Universidade Federal de São João Del Rei, Campus Alto Paraopeba

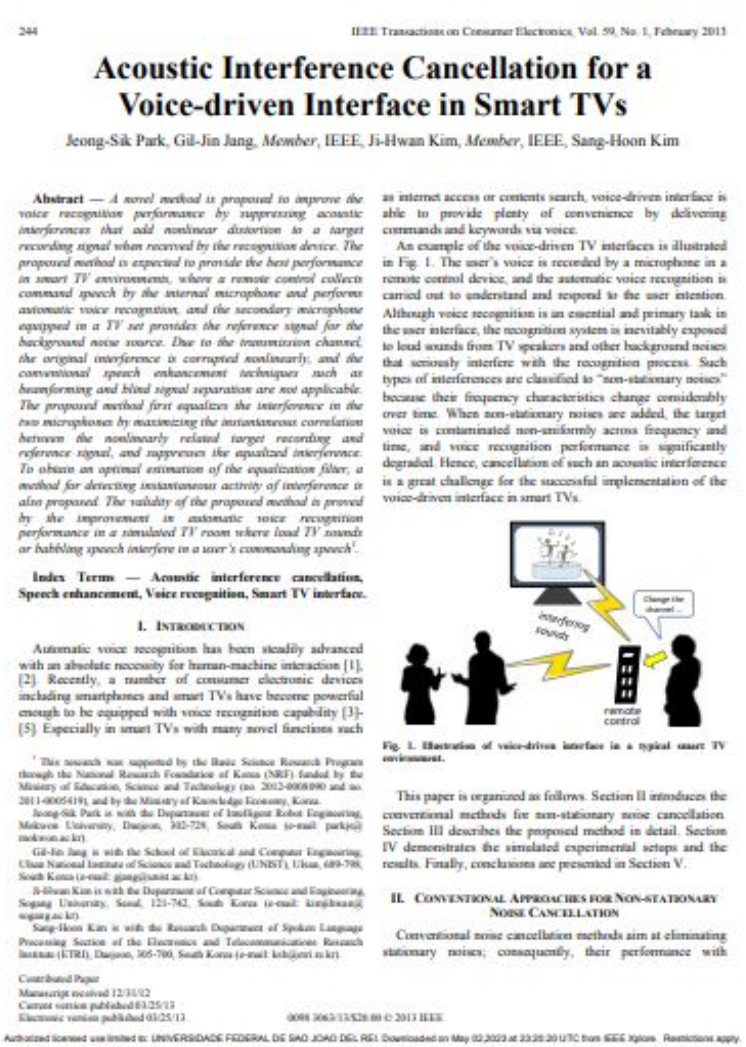
Apresentação do Artigo:

“Acoustic Interference Cancellation for a Voice-driven Interface in Smart TVs”

Cancelamento de Interferência Acústica para uma Interface Controlada por Voz em Smart TVs.

Autores: **Jeong-Sik Park, Gil-Jin Jang, Member, IEEE, Ji-Hwan Kim, Member, IEEE, Sang-Hoon Kim**

Disciplina: **Processamento Digital de Sinais.**
Docente: **Gustavo Fernandes Rodrigues.**
Discente: **Luciano Stork**



Acoustic Interference Cancellation for a Voice-driven Interface in Smart TVs

Jeong-Sik Park, Gil-Jin Jang, *Member*, IEEE, Ji-Hwan Kim, *Member*, IEEE, Sang-Hoon Kim

Abstract — *A novel method is proposed to improve the voice recognition performance by suppressing acoustic interferences that add nonlinear distortion to a target recording signal when received by the recognition device. The proposed method is expected to provide the best performance in smart TV environments, where a remote control collects command speech by the internal microphone and performs automatic voice recognition, and the secondary microphone equipped in a TV set provides the reference signal for the background noise source. Due to the transmission channel, the original interference is corrupted nonlinearly, and the conventional speech enhancement techniques such as beamforming and blind signal separation are not applicable. The proposed method first equalizes the interference in the two microphones by maximizing the instantaneous correlation between the nonlinearly related target recording and reference signal, and suppresses the equalized interference. To obtain an optimal estimation of the equalization filter, a method for detecting instantaneous activity of interference is also proposed. The validity of the proposed method is proved by the improvement in automatic voice recognition performance in a simulated TV room where loud TV sounds or babbling speech interfere in a user's commanding speech*[1].

Index Terms — Acoustic interference cancellation, Speech enhancement, Voice recognition, Smart TV interface.

I. Introduction

Automatic voice recognition has been steadily advanced with an absolute necessity for human-machine interaction [1], [2]. Recently, a number of consumer electronic devices including smartphones and smart TVs have become powerful enough to be equipped with voice recognition capability [3]-[5]. Especially in smart TVs with many novel functions such as internet access or contents search, voice-driven interface is able to provide plenty of convenience by delivering commands and keywords via voice.

An example of the voice-driven TV interfaces is illustrated in Fig. 1. The user's voice is recorded by a microphone in a remote control device, and the automatic voice recognition is carried out to understand and respond to the user intention. Although voice recognition is an essential and primary task in the user interface, the recognition system is inevitably exposed to loud sounds from TV speakers and other background noises that seriously interfere with the recognition process. Such types of interferences are classified to "non-stationary noises" because their frequency characteristics change considerably over time. When non-stationary noises are added, the target voice is contaminated non-uniformly across frequency and time, and voice recognition performance is significantly degraded. Hence, cancellation of such an acoustic interference is a great challenge for the successful implementation of the voice-driven interface in smart TVs.

Fig. 1. Illustration of voice-driven interface in a typical smart TV environment.

This paper is organized as follows. Section II introduces the conventional methods for non-stationary noise cancellation. Section III describes the proposed method in detail. Section IV demonstrates the simulated experimental setups and the results. Finally, conclusions are presented in Section V.

II. Conventional Approaches for Non-stationary Noise Cancellation

Conventional noise cancellation methods aim at eliminating stationary noises; consequently, their performance with

[1] This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (no. 2012-0008090 and no. 2011-0005419), and by the Ministry of Knowledge Economy, Korea.

Jeong-Sik Park is with the Department of Intelligent Robot Engineering, Mokwon University, Daejeon, 302-729, South Korea (e-mail: parkjs@mokwon.ac.kr).

Gil-Jin Jang is with the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 689-798, South Korea (e-mail: gjang@unist.ac.kr).

Ji-Hwan Kim is with the Department of Computer Science and Engineering, Sogang University, Seoul, 121-742, South Korea (e-mail: kimjihwan@sogang.ac.kr).

Sang-Hoon Kim is with the Research Department of Spoken Language Processing Section of the Electronics and Telecommunications Research Institute (ETRI), Daejeon, 305-700, South Korea (e-mail: ksh@etri.re.kr).

Contributed Paper
Manuscript received 12/31/12
Current version published 03/25/13
Electronic version published 03/25/13.

RESUMO:

- Proposta de um novo método para melhorar o desempenho de reconhecimento de voz em ambientes de smart TV.

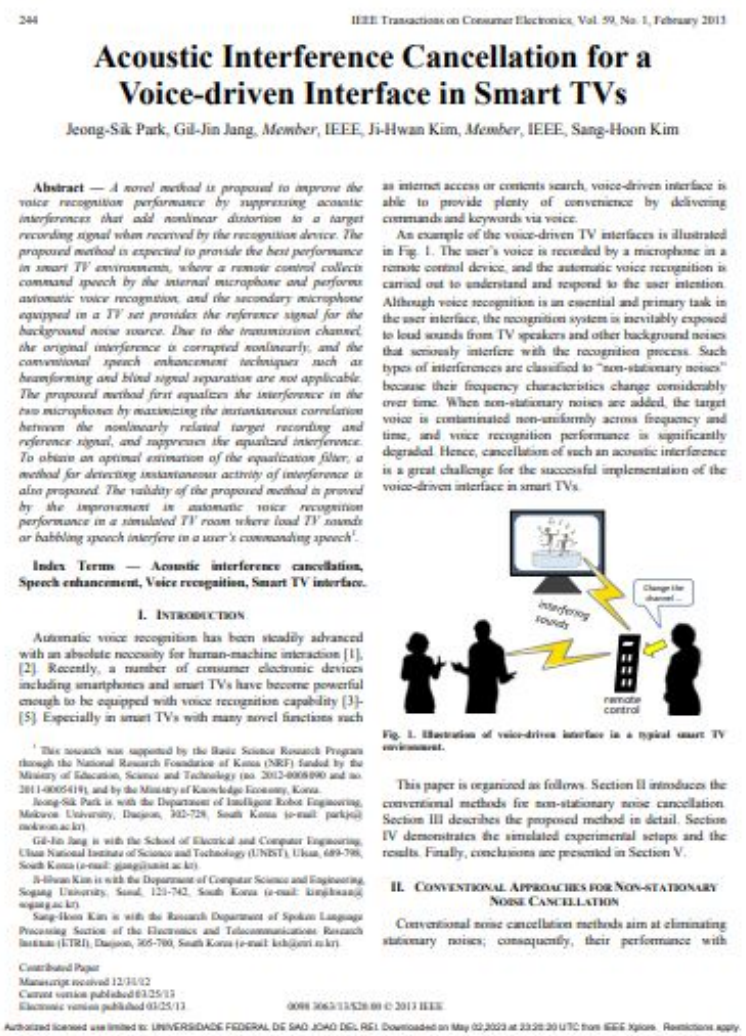
- O método visa suprimir interferências acústicas que causam distorção não linear no sinal de gravação recebido pelo dispositivo de reconhecimento.

- As técnicas convencionais de aprimoramento de fala, como beamforming (direcionamento) e separação cega de sinais (distinção), não são aplicáveis devido à corrupção não linear da interferência pelo canal de transmissão.

- O método proposto equaliza a interferência em dois microfones, maximizando a correlação instantânea entre o sinal de gravação alvo e o sinal de referência.

- Detecção da atividade instantânea da interferência para obter uma estimativa ótima do filtro de equalização.

- A validade do método é comprovada pela melhoria no desempenho de reconhecimento automático de voz em um ambiente simulado de uma sala de TV com interferências de sons altos da TV ou fala confusa.



244 IEEE Transactions on Consumer Electronics, Vol. 59, No. 1, February 2013

Acoustic Interference Cancellation for a Voice-driven Interface in Smart TVs

Jeong-Sik Park, Gil-Jin Jang, *Member*, IEEE, Ji-Hwan Kim, *Member*, IEEE, Sang-Hoon Kim

Abstract — *A novel method is proposed to improve the voice recognition performance by suppressing acoustic interferences that add nonlinear distortion to a target recording signal when received by the recognition device. The proposed method is expected to provide the best performance in smart TV environments, where a remote control collects command speech by the internal microphone and performs automatic voice recognition, and the secondary microphone equipped in a TV set provides the reference signal for the background noise source. Due to the transmission channel, the original interference is corrupted nonlinearly, and the conventional speech enhancement techniques such as beamforming and blind signal separation are not applicable. The proposed method first equalizes the interference in the two microphones by maximizing the instantaneous correlation between the nonlinearly related target recording and reference signal, and suppresses the equalized interference. To obtain an optimal estimation of the equalization filter, a method for detecting instantaneous activity of interference is also proposed. The validity of the proposed method is proved by the improvement in automatic voice recognition performance in a simulated TV room where loud TV sounds or babbling speech interfere in a user's commanding speech*[1].

Index Terms — Acoustic interference cancellation, Speech enhancement, Voice recognition, Smart TV interface.

I. Introduction

Automatic voice recognition has been steadily advanced with an absolute necessity for human-machine interaction [1], [2]. Recently, a number of consumer electronic devices including smartphones and smart TVs have become powerful enough to be equipped with voice recognition capability [3]-[5]. Especially in smart TVs with many novel functions such as internet access or contents search, voice-driven interface is able to provide plenty of convenience by delivering commands and keywords via voice.

An example of the voice-driven TV interfaces is illustrated in Fig. 1. The user's voice is recorded by a microphone in a remote control device, and the automatic voice recognition is carried out to understand and respond to the user intention. Although voice recognition is an essential and primary task in the user interface, the recognition system is inevitably exposed to loud sounds from TV speakers and other background noises that seriously interfere with the recognition process. Such types of interferences are classified to "non-stationary noises" because their frequency characteristics change considerably over time. When non-stationary noises are added, the target voice is contaminated non-uniformly across frequency and time, and voice recognition performance is significantly degraded. Hence, cancellation of such an acoustic interference is a great challenge for the successful implementation of the voice-driven interface in smart TVs.

Fig. 1. Illustration of voice-driven interface in a typical smart TV environment.

This paper is organized as follows. Section II introduces the conventional methods for non-stationary noise cancellation. Section III describes the proposed method in detail. Section IV demonstrates the simulated experimental setups and the results. Finally, conclusions are presented in Section V.

II. Conventional Approaches for Non-stationary Noise Cancellation

Conventional noise cancellation methods aim at eliminating stationary noises; consequently, their performance with

[1] This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (no. 2012-0008090 and no. 2011-0005419), and by the Ministry of Knowledge Economy, Korea.

Jeong-Sik Park is with the Department of Intelligent Robot Engineering, Mokwon University, Daejeon, 302-729, South Korea (e-mail: parkjs@mokwon.ac.kr).

Gil-Jin Jang is with the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, 689-798, South Korea (e-mail: gjang@unist.ac.kr).

Ji-Hwan Kim is with the Department of Computer Science and Engineering, Sogang University, Seoul, 121-742, South Korea (e-mail: kimjihwan@sogang.ac.kr).

Sang-Hoon Kim is with the Research Department of Spoken Language Processing Section of the Electronics and Telecommunications Research Institute (ETRI), Daejeon, 305-700, South Korea (e-mail: ksh@etri.re.kr).

Contributed Paper
Manuscript received 12/31/12
Current version published 03/25/13
Electronic version published 03/25/13

I - INTRODUÇÃO:

- O reconhecimento automático de voz é fundamental para a interação entre humanos e máquinas.

- Dispositivos eletrônicos, como smartphones e smart TVs, estão equipados com capacidade de reconhecimento de voz.

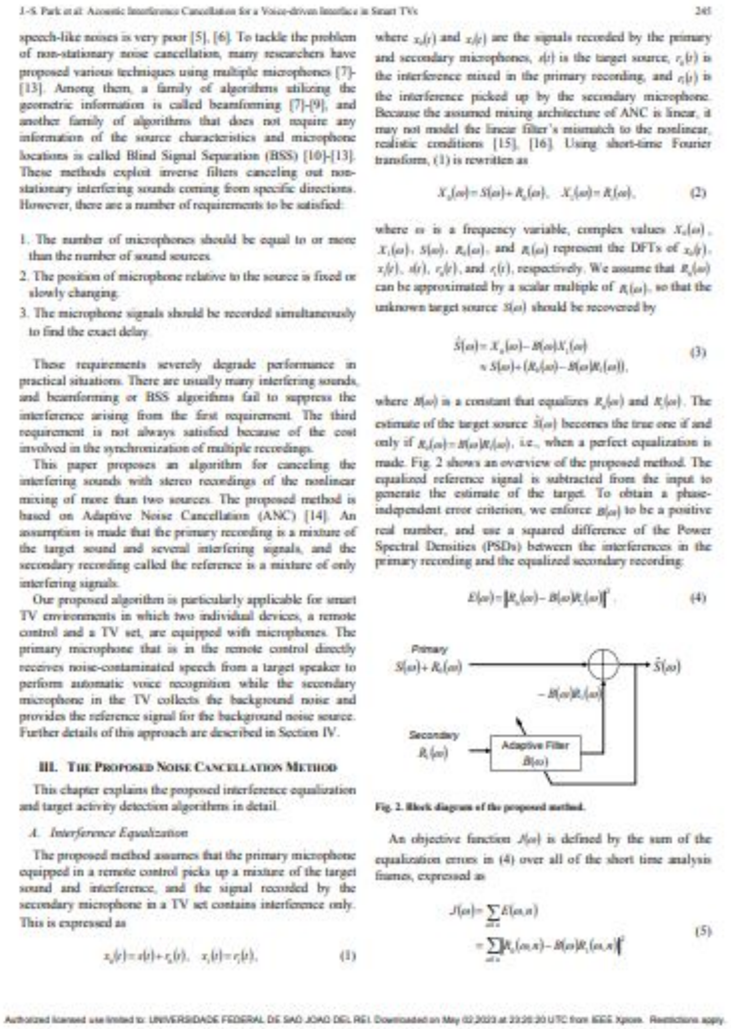
- A interface controlada por voz em smart TVs oferece conveniência ao permitir comandos e palavras-chave por meio da voz. ("Prompt por voz" - Acessibilidade).

- O sistema de reconhecimento de voz está exposto a ruídos não estacionários, como sons altos da TV e outros ruídos de fundo, que interferem no processo de reconhecimento.

- Essas interferências afetam a voz-alvo de maneira não uniforme em termos de frequência e tempo, degradando o desempenho do reconhecimento de voz. (Impossível prever).

- O cancelamento dessas interferências acústicas é um desafio para a implementação bem-sucedida da interface controlada por voz em smart TVs.

II - ABORDAGENS CONVENCIONAIS PARA CANCELAMENTO DE RUÍDO NÃO ESTACIONÁRIO:



J.-S. Park et al: Acoustic Interference Cancellation for a Voice-driven Interface in Smart TVs 245

speech-like noises is very poor [5], [6]. To tackle the problem of non-stationary noise cancellation, many researchers have proposed various techniques using multiple microphones [7]-[13]. Among them, a family of algorithms utilizing the geometric information is called beamforming [7]-[9], and another family of algorithms that does not require any information of the source characteristics and microphone locations is called Blind Signal Separation (BSS) [10]-[13]. These methods exploit inverse filters canceling out non-stationary interfering sounds coming from specific directions. However, there are a number of requirements to be satisfied:

1. The number of microphones should be equal to or more than the number of sound sources.
2. The position of microphone relative to the source is fixed or slowly changing.
3. The microphone signals should be recorded simultaneously to find the exact delay.

These requirements severely degrade performance in practical situations. There are usually many interfering sounds, and beamforming or BSS algorithms fail to suppress the interference arising from the first requirement. The third requirement is not always satisfied because of the cost involved in the synchronization of multiple recordings.

This paper proposes an algorithm for canceling the interfering sounds with stereo recordings of the nonlinear mixing of more than two sources. The proposed method is based on Adaptive Noise Cancellation (ANC) [14]. An assumption is made that the primary recording is a mixture of the target sound and several interfering signals, and the secondary recording called the reference is a mixture of only interfering signals.

Our proposed algorithm is particularly applicable for smart TV environments in which two individual devices, a remote control and a TV set, are equipped with microphones. The primary microphone that is in the remote control directly receives noise-contaminated speech from a target speaker to perform automatic voice recognition while the secondary microphone in the TV collects the background noise and provides the reference signal for the background noise source. Further details of this approach are described in Section IV.

III. The Proposed Noise Cancellation Method

This chapter explains the proposed interference equalization and target activity detection algorithms in detail.

A. Interference Equalization

The proposed method assumes that the primary microphone equipped in a remote control picks up a mixture of the target sound and interference, and the signal recorded by the secondary microphone in a TV set contains interference only. This is expressed as

$$x_0(t)=s(t)+r_0(t), \quad x_1(t)=r_1(t), \tag{1}$$

where $x_0(t)$ and $x_1(t)$ are the signals recorded by the primary and secondary microphones, $s(t)$ is the target source, $r_0(t)$ is the interference mixed in the primary recording, and $r_1(t)$ is the interference picked up by the secondary microphone. Because the assumed mixing architecture of ANC is linear, it may not model the linear filter's mismatch to the nonlinear, realistic conditions [15], [16]. Using short-time Fourier transform, (1) is rewritten as

$$X_0(\omega)=S(\omega)+R_0(\omega), \quad X_1(\omega)=R_1(\omega), \tag{2}$$

where ω is a frequency variable, complex values $X_0(\omega)$, $X_1(\omega)$, $S(\omega)$, $R_0(\omega)$, and $R_1(\omega)$ represent the DFTs of $x_0(t)$, $x_1(t)$, $s(t)$, $r_0(t)$, and $r_1(t)$, respectively. We assume that $R_0(\omega)$ can be approximated by a scalar multiple of $R_1(\omega)$, so that the unknown target source $S(\omega)$ should be recovered by

$$\begin{aligned}\hat{S}(\omega)&=X_0(\omega)-B(\omega)X_1(\omega)\\&\approx S(\omega)+(R_0(\omega)-B(\omega)R_1(\omega)),\end{aligned} \tag{3}$$

where $B(\omega)$ is a constant that equalizes $R_0(\omega)$ and $R_1(\omega)$. The estimate of the target source $\hat{S}(\omega)$ becomes the true one if and only if $R_0(\omega)=B(\omega)R_1(\omega)$, i.e., when a perfect equalization is made. Fig. 2 shows an overview of the proposed method. The equalized reference signal is subtracted from the input to generate the estimate of the target. To obtain a phase-independent error criterion, we enforce $B(\omega)$ to be a positive real number, and use a squared difference of the Power Spectral Densities (PSDs) between the interferences in the primary recording and the equalized secondary recording:

$$E(\omega)=\|R_0(\omega)-B(\omega)R_1(\omega)\|^2. \tag{4}$$

Fig. 2. Block diagram of the proposed method.

An objective function $J(\omega)$ is defined by the sum of the equalization errors in (4) over all of the short time analysis frames, expressed as

$$\begin{aligned}J(\omega)&=\sum_{\text{all } n}E(\omega,n)\\&=\sum_{\text{all } n}\|R_0(\omega,n)-B(\omega)R_1(\omega,n)\|^2\end{aligned} \tag{5}$$

- Os métodos convencionais de cancelamento de ruído não são eficientes para ruídos não estacionários, como ruídos semelhantes à fala.

- Diversas técnicas foram propostas para lidar com o cancelamento de ruído não estacionário usando múltiplos microfones.

- Beamforming e Blind Signal Separation (BSS) são duas famílias de algoritmos utilizados, mas possuem requisitos que limitam sua aplicabilidade prática.

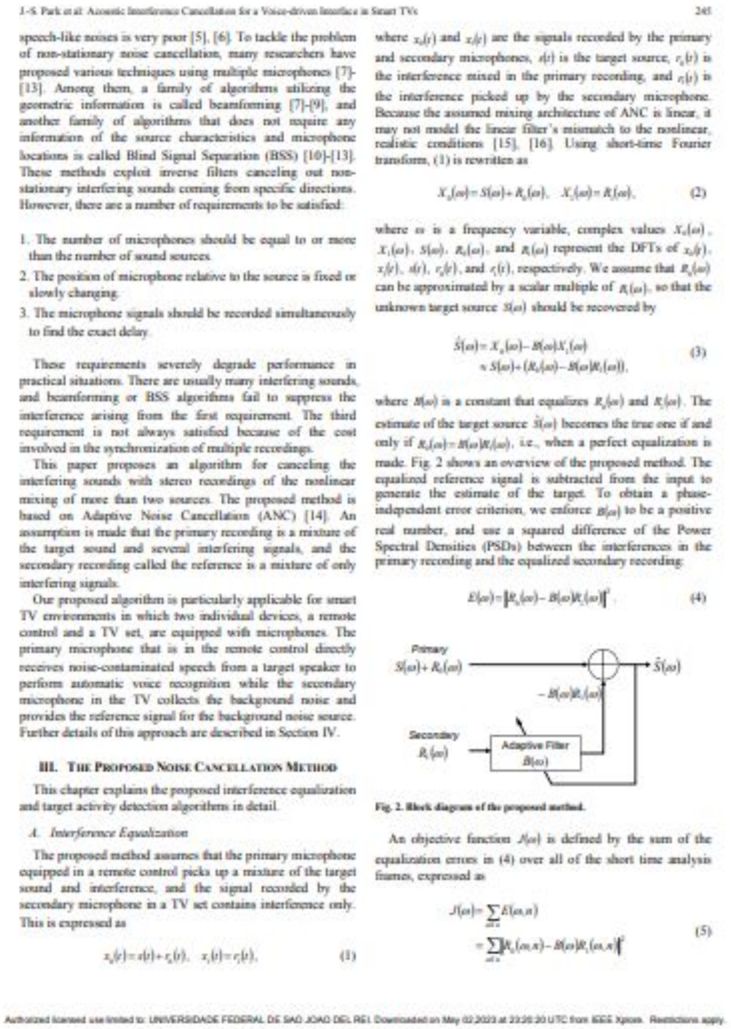
    1. O número de microfones deve ser igual ou superior ao número de fontes sonoras.
    2. A posição do microfone em relação à fonte é fixa ou muda lentamente.
    3. Os sinais dos microfones devem ser gravados simultaneamente para encontrar o atraso exato.

- Este artigo propõe um algoritmo baseado no Cancelamento Adaptativo de Ruído (Adaptive Noise Cancellation) para cancelar os sons interferentes em gravações estéreo de misturas não lineares de fontes sonoras.

- O algoritmo é aplicável a ambientes de smart TV, onde o controle remoto e a TV possuem microfones.

- O microfone do controle remoto recebe a fala contaminada por ruído para o reconhecimento automático de voz, enquanto o microfone da TV coleta o ruído de fundo e fornece o sinal de referência.

III - O MÉTODO PROPOSTO DE CANCELAMENTO DE RUÍDO:



J.-S. Park et al: Acoustic Interference Cancellation for a Voice-driven Interface in Smart TVs 245

speech-like noises is very poor [5], [6]. To tackle the problem of non-stationary noise cancellation, many researchers have proposed various techniques using multiple microphones [7]-[13]. Among them, a family of algorithms utilizing the geometric information is called beamforming [7]-[9], and another family of algorithms that does not require any information of the source characteristics and microphone locations is called Blind Signal Separation (BSS) [10]-[13]. These methods exploit inverse filters canceling out non-stationary interfering sounds coming from specific directions. However, there are a number of requirements to be satisfied:

1. The number of microphones should be equal to or more than the number of sound sources.
2. The position of microphone relative to the source is fixed or slowly changing.
3. The microphone signals should be recorded simultaneously to find the exact delay.

These requirements severely degrade performance in practical situations. There are usually many interfering sounds, and beamforming or BSS algorithms fail to suppress the interference arising from the first requirement. The third requirement is not always satisfied because of the cost involved in the synchronization of multiple recordings.

This paper proposes an algorithm for canceling the interfering sounds with stereo recordings of the nonlinear mixing of more than two sources. The proposed method is based on Adaptive Noise Cancellation (ANC) [14]. An assumption is made that the primary recording is a mixture of the target sound and several interfering signals, and the secondary recording called the reference is a mixture of only interfering signals.

Our proposed algorithm is particularly applicable for smart TV environments in which two individual devices, a remote control and a TV set, are equipped with microphones. The primary microphone that is in the remote control directly receives noise-contaminated speech from a target speaker to perform automatic voice recognition while the secondary microphone in the TV collects the background noise and provides the reference signal for the background noise source. Further details of this approach are described in Section IV.

III. The Proposed Noise Cancellation Method

This chapter explains the proposed interference equalization and target activity detection algorithms in detail.

A. Interference Equalization

The proposed method assumes that the primary microphone equipped in a remote control picks up a mixture of the target sound and interference, and the signal recorded by the secondary microphone in a TV set contains interference only. This is expressed as

$$x_0(t)=s(t)+r_0(t),\quad x_1(t)=r_1(t), \tag{1}$$

where $x_0(t)$ and $x_1(t)$ are the signals recorded by the primary and secondary microphones, $s(t)$ is the target source, $r_0(t)$ is the interference mixed in the primary recording, and $r_1(t)$ is the interference picked up by the secondary microphone. Because the assumed mixing architecture of ANC is linear, it may not model the linear filter's mismatch to the nonlinear, realistic conditions [15], [16]. Using short-time Fourier transform, (1) is rewritten as

$$X_0(\omega)=S(\omega)+R_0(\omega),\quad X_1(\omega)=R_1(\omega), \tag{2}$$

where ω is a frequency variable, complex values $X_0(\omega)$, $X_1(\omega)$, $S(\omega)$, $R_0(\omega)$, and $R_1(\omega)$ represent the DFTs of $x_0(t)$, $x_1(t)$, $s(t)$, $r_0(t)$, and $r_1(t)$, respectively. We assume that $R_0(\omega)$ can be approximated by a scalar multiple of $R_1(\omega)$, so that the unknown target source $S(\omega)$ should be recovered by

$$\begin{aligned}\hat{S}(\omega)&=X_0(\omega)-B(\omega)X_1(\omega)\\&\approx S(\omega)+(R_0(\omega)-B(\omega)R_1(\omega)),\end{aligned} \tag{3}$$

where $B(\omega)$ is a constant that equalizes $R_0(\omega)$ and $R_1(\omega)$. The estimate of the target source $\hat{S}(\omega)$ becomes the true one if and only if $R_0(\omega)=B(\omega)R_1(\omega)$, i.e., when a perfect equalization is made. Fig. 2 shows an overview of the proposed method. The equalized reference signal is subtracted from the input to generate the estimate of the target. To obtain a phase-independent error criterion, we enforce $B(\omega)$ to be a positive real number, and use a squared difference of the Power Spectral Densities (PSDs) between the interferences in the primary recording and the equalized secondary recording:

$$E(\omega)=\left\|R_0(\omega)-B(\omega)R_1(\omega)\right\|^2. \tag{4}$$

Fig. 2. Block diagram of the proposed method.

An objective function $J(\omega)$ is defined by the sum of the equalization errors in (4) over all of the short time analysis frames, expressed as

$$\begin{aligned}J(\omega)&=\sum_{\text{all } n}E(\omega,n)\\&=\sum_{\text{all } n}\left\|R_0(\omega,n)-B(\omega)R_1(\omega,n)\right\|^2\end{aligned} \tag{5}$$

- Apresentação do método proposto com base na equalização da interferência e detecção de atividade do sinal-alvo.

- A equalização de interferência é baseada na suposição de que o microfone primário captura uma mistura do som-alvo e interferência, enquanto o microfone secundário registra apenas a interferência.

$$x0(t) = s(t) + r0(t)$$
$$x1(t) = r1(t),$$

onde x0(t) e x1(t) são os sinais gravados pelos microfones primário e secundário, s(t) é a fonte-alvo, r0(t) é a interferência misturada na gravação primária e r1(t) é a interferência capturada pelo microfone secundário

- A partir de então, realiza-se a Transformada de Fourier com curta duração, obtendo os sinais no domínio da frequência.

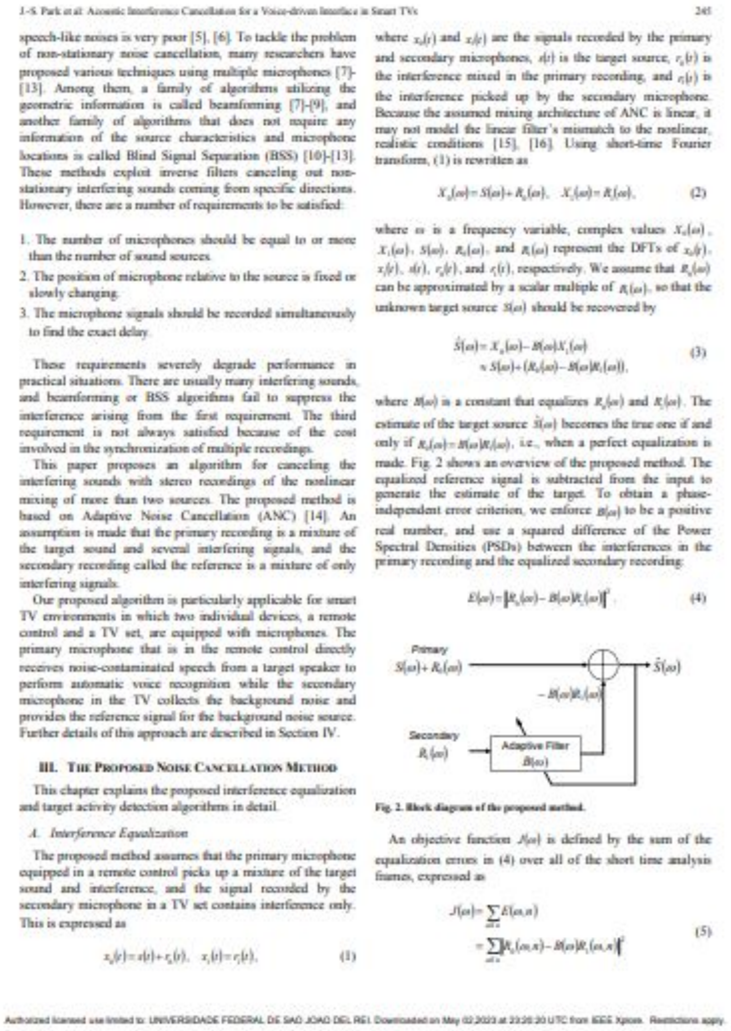
$$X0(w) = S(w) + R0(w)\ ;$$
$$X1(w) = R1(w).$$

- Assumimos que R0 pode ser aproximado por um múltiplo escalar de R1, de modo que a fonte-alvo desconhecida S deve ser recuperada pela equação

$$S'(w) = X0(w) - B(w)\ X1(w)$$
$$S'(w) = S(w)+[R0(w)-B(w)-R1(w)]$$

- A estimativa da fonte-alvo S'(w) se torna a verdadeira apenas se R0(w) = B(w)*R1(w), ou seja, quando uma equalização perfeita é feita.

III - O MÉTODO PROPOSTO DE CANCELAMENTO DE RUÍDO:



J.-S. Park et al: Acoustic Interference Cancellation for a Voice-driven Interface in Smart TVs 245

speech-like noises is very poor [5], [6]. To tackle the problem of non-stationary noise cancellation, many researchers have proposed various techniques using multiple microphones [7]-[13]. Among them, a family of algorithms utilizing the geometric information is called beamforming [7]-[9], and another family of algorithms that does not require any information of the source characteristics and microphone locations is called Blind Signal Separation (BSS) [10]-[13]. These methods exploit inverse filters canceling out non-stationary interfering sounds coming from specific directions. However, there are a number of requirements to be satisfied:

1. The number of microphones should be equal to or more than the number of sound sources.
2. The position of microphone relative to the source is fixed or slowly changing.
3. The microphone signals should be recorded simultaneously to find the exact delay.

These requirements severely degrade performance in practical situations. There are usually many interfering sounds, and beamforming or BSS algorithms fail to suppress the interference arising from the first requirement. The third requirement is not always satisfied because of the cost involved in the synchronization of multiple recordings.

This paper proposes an algorithm for canceling the interfering sounds with stereo recordings of the nonlinear mixing of more than two sources. The proposed method is based on Adaptive Noise Cancellation (ANC) [14]. An assumption is made that the primary recording is a mixture of the target sound and several interfering signals, and the secondary recording called the reference is a mixture of only interfering signals.

Our proposed algorithm is particularly applicable for smart TV environments in which two individual devices, a remote control and a TV set, are equipped with microphones. The primary microphone that is in the remote control directly receives noise-contaminated speech from a target speaker to perform automatic voice recognition while the secondary microphone in the TV collects the background noise and provides the reference signal for the background noise source. Further details of this approach are described in Section IV.

III. The Proposed Noise Cancellation Method

This chapter explains the proposed interference equalization and target activity detection algorithms in detail.

A. Interference Equalization

The proposed method assumes that the primary microphone equipped in a remote control picks up a mixture of the target sound and interference, and the signal recorded by the secondary microphone in a TV set contains interference only. This is expressed as

$$x_0(t)=s(t)+r_0(t),\quad x_1(t)=r_1(t), \qquad (1)$$

where $x_0(t)$ and $x_1(t)$ are the signals recorded by the primary and secondary microphones, $s(t)$ is the target source, $r_0(t)$ is the interference mixed in the primary recording, and $r_1(t)$ is the interference picked up by the secondary microphone. Because the assumed mixing architecture of ANC is linear, it may not model the linear filter's mismatch to the nonlinear, realistic conditions [15], [16]. Using short-time Fourier transform, (1) is rewritten as

$$X_0(\omega)=S(\omega)+R_0(\omega),\quad X_1(\omega)=R_1(\omega), \qquad (2)$$

where ω is a frequency variable, complex values $X_0(\omega)$, $X_1(\omega)$, $S(\omega)$, $R_0(\omega)$, and $R_1(\omega)$ represent the DFTs of $x_0(t)$, $x_1(t)$, $s(t)$, $r_0(t)$, and $r_1(t)$, respectively. We assume that $R_0(\omega)$ can be approximated by a scalar multiple of $R_1(\omega)$, so that the unknown target source $S(\omega)$ should be recovered by

$$\begin{aligned}\hat{S}(\omega)&=X_0(\omega)-B(\omega)X_1(\omega)\\&\approx S(\omega)+(R_0(\omega)-B(\omega)R_1(\omega)),\end{aligned} \qquad (3)$$

where $B(\omega)$ is a constant that equalizes $R_0(\omega)$ and $R_1(\omega)$. The estimate of the target source $\hat{S}(\omega)$ becomes the true one if and only if $R_0(\omega)=B(\omega)R_1(\omega)$, i.e., when a perfect equalization is made. Fig. 2 shows an overview of the proposed method. The equalized reference signal is subtracted from the input to generate the estimate of the target. To obtain a phase-independent error criterion, we enforce $B(\omega)$ to be a positive real number, and use a squared difference of the Power Spectral Densities (PSDs) between the interferences in the primary recording and the equalized secondary recording:

$$E(\omega)=\left\|R_0(\omega)\right|-B(\omega)\left|R_1(\omega)\right\|^2. \qquad (4)$$

Fig. 2. Block diagram of the proposed method.

An objective function $J(\omega)$ is defined by the sum of the equalization errors in (4) over all of the short time analysis frames, expressed as

$$\begin{aligned}J(\omega)&=\sum_{\text{all } n}E(\omega,n)\\&=\sum_{\text{all } n}\left\|R_0(\omega,n)\right|-B(\omega)\left|R_1(\omega,n)\right\|^2\end{aligned} \qquad (5)$$

- Para obter um critério de erro independente de fase, impomos que B seja um número real positivo e utilizamos a diferença ao quadrado das Densidades Espectrais de Potência (PSDs) entre as interferências na gravação primária e a gravação secundária equalizada.

PSDs = |R0(w)-B(w)*R1(w)|²

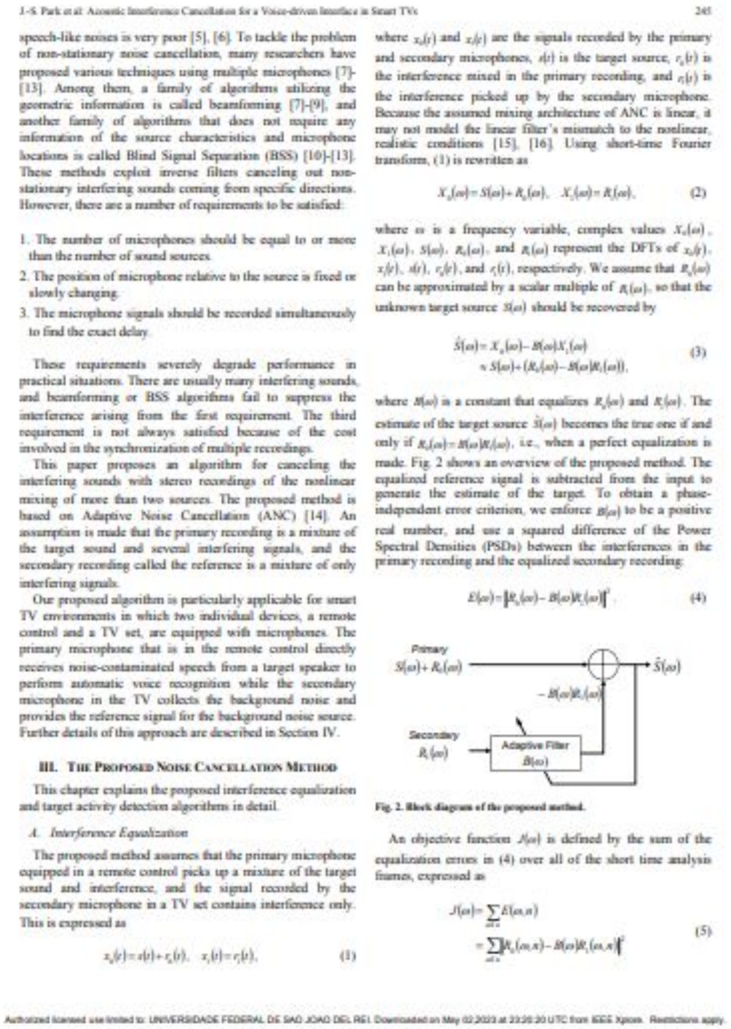
- Esse processo terá continuidade até que seja atingido uma boa aproximação das interferências R0(w) e R1(w), ou seja, quando tivermos um B(w) bem ajustado. Assim:

PSDs = 0

- A função J(w) é definida como a soma dos erros de equalização ao longo dos quadros de análise.

- A derivada de J(w) em relação a B(w) é calculada para encontrar um filtro de equalização ótimo. Igualar a zero e aproximar garantem que B(w) seja um valor positivo e real. Isso implica que apenas as PSDs dos dois canais são consideradas e a diferença de fase entre elas é ignorada.

III - O MÉTODO PROPOSTO DE CANCELAMENTO DE RUÍDO:



speech-like noises is very poor [5], [6]. To tackle the problem of non-stationary noise cancellation, many researchers have proposed various techniques using multiple microphones [7]-[13]. Among them, a family of algorithms utilizing the geometric information is called beamforming [7]-[9], and another family of algorithms that does not require any information of the source characteristics and microphone locations is called Blind Signal Separation (BSS) [10]-[13]. These methods exploit inverse filters canceling out non-stationary interfering sounds coming from specific directions. However, there are a number of requirements to be satisfied:

1. The number of microphones should be equal to or more than the number of sound sources.
2. The position of microphone relative to the source is fixed or slowly changing.
3. The microphone signals should be recorded simultaneously to find the exact delay.

These requirements severely degrade performance in practical situations. There are usually many interfering sounds, and beamforming or BSS algorithms fail to suppress the interference arising from the first requirement. The third requirement is not always satisfied because of the cost involved in the synchronization of multiple recordings.

This paper proposes an algorithm for canceling the interfering sounds with stereo recordings of the nonlinear mixing of more than two sources. The proposed method is based on Adaptive Noise Cancellation (ANC) [14]. An assumption is made that the primary recording is a mixture of the target sound and several interfering signals, and the secondary recording called the reference is a mixture of only interfering signals.

Our proposed algorithm is particularly applicable for smart TV environments in which two individual devices, a remote control and a TV set, are equipped with microphones. The primary microphone that is in the remote control directly receives noise-contaminated speech from a target speaker to perform automatic voice recognition while the secondary microphone in the TV collects the background noise and provides the reference signal for the background noise source. Further details of this approach are described in Section IV.

III. The Proposed Noise Cancellation Method

This chapter explains the proposed interference equalization and target activity detection algorithms in detail.

A. Interference Equalization

The proposed method assumes that the primary microphone equipped in a remote control picks up a mixture of the target sound and interference, and the signal recorded by the secondary microphone in a TV set contains interference only. This is expressed as

$$x_0(t)=s(t)+r_0(t), \quad x_1(t)=r_1(t), \tag{1}$$

where $x_0(t)$ and $x_1(t)$ are the signals recorded by the primary and secondary microphones, $s(t)$ is the target source, $r_0(t)$ is the interference mixed in the primary recording, and $r_1(t)$ is the interference picked up by the secondary microphone. Because the assumed mixing architecture of ANC is linear, it may not model the linear filter's mismatch to the nonlinear, realistic conditions [15], [16]. Using short-time Fourier transform, (1) is rewritten as

$$X_0(\omega)=S(\omega)+R_0(\omega), \quad X_1(\omega)=R_1(\omega), \tag{2}$$

where ω is a frequency variable, complex values $X_0(\omega)$, $X_1(\omega)$, $S(\omega)$, $R_0(\omega)$, and $R_1(\omega)$ represent the DFTs of $x_0(t)$, $x_1(t)$, $s(t)$, $r_0(t)$, and $r_1(t)$, respectively. We assume that $R_0(\omega)$ can be approximated by a scalar multiple of $R_1(\omega)$, so that the unknown target source $S(\omega)$ should be recovered by

$$\begin{aligned}\hat{S}(\omega)&=X_0(\omega)-B(\omega)X_1(\omega)\\&\approx S(\omega)+(R_0(\omega)-B(\omega)R_1(\omega)),\end{aligned} \tag{3}$$

where $B(\omega)$ is a constant that equalizes $R_0(\omega)$ and $R_1(\omega)$. The estimate of the target source $\hat{S}(\omega)$ becomes the true one if and only if $R_0(\omega)=B(\omega)R_1(\omega)$, i.e., when a perfect equalization is made. Fig. 2 shows an overview of the proposed method. The equalized reference signal is subtracted from the input to generate the estimate of the target. To obtain a phase-independent error criterion, we enforce $B(\omega)$ to be a positive real number, and use a squared difference of the Power Spectral Densities (PSDs) between the interferences in the primary recording and the equalized secondary recording:

$$E(\omega)=\|R_0(\omega)-B(\omega)R_1(\omega)\|^2. \tag{4}$$

Fig. 2. Block diagram of the proposed method.

An objective function $J(\omega)$ is defined by the sum of the equalization errors in (4) over all of the short time analysis frames, expressed as

$$\begin{aligned}J(\omega)&=\sum_{\text{all } n}E(\omega,n)\\&=\sum_{\text{all } n}\|R_0(\omega,n)-B(\omega)R_1(\omega,n)\|^2\end{aligned} \tag{5}$$

- Como o sinal é misturado [X0(w) = S(W)+ R0(W)], R0(w) não está disponível. Portanto, utiliza-se um método para detectar a atividade da fonte-alvo na gravação primária, de modo que o alvo seja excluído do cálculo do filtro de equalização.

- Tem-se a divisão da análise dos sinais em quadros, que são classificados em duas classes: interferência com alvo[Cr+s(w)] e somente interferência[Cr(w)] . Regra de decisão:

PSDs = R0(w)-B(w)*R1(w)

n pertence a Cr+s(w) se R0(w) > B(w)*R1(w) [PSDs > zero]

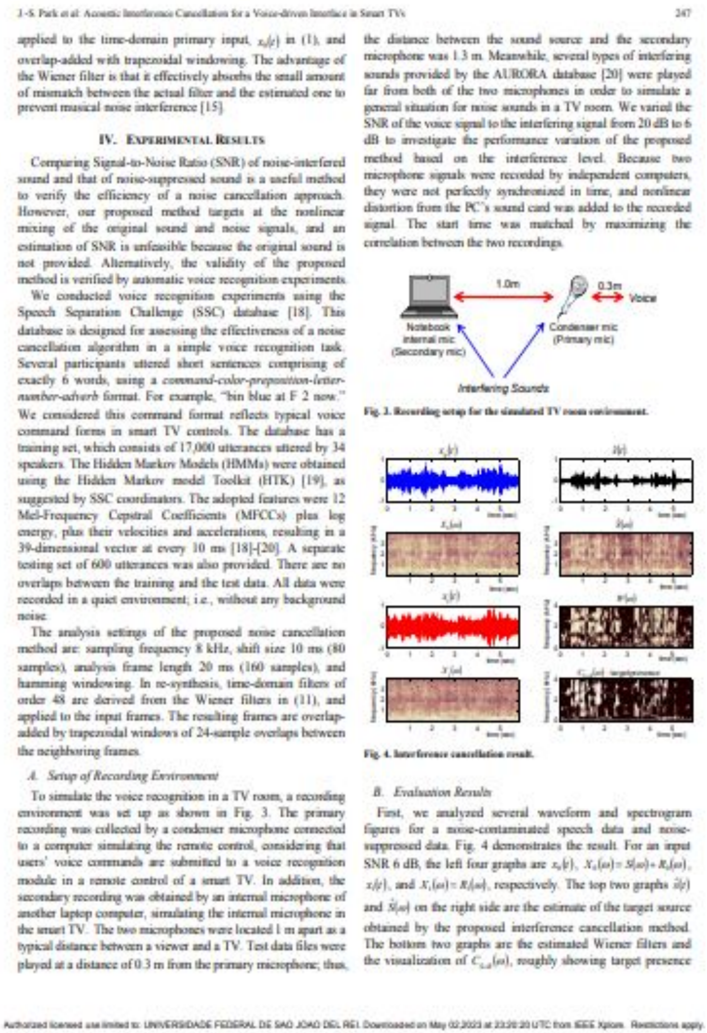
n pertence a Cr(w) se R0(w) ≤ B(w)*R1(w) [PSDs < zero]

- O filtro de equalização é atualizado usando apenas os quadros cujos índices estão em Cr(w), ou seja, quadros que contenham somente interferência.

- Uma vez que os coeficientes do filtro de equalização B’(w) são obtidos, um filtro de Wiener é derivado para suprimir a interferência estimada a partir da PSD do canal primário, garantindo um ganho adequado.

- Então, ele é convertido em um filtro FIR no domínio do tempo e é aplicado à entrada primária x0(t), com sobreposição usando janelamento trapezoidal, absorvendo a pequena diferença entre o filtro real e o estimado, evitando interferência de ruído musical.

IV e V - RESULTADOS EXPERIMENTAIS e CONCLUSÕES:



applied to the time-domain primary input, $x_0(t)$ in (1), and overlap-added with trapezoidal windowing. The advantage of the Wiener filter is that it effectively absorbs the small amount of mismatch between the actual filter and the estimated one to prevent musical noise interference [15].

IV. Experimental Results

Comparing Signal-to-Noise Ratio (SNR) of noise-interfered sound and that of noise-suppressed sound is a useful method to verify the efficiency of a noise cancellation approach. However, our proposed method targets at the nonlinear mixing of the original sound and noise signals, and an estimation of SNR is unfeasible because the original sound is not provided. Alternatively, the validity of the proposed method is verified by automatic voice recognition experiments.

We conducted voice recognition experiments using the Speech Separation Challenge (SSC) database [18]. This database is designed for assessing the effectiveness of a noise cancellation algorithm in a simple voice recognition task. Several participants uttered short sentences comprising of exactly 6 words, using a *command-color-preposition-letter-number-adverb* format. For example, "bin blue at F 2 now." We considered this command format reflects typical voice command forms in smart TV controls. The database has a training set, which consists of 17,000 utterances uttered by 34 speakers. The Hidden Markov Models (HMMs) were obtained using the Hidden Markov model Toolkit (HTK) [19], as suggested by SSC coordinators. The adopted features were 12 Mel-Frequency Cepstral Coefficients (MFCCs) plus log energy, plus their velocities and accelerations, resulting in a 39-dimensional vector at every 10 ms [18]-[20]. A separate testing set of 600 utterances was also provided. There are no overlaps between the training and the test data. All data were recorded in a quiet environment; i.e., without any background noise.

The analysis settings of the proposed noise cancellation method are: sampling frequency 8 kHz, shift size 10 ms (80 samples), analysis frame length 20 ms (160 samples), and hamming windowing. In re-synthesis, time-domain filters of order 48 are derived from the Wiener filters in (11), and applied to the input frames. The resulting frames are overlap-added by trapezoidal windows of 24-sample overlaps between the neighboring frames.

A. Setup of Recording Environment

To simulate the voice recognition in a TV room, a recording environment was set up as shown in Fig. 3. The primary recording was collected by a condenser microphone connected to a computer simulating the remote control, considering that users' voice commands are submitted to a voice recognition module in a remote control of a smart TV. In addition, the secondary recording was obtained by an internal microphone of another laptop computer, simulating the internal microphone in the smart TV. The two microphones were located 1 m apart as a typical distance between a viewer and a TV. Test data files were played at a distance of 0.3 m from the primary microphone; thus, the distance between the sound source and the secondary microphone was 1.3 m. Meanwhile, several types of interfering sounds provided by the AURORA database [20] were played far from both of the two microphones in order to simulate a general situation for noise sounds in a TV room. We varied the SNR of the voice signal to the interfering signal from 20 dB to 6 dB to investigate the performance variation of the proposed method based on the interference level. Because two microphone signals were recorded by independent computers, they were not perfectly synchronized in time, and nonlinear distortion from the PC's sound card was added to the recorded signal. The start time was matched by maximizing the correlation between the two recordings.

1.0m
0.3m
Voice
Notebook internal mic (Secondary mic)
Condenser mic (Primary mic)
Interfering Sounds

Fig. 3. Recording setup for the simulated TV room environment.

Fig. 4. Interference cancellation result.

B. Evaluation Results

First, we analyzed several waveform and spectrogram figures for a noise-contaminated speech data and noise-suppressed data. Fig. 4 demonstrates the result. For an input SNR 6 dB, the left four graphs are $x_0(t)$, $X_0(\omega)=S(\omega)+R_0(\omega)$, $x_1(t)$, and $X_1(\omega)=R_1(\omega)$, respectively. The top two graphs $\hat{s}(t)$ and $\hat{S}(\omega)$ on the right side are the estimate of the target source obtained by the proposed interference cancellation method. The bottom two graphs are the estimated Wiener filters and the visualization of $C_{s,n}(\omega)$, roughly showing target presence

- Um ambiente de gravação foi configurado para simular o reconhecimento de voz em uma sala de TV.

- A gravação primária foi feita com um microfone conectado a um computador [simulando o controle remoto], posicionado a um metro de distância da gravação secundária, que foi feita com um microfone interno de outro computador [simulando o microfone interno da smart TV].

- Sons interferentes fornecidos por banco de dados foram reproduzidos para simular uma situação realista de ruídos em uma sala de TV, variando, inclusive, a Relação Sinal-Ruído (SNR) entre o sinal de voz e o sinal interferente para avaliar o desempenho do método em diferentes níveis de interferência.

- O método utiliza a máscara das densidades espectrais de potência do sinal de entrada e a gravação de referência de outro dispositivo para reduzir a interferência estimada, conseguindo suprimir com sucesso os sons interferentes e rastrear bem a fala alvo original, mostrando uma melhoria significativa no desempenho do reconhecimento de voz em diferentes níveis de SNR sem diminuir o desempenho na condição limpa, sendo capaz de lidar com sons de ruído não-estacionário e múltiplas fontes de interferência.

- Apesar da notável melhoria, não é uma boa prática utilizar esse método para SNR abaixo de 6 dB devido à grave distorção da fala causada pela interferência.