Automatic Speech Translation System Selecting Target Language by Direction-of-Arrival Information

Masanori Tsujikawa
Department of Electrical, Electronic and Information Engineering, Faculty of Engineering Science
Kansai University
Osaka, Japan

Biometrics Research Laboratoties
NEC Corporation
Kanagawa, Japan
tujikawa@cb.jp.nec.com

Koji Okabe
Biometrics Research Laboratoties
NEC Corporation
Kanagawa, Japan

Ken Hanazawa
Biometrics Research Laboratoties
NEC Corporation
Kanagawa, Japan

Yoshinobu Kajikawa
Department of Electrical, Electronic and Information Engineering, Faculty of Engineering Science
Kansai University
Osaka, Japan
kaji@kansai-u.ac.jp

***Abstract*—In this paper, we propose an automatic speech translation system that selects its target language on the basis of the direction-of-arrival (DOA) information. The system uses two microphones to detect speech signals arriving from specific directions. The target language for speech recognition is selected on the basis of the DOA. Both the speech detection and target language selection relieves users from operations normally required for individual utterances, without serious increase in computational costs. In a speech-recognition evaluation of the proposed system, 80% word accuracy was achieved for utterances recorded with two microphones that were 40cm distant from speaker positions. This accuracy is nearly equivalent to that in which the time frame and target language of a user's speech are given in advance.**

Keywords—automatic speech translation, speech recognition, language identification, direction of arrival, speech detection, microphone array

I. Introduction

We have developed a compact bidirectional Japanese-English automatic translation system that runs on such terminals as mobile phones [1]. On a small-screen mobile phone terminal, it may ordinarily take a long time from when a user A utters something to when a user B sees the translation results, i.e., the following cumbersome procedures 1 to 5 may be required.

1. User A performs the operations to select a target language and then to start speech input.
2. User A utters something.
3. User A confirms the speech recognition and translation results.
4. User A shows the screen of the mobile phone to user B.
5. User B sees the translation results.

Tablet terminals equipped with larger screens than mobile phones are also widely used. When running an automatic translation system on a tablet [1]-[3], two users can talk while confirming the translation results at the same time. That is, user B can see the translation (procedure 5) without waiting for procedures 3 and 4, which shortens the required time and smooths a translated conversation.

Even with tablets, however, the process is still insufficiently smooth due to the necessity of procedure 1, especially since the procedure is necessary before speaker change, which frequently occurs in conversation. To avoid the need for procedure 1, both accurate speech detection [4]-[7] and accurate language discrimination [8] need to be achieved without any button operations. Further, with two users simultaneously viewing the screen, noise-robust speech recognition [9] capable of handling speech at distances of tens of centimeters is also required.

This paper proposes an automatic speech translation system that selects a target language on the basis of the direction-of-arrival (DOA) information and performs both two-microphone speech detection and language discrimination. For distant speech recognition, speech-model-based noise suppression and multi-condition training of acoustic models [9] are used. The following sections describe in detail the proposed system, our evaluation methods, and evaluation results.

II. Translation Tablets

In this paper, we refer to a tablet terminal equipped with an automatic translation application as a translation tablet. It is assumed that two speakers of different mother tongues face each other and employ the translation tablet at such places, as reception desks, ticket sales windows, offices, or commercial facility counters. Fig. 1 shows an example in which a translation tablet is placed between two people and both recognition and translation results are mutually observed as a conversation is conducted.

In actual use, we may assume there to be background noise that will degrade speech recognition accuracy. We also assume here that another customer and clerk are talking in the same way at a neighboring counter. Their voices (interference sound) will

ISBN 978-90-827970-1-5

further reduce speech recognition accuracy. For conversational smoothness in such a high noise-level environment, accurate speech detection, language discrimination, and speech recognition need to be performed automatically.

III. Proposed Method

Assuming the usage scenario described above, a translation tablet needs to satisfy the following three requirements.

- Accurate, noise-robust automatic speech detection
- Accurate, noise-robust automatic language discrimination
- Accurate, noise-robust automatic speech recognition

With respect to the speech detection and language discrimination requirements, we propose the use of two-microphone speech detection. The requirement for noise-robust speech recognition can be satisfied by combining speech-model-based noise suppression with multi-noise-condition training for acoustic models. Our methods are explained below.

A. Two-microphone speech detection and language discrimination on the basis of DOA information

In order to perform accurate noise-robust automatic speech detection and accurate noise-robust automatic language discrimination, we use the 2-microphone speech detection proposed in [4], which detects speech arriving from specific directions.

Fig.2 shows the structure of the two-microphone speech detection. Speech is detected on the basis of a ratio of the output of the upper-side filter (Beamfomer), which emphasizes sound arriving from a certain direction α, and the output of the lower-side filters (Null-Beamformer and Spatial spectral subtraction), which remove sound arriving from that direction α. When the ratio is larger than a given threshold, it is detected as speech arriving from direction α [rad], and, conversely, when the ratio is below the threshold, it is judged that the speech has not arrived from direction α. Using both a complex frequency spectrum $X_1(f,t)$ of signals from microphone 1 and that $X_2(f,t)$ of signals from microphone 2, the output amplitude spectra $|Y_1(f,t)|$ of Beamformer and $|Y_2(f,t)|$ of Null-Beamformer are respectively obtained as follows, where f is frequency bin number and t is time frame number.

$$|Y_1(f,t)|=|W_1(f)X_1(f,t)+W_2(f)X_2(f,t)|, \quad (1)$$

$$|Y_2(f,t)|=|W_1(f)X_1(f,t)-W_2(f)X_2(f,t)|, \quad (2)$$

where,

$$W_1(f)=\exp\{-j2\pi f(f_s/N)d_1 \sin\alpha/c\}, \quad (3)$$

$$W_2(f)=\exp\{-j2\pi f(f_s/N)d_2 \sin\alpha/c\}. \quad (4)$$

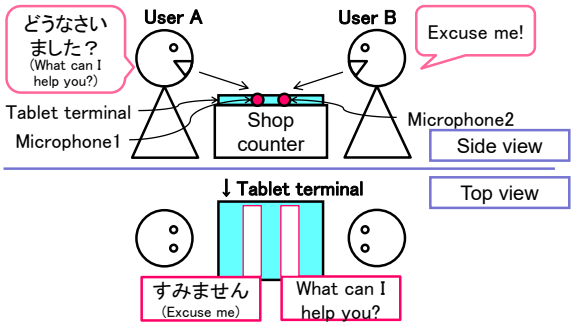

Fig. 1. Example use of translation tablet.

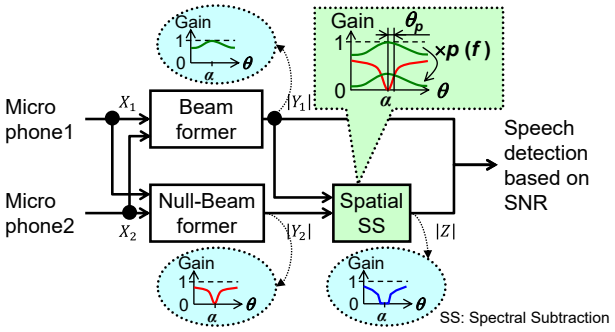

Fig. 2. Structure of two-microphone speech detection.

In (3) and (4), f_s [Hz], N, and c [m/s] are, respectively, the sampling frequency, the point of a Fourier transform, and sound speed. d_1 and d_2 are the positions [m] of microphones 1 and 2, respectively. Next, using $|Y_1(f,t)|$ and $|Y_2(f,t)|$, the output $|Z(f,t)|$ of Spatial spectral subtraction can be obtained as follows.

$$|Z(f,t)|=\max[|Y_2(f,t)|-p(f)|Y_1(f,t)|,0], \quad (5)$$

$$p(f)=\sqrt{D_2(f,\alpha+\theta_p,\alpha)/D_1(f,\alpha+\theta_p,\alpha)}, \quad (6)$$

$$D_1(f,\alpha+\theta_p,\alpha)=\cos^2[2\pi f(f_s/N)d_1\{\sin(\alpha+\theta_p)-\sin\alpha\}/c], \quad (7)$$

$$D_2(f,\alpha+\theta_p,\alpha)=\sin^2[-2\pi f(f_s/N)d_2\{\sin(\alpha+\theta_p)-\sin\alpha\}/c]. \quad (8)$$

In (6), $D_1(f,\alpha+\theta_p,\alpha)$ and $D_2(f,\alpha+\theta_p,\alpha)$ are the directivity patterns for Beamformer and Null-Beamformer, respectively, at frequency bin f and direction $\alpha+\theta_p$. Speech is detected on the basis of a ratio of $|Y_1(f,t)|$ and $|Z(f,t)|$. One important feature of this method is the removal of sound arriving from specific directions in two stages, i.e., the complex spectral domain in (2) and the amplitude spectral domain in (5). This two-stage removal makes speech detection robust in cases in which the DOA of speech deviates within a certain range $\pm\theta_p$ [rad] from the assumed direction α.

Fig.3 shows the structure of the proposed system for selecting a target language using the above two-microphone

ISBN 978-90-827970-1-5 © EURASIP 2018

speech detection. With respect to the translation tablet, we assume that the direction of user A is direction 1 and that of user B is direction 2. The proposed system has two blocks of two-microphone speech detection, for detecting speech arriving either only from direction 1 or only from direction 2. By discarding sound arriving from other directions, it is possible to reduce detection error due to background noise and interference speech over that possible with only single-microphone use.

Assuming that directions 1 and 2 are sufficiently different, two-microphone speech detection for direction 1 can also reject speech arriving from direction 2, making it is possible to discriminate among languages on the basis of DOA. For example, as shown in Fig. 3, speech arriving from direction 1 is recognized as Japanese, and that from direction 2 as English. The conventional language discrimination method in [8], which compares likelihoods of recognition results for both languages is not necessarily high in discrimination accuracy, and the costs of calculating likelihoods for both languages is very high. Also, DOA discrimination offers higher accuracy.

B. *Speech-model-based noise suppression and multi-noise-condition training for acoustic models*

Since the volume of speech from a distance of several tens of centimeters from microphones is relatively low, the SNR (speech to noise ratio) is similarly low. Recognizing low-SNR speech with acoustic models trained using a large amount of high-SNR speech data results in low recognition accuracy, but by combining speech-model based noise suppression with multi-noise-condition training for acoustic models [9], we are able to achieve accurate noise-robust speech recognition. In this paper, due to space limitations, we omit explanation of speech-model-based noise suppression.

The speech data used for multi-noise-condition training of acoustic models are as follows. For clean speech data recorded in a quiet environment, noise data was added so that the SNR would be three normal distributions. The parameter for the three distributions was set so that (mean, standard deviation) = (5dB, 3dB), (15dB, 3dB), and (25dB, 3dB). Noisy speech data was used for training noisy acoustic models, which were mixed with clean acoustic models that had been trained using clean speech data. These mixed acoustic models were used for noise-robust speech recognition.

IV. Recording Speech Data for Evaluations

We assumed actual-use environments in experimentally evaluating the effectiveness of the proposed method. Recording was performed using a mock-up whose shape and size were the same as a 7-inch tablet terminal. Two microphones were placed at an interval of 3 cm, as shown in Fig. 4. For recording, we used a soundproof, 5.0 x 5.0 x 2.2m room having a reverberation time of 0.3s.

As shown in Fig. 4, in addition to the tablet terminal mock-up equipped with two microphones, we used two loudspeakers for playing speech files. In a quiet environment (a noise level of 28 dBA), Japanese-speech files and English-speech files were alternately played from two loudspeakers and recorded through the two microphones. θ was varied through 0, 15, 30, 45, and 60 degrees. L was varied through 20, 30, and 40 cm. The speech

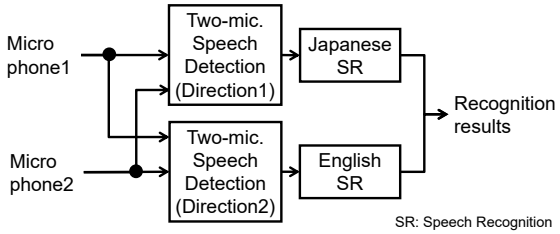

Fig. 3. Structure of proposed system that selects target language on the basis of the DOA information.

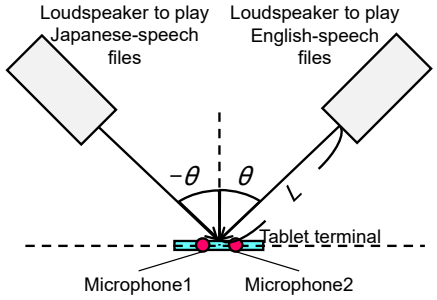

Fig. 4. Arrangement of microphones and loudspeakers for recording speech data for use in evaluations.

files contained travel-related conversations, and for each loudspeaker position, 40 speech files were played and recorded. They had been prerecorded as uttered, five times each, by four Japanese speakers and four English speakers.

Noise files were also played through other loudspeakers located farther away (near room walls) so as to diffuse the noise over the room space. The volumes of three types of background noise were set to 50dBA for office noise, 45dBA for lobby noise, and 57dBA for sales-window noise. Recording of the diffused background noise was also done through the two tablet mock-up microphones. The background noise thus recorded was artificially added to the above-mentioned speech data to create sound data for our evaluations.

We assumed a conversation also to be taking place between another customer and another clerk at a neighboring counter, defined that speech as interference, and recorded it as follows. In arranging the two loudspeakers, we used θ=45 degrees, L=30 cm, and a horizontal distance of 1 m between the tablet mock-up and loudspeakers. Evaluation data containing interference speech were created by artificially adding interference speech data to the above-mentioned evaluation data that contained no interference speech. Both the ratio of speech to be recognized to background noise, and the ratio of speech to be recognized to interference speech are shown in Table 1.

V. Experimental Evaluations

A. *Speech detection evaluations*

We used the speech data, background noise data, and interference speech data described in the previous section to evaluate two-microphone speech detection. As evaluation

ISBN 978-90-827970-1-5 © EURASIP 2018

indices, we used the detection rate and rejection rate defined by the following equations.

$$\text{Detection rate } [\%] = N_{\text{utt}}(b \cap c) / N_{\text{utt}}(a) \times 100,$$

$$\text{Rejection rate } [\%] = N_{\text{utt}}(b \cap d) / N_{\text{utt}}(a) \times 100,$$

where $N_{\text{utt}}(x)$ is the number of utterances satisfying the condition of x. The conditions for a, b, c, and d are as follows:

a. All utterances (= 40 utterances)

b. 90% or more of an overall duration in which there is no speech to be detected is correctly judged as being the duration.

c. 90% or more of an overall duration in which there is speech to be detected is correctly judged as being the duration.

d. 90% or more of an overall duration in which there is speech to be rejected is correctly judged as being the duration.

Since this two-microphone speech detection detects speech arriving only from specific directions, it is desirable that the detection rate for those directions should be high and the rejection rate low. Conversely, the detection rate for other directions should be low and the rejection rate high.

Fig. 5 shows the evaluation results for speech detection without interference speech, and Fig. 6 shows that with interference speech. Both Figs. show the average of results for the three types of background noise. The left-hand side of each of the Figs. shows the detection and rejection rates in the case of setting to detect speech arriving from the direction -45degrees, while the right sides show those in the case of 45degrees. Results in both Figs. are those for a distance L=30cm between the loudspeakers and microphones. With the -45-degree setting in Fig. 5, for DOAs of -60 degrees and -45 degrees, the detection rates are both 85% or more, and the rejection rates are 0%. For DOAs of from -15 degrees to 60 degrees, the detection rates are 0% and the rejection rates are 95% or more. As expected, speech arriving from directions close to -45 degrees was accurately detected, and speech arriving from other directions was accurately rejected. The difference between Figs. 5 and 6 is small. In the case of a -45-degree setting for two-microphone speech detection, interference speech was sufficiently rejected, as was the cases 45 degrees as well.

Figs. 5 and 6 show that, with an advance setting of which speech language will arrive from which direction, e.g., Japanese speech arriving from -45degrees and English speech arriving from 45 degrees, it is possible for the two-microphone arrangement to achieve both accurate, noise-robust automatic speech detection and accurate, noise-robust automatic language discrimination.

B. *Speech recognition evaluations*

We also evaluated speech recognition. Parameters were set to θ= 45 degrees and L=20, 30, 40 cm. Conditions for

TABLE I. THE RATIO OF SPEECH TO BE RECOGNIZED TO BACKGROUND NOISE, AND THE RATIO OF SPEECH TO BE RECOGNIZED TO INTERFERENCE SPEECH.

	Distance L	20cm	30cm	40cm
Type of background noise	Quiet	40.4dB	37.0dB	35.0dB
	Office	21.2dB	17.8dB	15.8dB
	Lobby	21.1dB	17.7dB	15.7dB
	Sales Window	12.6dB	9.1dB	7.1dB
	Interference speech	14.6dB	11.2dB	9.2dB

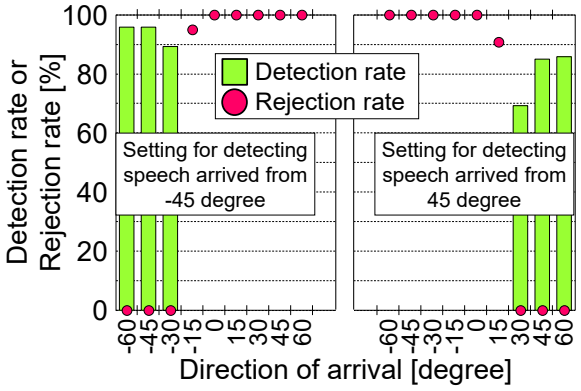

Fig. 5. Evaluation results for speech detection without interference speech.

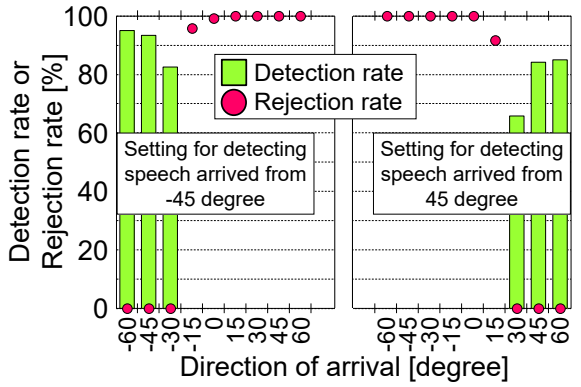

Fig. 6. Evaluation results for speech detection with interference speech.

background noise and interference speech were the same as in the previous section. Speech recognition accuracies (word accuracies) with methods A, B, and C in Table 2 were compared. The single-microphone speech detection method, which was internally developed, is based both on speech Gaussian mixture models (GMMs) and on noise GMMs. A parallel search with a likelihood criterion selects one of two recognition results, that for the output of a Japanese speech recognition system or that for an English speech recognition system output, on the basis of likelihoods. That is, it is necessary to use two speech recognition systems to process a single utterance, which results in seriously high computational costs.

In word accuracy calculation, when language-discrimination errors occurred, it was assumed that deletion errors for the

ISBN 978-90-827970-1-5 © EURASIP 2018

uttered language had occurred and that insertion errors for the other language had occurred.

Fig. 7 shows evaluation results for speech recognition without interference speech, and Fig. 8 shows those with interference speech. Both Figs. show the average of results for the three types of background noise. Word accuracies for Japanese speech are shown on the left-hand sides of both Figs., and those for English speech are shown on the right-hand sides.

With two-microphone speech detection, word accuracies for B were much better than those for A. The average error reduction rate was 44.0% without interference speech and 70.2% with interference speech, i.e., two-microphone speech detection can suppress insertion error due to erroneous detection of either background noise or interference speech.

As compared to B, use of DOA information for language discrimination in C resulted in the average error reduction rates of 5.1% for English and 3.7% for Japanese. With language discrimination based on a likelihood criterion, Japanese was often taken to be English, but use of DOA information made more accurate discrimination possible, and insertion errors in English were few. Further, it is noteworthy that, in speech recognition, C requires only half the computational cost required by B.

With C, an accuracy of word correctness of roughly 80% can be obtained for both Japanese and English at an L=40 cm distance under interference sound conditions. This accuracy is nearly equivalent to the case in which the time frame and target language of a user's speech are given in advance. Our proposed system successfully achieves both conversational smoothness and high word accuracy.

VI. Conclusion

For the purpose of achieving accurate, conversationally smooth automatic speech translation, we have developed a system that selects target language on the basis of DOA information. Its automatic speech detection and automatic language discrimination relieves users from annoying button operations that would otherwise be required for individual utterances, without serious increase in computational costs. In speech recognition evaluation, word accuracy of 80% has been achieved for utterances recorded with two microphones located at 40 cm distances from speaker positions.

TABLE II. Methods for comparison

	A	B	C (Proposed)
Speech detection	Single-microphone	Two-microphone	Two-microphone
Language discrimination	Parallel search with likelihood criterion	Parallel search with likelihood criterion	Direction of arrival information

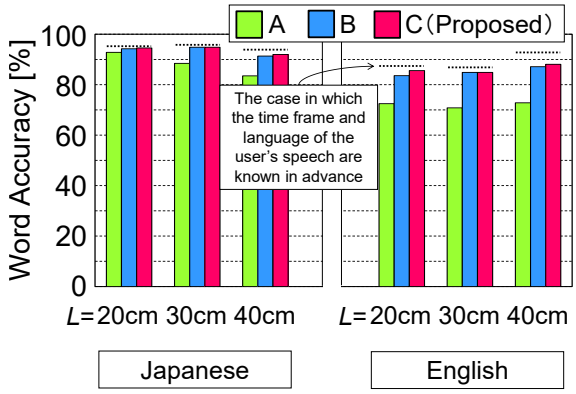

Fig. 7. Evaluation results of speech recognition without interference speech.

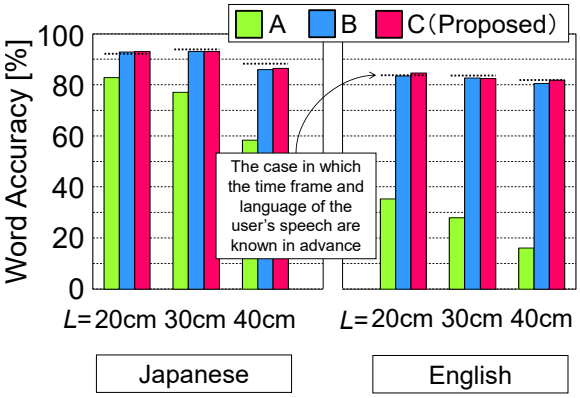

Fig. 8. Evaluation results of speech recognition with interference speech.

References

[1] Ken Hanazawa, Akitoshi Okumura, Koji Okabe, Shinichi Ando, "Development and Evaluation of the Fast and Accurate Compact-scalable Speech Translation Software," Information Processing Society of Japan Transactions on Consumer Devices and Systems (CDS), vol.2, no.2, pp.10-18, July 2012 (in Japanese).

[2] Matthias Eck, Ian lane, Ying Zhang, and Alex Waibel, "Jibbigo: Speech-to-speech Translation on mobile devices," Proceedings of IEEE Spoken Language Technology Workshop, pp.165-166, Dec. 2010.

[3] Shigeki Matsuda, Teruaki Hayashi, Yutaka Ashikari, Yoshinori Shiga, Hidenori Kashioka, Keiji Yasuda, Hideo Okuma, Masao Uchiyama, Eiichiro Sumita, Hisashi Kawai, and Stoshi Nakamura, "Development of the "VoiceTra" Multi-Lingual Speech Translation System," IEICE Transaction on Information & Systems, vol.E100-D, no.4, pp.621-632, Apr. 2017.

[4] Masanori Tsujikawa, "Robust Speech Detection Using 2-Microphone Outputs for Hands-Free Speech Recognition," Proceedings of Acoustic Society of Japan 2005 Spring Conference, pp.121-122, Mar. 2005 (in Japanese).

[5] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," HSCMA2008, pp.29--32, 2008.

[6] Yanmeng Guo, Kai Li, Qiang Fu, Yonghong Yan, "A Two-Microphone Based Voice Activity Detection for Distant-Talking Speech in Wide Range of Direction of Arrival, " Proc. of ICASSP 2012, pp.4901-4904, Mar. 2012.

[7] Elias Nemer and Ashutosh Pandey, "A Dual-Microphone Subband-Based Voice Activity Detector Using Higher-Order Cumulants, " Proc. of ICASSP 2012, pp.5995-5999, May 2014.

[8] Marc A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," IEEE Transactions on Speech and Audio Processing, vol.4, no.1, pp.31-44, Jan.1996.

[9] Masanori Tsujikawa, Takayuki Arakawa, and Ryosuke Isotani, "In-Car Speech Recognition Using Model-Based Wiener Filter and Multi-Condition Training, " Proc. of Interspeech 2008, pp.972-975, Sep. 2008.

ISBN 978-90-827970-1-5 © EURASIP 2018