{W}}_n(l+1,k) = \widehat{\boldsymbol{W}}_n(l,k) + diag\{\widehat{\boldsymbol{a}}_n\}\widehat{\boldsymbol{U}}_n(l,k)\boldsymbol{e}(l,k) \quad (5)$$

http://dx.doi.org/10.21437/Interspeech.2020-2131

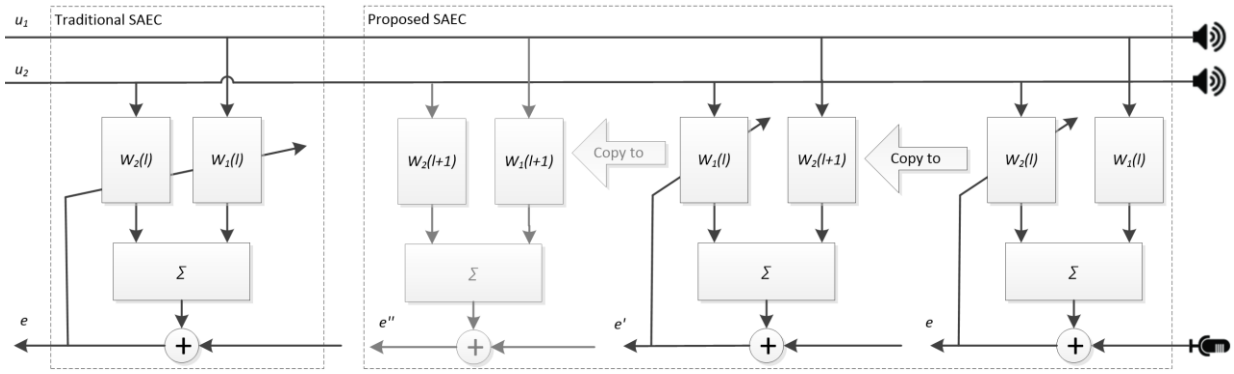

Figure 1: *Schematic diagram for the traditional and proposed method for SAEC system*

$$\widehat{\boldsymbol{a}}_n(l,k) = [\boldsymbol{a}_n(l,k), \dots, \boldsymbol{a}_n(l-B,k)]^T \quad (6)$$

Where $\boldsymbol{a}_n(l,k) \equiv \frac{\mu}{E\|\boldsymbol{U}_n(l,k)\|^2+\epsilon}$, μ is the step size and ϵ is regularization factor. While a variable step-size is usually use, the derivation for it is beyond the scope of this paper.

3. IEPM-based SAEC

3.1. IEPM-based solution

The proposed independent echo path modeling(IEPM) SAEC algorithm is modified from above mentioned fast block NLMS approach. The computation steps are listed in Table 1, where **0** is M-by-1 null vector.

Table 1: *The proposed IEPM SAEC Algorithm*

Computation Steps (For each new block of *M* input samples)

Step 1) Filtering:

$$\boldsymbol{U}_n(l,k) = STFT\{[\boldsymbol{u}_n(l*M-M+1), \dots, \boldsymbol{u}_n(l*M)]^T\}$$

$$\widehat{\boldsymbol{U}}_n(l,k) = [\boldsymbol{U}_n(l,k), \boldsymbol{U}_n(l-1,k), \dots, \boldsymbol{U}_n(l-B,k)]$$

$$\widehat{\boldsymbol{W}}_n(l,k) = [\boldsymbol{W}_n(1,k), \dots, \boldsymbol{W}_n(B,k)]^T$$

$$\boldsymbol{Y}_n(l,k) = \widehat{\boldsymbol{W}}_n^H(l,k)\ \widehat{\boldsymbol{U}}_n(l,k)$$

$$\boldsymbol{y}_n(l) = first\ of\ half\ of\ ISTFT[\boldsymbol{Y}_n(l,k)]$$

Step 2) Error estimation:

$$\boldsymbol{e}(l) = \boldsymbol{d}(l) - \boldsymbol{y}_1(l) - \boldsymbol{y}_2(l)$$

$$\boldsymbol{E}_2(l,k) = STFT\left\{\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{e}(l)\end{bmatrix}\right\}$$

Step 3) Signal-power estimation:

$$\boldsymbol{P}(l,k) = \gamma\boldsymbol{P}(l-1,k) + (1-\gamma)\sum_{n=1}^{2}|\boldsymbol{U}_n(l,k)|^2$$

$$\boldsymbol{PD}(l) = \gamma\boldsymbol{PD}(l-1,k) + (1-\gamma)\sum_{n=1}^{2}\sum_{k=1}^{K}|\boldsymbol{U}_n(l,k)|^2$$

$$\widehat{\boldsymbol{D}}(l,k) = diag\{[P^{-1}(l,k), \dots, P^{-1}(l-B,k)]\}$$

Step 4) Filter2 independent adaptation:

$$\widehat{\boldsymbol{\phi}}_2(l,k) = first\ half\ of\ ISTFT\left[\widehat{\boldsymbol{D}}(l,k)\widehat{\boldsymbol{U}}_2^H(l,k)\boldsymbol{E}_2(l,k)\right]$$

$$\widehat{\boldsymbol{W}}_2(l+1,k) = \widehat{\boldsymbol{W}}_2(l,k) + \alpha * STFT\left\{\begin{bmatrix}\widehat{\boldsymbol{\phi}}_2(l,k), & if\ \boldsymbol{PD}(l) > threshold\\ \boldsymbol{0}, & otherwise\end{bmatrix}\right\}$$

Step 5) Filtering using new filter2:

$$\boldsymbol{y}_2'(l) = first\ half\ of\ ISTFT\{\widehat{\boldsymbol{W}}_2^H(l,k)\ \widehat{\boldsymbol{U}}_2(l,k)\}$$

Step 6) Error re-estimation:

$$\boldsymbol{e}'(l) = \boldsymbol{d}(l) - \boldsymbol{y}_1(l) - \boldsymbol{y}_2'(l)$$

$$\boldsymbol{E}_1(l,k) = STFT\left\{\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{e}'(l)\end{bmatrix}\right\}$$

Step 7) Filter 1 independent adaptation:

$$\widehat{\boldsymbol{\phi}}_1(l,k) = first\ half\ of\ ISTFT\left[\widehat{\boldsymbol{D}}(l,k)\widehat{\boldsymbol{U}}_1^H(l,k)\boldsymbol{E}_1(l,k)\right]$$

$$\widehat{\boldsymbol{W}}_1(l+1,k) = \widehat{\boldsymbol{W}}_1(l,k) + \alpha * STFT\left\{\begin{bmatrix}\widehat{\boldsymbol{\phi}}_1(l,k), & if\ \boldsymbol{PD}(l) > threshold\\ \boldsymbol{0}, & otherwise\end{bmatrix}\right\}$$

Step 8) Filtering using new filter1:

$$\boldsymbol{y}_1'(l) = first\ half\ of\ ISTFT\{\widehat{\boldsymbol{W}}_1^H(l,k)\widehat{\boldsymbol{U}}_1(l,k)\}$$

Step 9) Final error estimation:

$$\boldsymbol{e}''(l) = \boldsymbol{d}(l) - \boldsymbol{y}_1'(l) - \boldsymbol{y}_2'(l)$$

The proposed method in Table 1, $\mathbf{e}''(l)$ will be used for the output. And the computational complexity is obvious doubled in comparing to traditional fast block NLMS algorithm. To reduce computational cost, step 8) and step 9) can be avoided because the two filters have already been updated and $\mathbf{e}'(l)$ and can be then used as the SAEC output. This simplified method is denoted as ProposedS. The process of the traditional and the proposed methods are illustrated in Figure 1.

4. Experiments

To illustrate the performance improvements by the proposed SAEC algorithms, we conducted tests with using real life test samples by measuring ERLE and wakeup word detection(WWD) rate. We defaulted same filter length equal to 2048 with frame size 256 and NFFT size 512 for obtaining below test results.

Typical constants used in the simulator are block number setting to 8, power smoothing factor γ setting to 0.99 and μ

Table 2*: ERLE under quiet and noisy conditions(dB)*

	no noise	babble		destroyerengine		facttory2		volvo		white		Average
		6dB	20dB	6dB	20dB	6dB	20dB	6dB	20dB	6dB	20dB	
Traditional	17.48	14.73	17.29	13.35	17.15	12.60	17.05	12.51	17.03	11.70	16.90	15.03
Proposed	18.58	15.62	18.37	14.15	18.20	13.50	18.10	13.40	18.09	12.51	17.94	15.99
ProposedS	18.17	15.31	17.98	13.89	17.82	13.25	17.73	13.15	17.71	12.29	17.57	15.67
Proposed Δ	1.10	0.89	1.07	0.80	1.05	0.90	1.06	0.89	1.06	0.81	1.04	0.96
ProposedS Δ	0.69	0.59	0.68	0.54	0.67	0.64	0.68	0.64	0.68	0.59	0.67	0.64

setting to 0.008, ϵ is 0.001. All test samples were recorded and processed at sampling rate of 16kHz.

The recordings were collected from a conference room with surrounded by concrete walls. The device used for conducting the test is Tencent Dingdang smart speaker with display. And it has two stereo loudspeakers. Playback level was set to between 70% to 100% of max volume. Totally 421 utterances of the wakeup word is recorded from 12 people while music is playing. Averaged SER is around -13.20dB which was measured using the microphone raw input signal. Without applying AEC, overall WWD accuracy is < 6% under such low SER condition. With playback disabled, WWD rate could be above 90% in quiet meeting room. Hence AEC is necessary in order to achieve acceptable wakeup word detection accuracy rate.

4.1. ERLE test

To reduce the potential biased error due to initial convergence, last 2 seconds of raw mic and SAEC output recordings were taken for steady state ERLE measure.

We also artificially added 5 most representative noise samples (carefully selected from noiseX-92 database, they are speech babble, destroyer engine, factory2, volvo and white noise) to the recordings to cross check how robustness of proposed solution is. We mixed the recordings with two SNRs, 6dB and 20dB. Same last 2 seconds of raw and SAEC output recordings were used for measuring the ERLE.

Table 2 describes the overall ERLE improvements under quiet environment and artificially noisy conditions with SNR at 6 and 20dB, consistent improvement is observed.

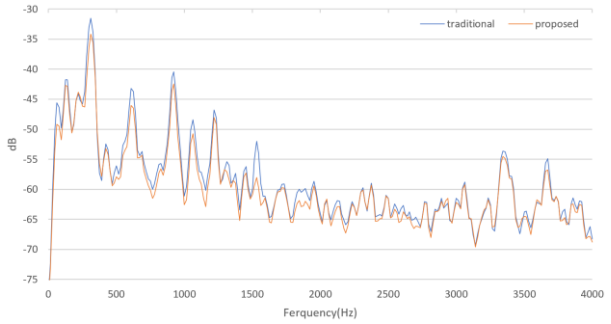

Figure 2: Residue *echo power spectrum of traditional (blue line) and proposed (red line) method（only 0~4kHz is shown）.*

Although overall ERLE improvement by the proposed solution is not huge, however we found proposed solution often offers noticeable improvement under two mentioned scenarios in below.

1. Content quickly changes from vocal to music and vice versus
2. Content level abruptly increases/decreases

As we can see in Figure 2, residual echo was lower in proposed solution than in traditional solution. The residual echo level was measured right after the content changes from music to vocal.

Figure 3 in below illustrates the improvement seen by scenario 2. During 0.5~1.5 sec and 3.1~4 sec, near end workup word is present, and at the steady state, ERLE improvement is around 1.3dB comparing with that done by traditional method, but the proposed solution offers > 5dB of ERLE improvement than in traditional solution at around 1.9 sec. where the playback song is transiting from verse to chorus and many more music instruments join in performance, which abruptly changes the echo level.

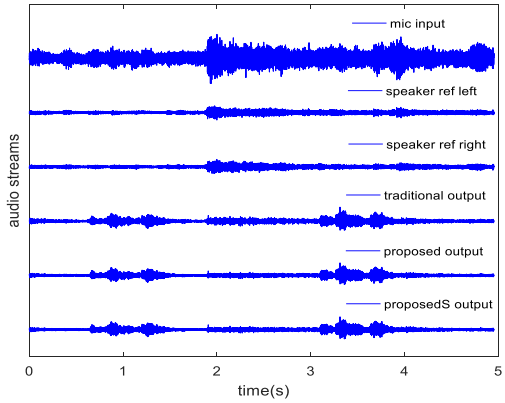

Figure 3: Residual echo level measure with music echo sharp burst.

4.2. Wakeup Word Detection

To reuse the wakeup model in Tencent Dingdang smart speaker with display, the wakeup word including 2 Chinese characters (each one is repeated once) is employed in this work, with their Chinese pinyin representation as “ding1 dang1 ding1 dang1”.

The baseline WWD system employed in this work consists of convolutional, LSTM and fully connected layers. From bottom to top, they are one convolutional layer with max pooling[15], two LSTM layers with 256 hidden units, each LSTM layer is followed by a batch normalization operation, one fully connected layers with 128 hidden units and a softmax

Table 3: *Wakeup word detection rate under different noise conditions (%)*

	no noise	babble		destroyerengine		facttory2		volvo		white		Average
		6dB	20dB	6dB	20dB	6dB	20dB	6dB	20dB	6dB	20dB	
Traditional	87.89	72.68	89.07	70.07	88.12	64.85	87.89	64.13	87.89	48.93	87.65	77.20
Proposed	90.26	75.30	90.97	72.45	89.07	67.46	88.60	66.27	88.84	55.11	90.02	79.49
ProposedS	91.21	75.53	89.79	71.26	90.26	66.98	88.84	65.08	88.84	55.82	89.31	79.36
Proposed Δ	2.38	2.61	1.90	2.38	0.95	2.61	0.71	2.14	0.95	6.18	2.38	2.29
ProposedS Δ	3.33	2.85	0.71	1.19	2.14	2.14	0.95	0.95	0.95	6.89	1.66	2.16

layer. 40 dimensional log-mel filterbank features with its delta and delta-delta appended are computed every 25ms with a 10ms frame shift. At each frame, we stack 10 frames to the left and 5 frames to the right as the input feature to the convolutional layer.

The WWD model is pre-trained on a 100k-hour Chinese ASR multi-condition training set that contains both clean and far-field noisy data. The output layer has 3 output units representing the 2 Chinese characters of the wakeup word and one non-wakeup-word filler. A wakeup word specific data set of more than 1000 hours from more than 10 thousand human speakers, is used as positive examples. So the model is well adapted to a wide range of smart box users. Please refer to [16] for more details about network and training settings and to [17] for more details about posterior handling

Also to reduce the potential biased error due to initial convergence, all SAEC filters start from convergent state by initializing with filter coefficients dumped from steady state.

From table 3, WWD test result also shows consistent positive observation that proposed solution has improved WWD rate over traditional solution by > 2% on average.

It is also shown that although the simplified ProposedS algorithm has lower ERLE improvement(average Δ is 0.96dB vs. 0.64dB) as shown in Table 2, it has comparable improvement in terms of WWD rate with full-process model(average Δ is 2.29% vs. 2.16%), and even higher in no noise cases(2.38% vs. 3.33%). That means with the ProposedS algorithm, computation is saved without much sacrificing the WWD performance.

5. Conclusions

To improve the performance of the traditional NLMS based SAEC system, we propose a method to independently update the adaptive filters by mutually pre-whitening the residue error signals. Consistent ERLE improvement is observed on the recordings from commercial smart speak product. Because of its improved stability especially at the abrupt content changes in echo signal, wakeup word detection rate is also reasonably improved. Computationally simplified version of the proposed method also shows comparable improvement in WWD task.

6. References

[1] M. M. Sondhi, D. R. Morgan and J. L. Hall, "*Stereophonic acoustic echo cancellation-an overview of the fundamental problem,*" in IEEE Signal Processing Letters, vol. 2, no. 8, pp. 148-151, Aug. 1995.

[2] S. Makino, *"Stereophonic acoustic echo cancellation: An overview and recent solution,"* in Proc. The 1999 IEEE Workshop on Acoustic Echo and Noise Control, 1999, pp. 12–19.

[3] A. W. H. Khong, J. Benesty, and P. A. Naylor, *"Stereophonic acoustic echo cancellation: Analysis of the misalignment in the frequency domain,"* IEEE Signal Process. Lett., vol. 13, no. 1, pp. 33–36, Jan. 2006

[4] C. Stanciu, C. Paleologu, J. Benesty, S. Ciochina and F. Albu, "*Variable-forgetting factor RLS for stereophonic acoustic echo cancellation with widely linear model,*" Proceedings of the 20th European Signal Processing Conference(EUSIPCO), Bucharest, 2012, pp. 1960-1964.

[5] D. R. Morgan, J. L. Hall and J. Benesty, "*Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation,*" in IEEE Transactions on Speech and Audio Processing, vol. 9, no. 6, pp. 686-696, Sept. 2001.

[6] M. Ali, "*Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation,*" Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98, 1998, pp. 3689-3692 vol.6.

[7] S. Emura, Y. Haneda, A. Kataoka, and S. Makino, *"Stereo echo cancellation algorithm using adaptive update on the basis of enhanced inputsignal vector,"* Signal Process., vol. 86, pp. 1157–1167, Jun. 2006.

[8] Q. Lei, H. Chen, J. Hou, L. Chen, L. Dai. "*Deep Neural Network Based Regression Approach for Acoustic Echo Cancellation.*" Proc. 4th International Conference on Multimedia Systems and Signal Processing, 2019, 94-98

[9] M. Bekrani, A. W. H. Khong and M. Lotfizad, *"A Linear Neural Network-Based Approach to Stereophonic Acoustic Echo Cancellation,"* in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 6, pp. 1743-1753, Aug. 2011

[10] J.-S. Soo and K. Pang, "*Multidelay block frequency domain adaptive filter,*" IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 38, no. 2, pp. 373–376, 1990.

[11] H. Buchner, J. Benesty, W. Kellermann, "*An Extended Multidelay Filter: Fast Low-Delay Algorithms for Very High-Order Adaptive Systems*," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003.

[12] JM P Borrallo, M G Otero, "*On the implementation of a partitioned block frequency domain adaptive filter (PBFDAF) for long acoustic echo cancellation*", Signal Processing,Vol. 27, Issue 3, pp. 301-315, 1992

[13] H. O. Simon, *Sec. 8.2, Adaptive Filter Theory, 5th Edition,* Pearson, 2014

[14] J. Valin, "*On Adjusting the Learning Rate in Frequency Domain Echo Cancellation with Double-Talk,*" in IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 3, pp. 1030-1034, March 2007.

[15] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "*Convolutional neural networks for speech recognition*," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 10, pp. 1533–1545, 2014.

[16] M. Yu, X. Ji, Y. Gao, L. Chen, J. Chen, J. Zheng, D. Su, D. Yu, *"Text-Dependent Speech Enhancement for Small-Footprint Robust Keyword Detection."* 2613-2617. 10.21437/Interspeech.2018-1668.

[17] G. Chen, C. Parada, and G. Heigold, "*Small-footprint keyword spotting using deep neural networks,*" in the Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 4087–4091.