Pseudo-pitch-synchronized phase information extraction and its application for robust speaker recognition

1st Longbiao Wang
Tianjin Key Lab. of Cognitive Computing and Application
Tianjin University
Tianjin, China
longbiao_wang@tju.edu.cn

2nd Seiichi Nakagawa
Toyohashi University of Technology
Toyohashi, Japan
nakagawa@slp.cs.tut.ac.jp

3rd Jianwu Dang
Tianjin Key Lab. of Cognitive Computing and Application
Tianjin University
Tianjin, China
dangjianwu@tju.edu.cn

4th Jianguo Wei
Tianjin Key Lab. of Cognitive Computing and Application
Tianjin University
Tianjin, China
jianguo.fr@gmail.com

5th Tongtong Shen
Tianjin Key Lab. of Cognitive Computing and Application
Tianjin University
Tianjin, China
ttshen@tju.edu.cn

6th Lantian Li
Tsinghua University
Beijing, China
lilt@cslt.riit.tsinghua.edu.cn

7th Thomas Fang Zheng
Tsinghua University
Beijing, China
fzheng@tsinghua.edu.cn

***Abstract*— Recent studies have shown that phase information contains speaker-dependent characteristics and is effective for speaker recognition. In this paper, we summarize a robust phase feature extracted from Fourier spectrum (including pitch non-synchronized phase information and pseudo-pitchsynchronized phase information) and its application for speaker recognition for different speaking rate speech and noisy speech, and add the evaluation of speaker identification for short duration speech of training dataset and test set and speaker verification for telephone speech with channel variability. For pseudopitch-synchronized phase information extraction, the maximum amplitude of each frame is adopted as the center of the next window. Experiments were conducted using the Japanese Newspaper Article Sentence (JNAS) database, NTT database and NIST SRE 2003 database. The pseudo-pitch-synchronized phase information significantly outperformed than our proposed conventional pitch non-synchronized phase information for all cases. By combining the proposed phase information with MFCC, the speaker recognition performance was remarkably improved than that of MFCC.**

***Keywords*—speaker recognition; phase information; GMM; noisy speech; short duration speech**

I. Introduction

In conventional speaker identication methods based on mel-frequency cepstral coefcients (MFCCs), only the magnitude of the Fourier transform in time-domain speech frames has been used. This means that the phase component is ignored. MFCCs capture not only speaker-specic vocal tract information, but also some vocal source characteristics. However, speaker characteristics in the voice source are not captured completely by the MFCC. Therefore, feature parameters extracted from excitation source characteristics are also useful for speaker identication [1][2][3][4]. Plumpe et al. proposed an automatic technique for estimating and modeling the glottal ow derivative source waveform of speech and applying the model parameters to speaker identication [1]. The complementary nature of speaker-specic information in the residual phase and information in conventional MFCCs was demonstrated in [2]. In [3], two proposed glottal signatures for speaker identification are proposed. The wavelet octave coefficients of residues [4] is presented to capture the spectro-temporal source excitation characteristics. The residual phase was derived from speech signals by linear prediction analysis. Recently, many speaker recognition studies using group-delay-based phase information have been proposed [5][6]. Group delay is dened as the negative derivative of the phase of the Fourier transform of a signal. In [5], the authors analytically showed why the group delay based phase are robust to noise. Actually, the group delay based phase contains both the power spectrum and phase information [5][6][7], and thus the complementary nature of the power spectrum-based MFCC and group delay phase was not sufcient enough.

In this paper, we summarize a robust Fourier spectrum based phase information extraction method and its application for speaker identication with normal-duration speech and noisy speech, and add the evaluation of speaker recognition for

978-1-5090-4045-2/17/$31.00 ©2017 IEEE

short duration speech and telephone speech. Previously, we proposed a speaker identication system using a combination of MFCCs and phase information [8][9][10], directly extracted from the limited bandwidth of the Fourier transform of the speech wave. The experimental results showed that the phase information is effective for speaker identication in clean and noisy environments [8][9][10]. However, there are still some problems occurred in extracting the phase information because of the inuence of the windowing position. In [11][12][13][14], we proposed a new method to extract pseudo-pitch-synchronized phase information with maximum amplitude synchronization. In this paper, we compare the speaker recognition performance using pseudo-pitchsynchronized phase information and pitch non-synchronized phase information for different speaking rate speech, noisy speech, short duration speech and telephone speech.

The rest of the paper is organized as follows. Section 2 gives a brief introduction of the Phase information extraction. Section 3 describes Combination method and decision method. The experiments are given in Section 4. Finally, we conclude the research and the future work in Section 5.

II. Phase information extraction

A. Formulas [10]

The spectrum $S(\omega,t)$ of a signal is obtained by DFT of an input speech signal sequence

$$\begin{aligned} S(\omega,t) &= X(\omega,t) + jY(\omega,t) \\ &= \sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times e^{j\theta(\omega,t)}. \end{aligned} \tag{1}$$

However, the phase changes, depending on the clipping position of the input speech even at the same frequency ω. To overcome this problem, the phase of a certain basis frequency ω is kept constant, and the phases of other frequencies are estimated relative to this. For example, by setting the basis frequency ω to 0, we obtain

$$S^{'}(\omega,t) = \sqrt{X^2(\omega,t)+Y^2(\omega,t)} \times e^{j\theta(\omega,t)} \times e^{j(-\theta(\omega,t))}. \tag{2}$$

whereas for the other frequency $\omega^{'} = 2\pi f^{'}$, the spectrum becomes [10]

$$\begin{aligned} &S^{'}(\omega^{'},t) \\ &= \sqrt{X^2(\omega^{'},t)+Y^2(\omega^{'},t)} \times e^{j\theta(\omega^{'},t)} \times e^{j\frac{\omega^{'}}{\omega}(-\theta(\omega,t))}. \end{aligned} \tag{3}$$

This way, the phase can be normalized, and the normalized phase information becomes

$$\widetilde{\theta} = \theta(\omega^{'},t) + \frac{\omega^{'}}{\omega}(-\theta(\omega,t)). \tag{4}$$

In the experiments described in this paper, the basis frequency ω is set to $2\pi\times$ 1000 Hz. In the previous study, we used phase information only in a sub-band frequency range to reduce the number of feature parameters. However, a problem arose with this method when comparing two phase values. For example, for two values $\pi - \widetilde{\theta}_1$ and $\widetilde{\theta}_2 = -\pi + \widetilde{\theta}_1$, the difference is $2\pi - 2\widetilde{\theta}_1$. If $\widetilde{\theta}_1 \approx \theta$, then the difference $\approx 2\pi$, despite the two phases being very similar to each other. Therefore, we modied the phase into coordinates on a unit circle [10], that is,

$$\widetilde{\theta} \rightarrow \left\{ cos\widetilde{\theta}, sin\widetilde{\theta} \right\}. \tag{5}$$

B. Pseudo-pitch-synchronized phase information

We can reduce the phase variation using the relative phase extraction method that normalizes the phase variation by cutting positions. However, the normalization of phase variation is still inadequate. For example, for a 1000-Hz periodic wave (16 samples per cycle for a 16-kHz sampling frequency), if one sample point shifts in the cutting position, the phase shifts only by $\frac{2\pi}{16}$, while for a 500-Hz periodic wave, the phase shifts only by $\frac{2\pi}{32}$ with this single sample cutting shift. On the other hand, if the 17 sample points shift, the phases of the 1000-Hz and 500-Hz waves will shift by $\frac{17\dot{2}\pi}{16}(\text{mod } 2\pi) = \frac{2\pi}{16}$ and $\frac{34\pi}{32}$, respectively. Therefore, the values of the relative phase information for different cutting positions are very different from those of the original cutting position. The phase variation is summarized in Table I. We have partly addressed such variations using a statistical GMM [15][10].

TABLE I
Phase variation related to the frequency and sample points Δ of shifted position.

Period	Frequency	Phase variation
T	$\omega = \frac{2\pi}{T}$	$\frac{\Delta}{T}2\pi$

If we could split the utterance by each pitch cycle, changes in the phase information would be further obviated. Thus, we proposed a new extraction method that synchronizes the splitting section with a pseudo-pitch cycle [12]. With respect to how to unite the cutting sections in the time domain, the proposed method looks for the maximum amplitude at the center of the conventional target splitting section of an utterance waveform, and the peak of the utterance waveform in this range is adopted as the center of the next window. This means that the center of the frame has maximum amplitude in all frames. We expect an improvement over our proposed conventional phase information [9].

III. Combination method and decision method

In this paper, the GMM [16] based on MFCCs is combined with the GMM based on phase information. When a combination of the two methods is used to identify the speaker, the likelihood of the MFCC-based GMM is linearly coupled with that of the GMM based on phase information to produce a new score L^n_{comb} given by

$$L^n_{comb} = (1-\alpha)\, L^n_{MFCC} + \alpha L^n_{phase}, n = 1, 2, \cdots, N, \tag{6}$$

where L^n_{MFCC} and L^n_{phase} are the likelihoods produced by the nth MFCC-based speaker model and phase information-based speaker model, respectively. N is the number of speakers registered and α denotes the weighting coefficients, which are determined empirically. For speaker identification, the speaker (or speaker model) with maximum likelihood is judged to

978-1-5090-4045-2/17/$31.00 ©2017 IEEE

be the target speaker. For speaker verification, GMM with cohortbased normalization [10] was used.

IV. Experiments

A. Database and speech analysis

We evaluate our proposed method for speaker recognition using the NTT (Nippon Telegraph and Telephone) database [9][10] with different speaking rate speech of different sessions, the JNAS (Japanese Newspaper Article sentence) database [17] with short duration speech and noisy speech, and NIST SRE 2003 database [18] with telephone speech.

The NTT clean database consists of recordings of 35 speakers (22 males and 13 females), collected in five sessions over 10 months (1990.8, 1990.9, 1990.12, 1991.3, and 1991.6) in a sound-proof room. To train the models, the same five sentences were used for all speakers in one session (1990.8). These sentences were uttered at a normal speaking rate. Five different sentences at each of the other four sessions were uttered at normal, fast, and slow speaking rates and used as test data. In total, the test corpus consisted of 2100 trials ($5 \times 4 \times 3 \times 35$) for speaker identification. The average duration of the sentences was approximately four seconds.

The JNAS corpus consists of the recordings of 270 speakers (135 males and 135 females). To train the speaker models, 5 clean sentences for clean test data and 10 clean sentences for noisy test data were used for each speaker. About 90 other sentences were used as test data. To obtain the noisy speech, we added stationary noise (in a computer room) and non-stationary noise (in an exhibition hall) from the JEIDA Noise Database [19] to the test speech with average signal-to-noise (SN) ratios of 20 dB and 10 dB. In total, the test corpus consisted of about 24,000 (90×270) trials for each condition. The average duration of the sentences was approximately 3.5 seconds. For a clean condition, we also used short duration test data by cutting whole utterance into 2, 1 and 0.5 seconds, in addition to whole one. For speaker model with short duration training dataset, 5, 10, 15 and 20 seconds speech segments were used.

The NIST 2003 SRE database consists of recordings of 356 speakers (149 males and 207 females) for training, recorded in multiple conditions which include six transmission methods (CDMA, LAND, GSM, TDMA, CELLULAR and UNK), multiple telephones, multiple places, etc. Almost all the data for every speaker were recorded by different environments. Test data includes 3284 trials. Therefore, this speaker verification task is very difficult.

The input speech was sampled at 16 kHz. A total of 25 dimensions (12 MFCCs, 12 ΔMFCCs and Δpower) were calculated every 10 ms with a window of 25 ms. A spectrum with 128 components consisting of magnitude and phase was calculated by DFT for every 256 samples. Phase information was calculated every 5 ms with a window of 12.5 ms. For phase information, we used the first 12 phase components (24 feature parameters in total), that is, from the first to the 12th component of the phase spectrum (frequency range: 62.5 Hz - 750 Hz), which achieved the best identification performance among all the other sub-band frequency ranges [10]. For the proposed pseudo-pitch-synchronized phase information extraction method, the range for searching the peak amplitude point is 2.5 ms (half of the frame shift) for JNAS database, and 2.0 ms for NTT database and NIST SRE 2003 database.

B. Speaker recognition results

We conducted a speaker identification experiment on the NTT database and JNAS database using phase information. GMMs with 32 mixtures for NTT database and 128 mixtures for JNAS database were used as speaker models. Speaker verification was also conducted on NIST SRE 2003 database, and mixture number of GMMs was 256. The robustness of our proposed method was evaluated on speaker recognition for utterance with different speaking rate (NTT database), short duration training dataset and test set (JNAS database), noisy speech (JNAS database), telephone speech with channel variability (NIST SRE 2003 database).

TABLE II
Speaker identification results for NTT database with different speaking rate speech (%).

Feature	Speaking rate			
	Normal	Slow	Fast	Average
MFCC	97.3	95.4	95.6	96.1
Conv. phase	74.7	72.1	73.0	73.4
Syn. phase	80.6	74.6	77.3	77.5
MFCC+Conv. phase	98.7	97.1	98.1	98.0
MFCC+Syn. phase	99.0	97.9	98.4	98.4

1) Speaker identification results for different speaking rate speech: The speaker identification results for different speaking rate speech (NTT database) obtained from individual methods and combination methods are shown in Table II. In Table II, "Conv. phase corresponds to the conventional pitch nonsynchronized phase extraction method [10] and "Syn. phase denotes the pseudo-pitch-synchronized phase information [12]. The average recognition rate of the conventional phase (pitch nonsynchronized phase) is 73.4%. On the other hand, by using the pseudo-pitch-synchronized method, the average performance of the newly proposed extraction technique is improved to 77.5%. By combining the pseudo-pitch-synchronized phase with MFCC, the error reduction rate compared with MFCC is 59.0% (96.1% to 98.4%). This improvement is due to phase information including rich vocal source information. And in the improvement for every speed utterances, these improvements have a relationship with adjusting the cutting position to be roughly synchronized the pitch period.

2) Speaker identification results for JNAS with short duration training data and test data: Speaker results for short duration test data are shown in Table III. The performance of individual method based on phase information or MFCC was degraded significantly. Identification rates were improved from MFCC by the combination method. For 0.5 sec. utterances, identification rates were improved from 66.2% to 79.8% (40.2% error reduction). The identification rate was over 90%

978-1-5090-4045-2/17/$31.00 ©2017 IEEE

TABLE III
SPEAKER IDENTIFICATION RESULTS FOR JNAS DATABASE WITH SHORT DURATION TEST DATA: 5 SENTENCES (ABOUT 17.5 SECONDS OF VOICED SOUND) / SPEAKER MODEL (%).

Feature	Whole	2 sec	1 sec	0.5 sec
MFCC	95.1	89.8	81.6	66.2
Conv. phase	80.3	66.8	51.2	36.1
Syn. phase	90.8	79.3	64.0	46.2
MFCC+Conv. phase	97.8	94.6	88.7	75.7
MFCC+Syn. phase	98.4	96.0	91.1	79.8

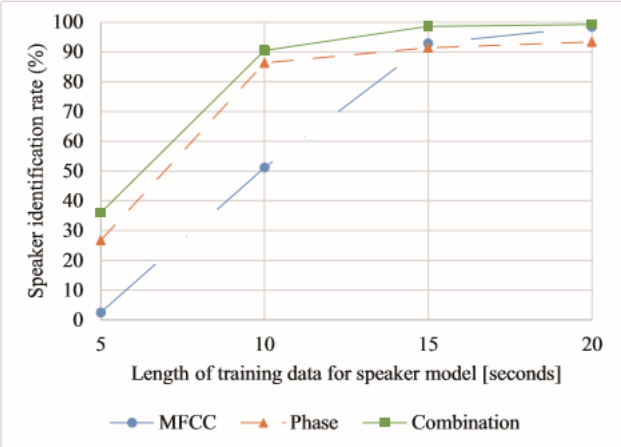

Fig. 1. Speaker identification results for JNAS with short duration training data; test data: 1 sentence (about 3.5 seconds of voiced sound) / trial.

even only using one second utterance by combination method. Therefore, it is necessary to use both of the vocal tract information based MFCC and vocal source information based phase feature.

The speaker identification results of GMMs trained by short duration training data are shown in Fig. 1. The test data is fixed to one sentence with about 3.5 seconds of voiced sound. The length of training data are 5, 10, 15 and 20 seconds, respectively. MFCC is better than phase information when sufficient training data (15 and 20 seconds in this case) is available. When the training data is not sufficient (less than 10 seconds), the performance of MFCC is degraded remarkably and it is much worse than that of the phase information. This indicated that the phase information including rich vocal source information has less variation within speaker (that is, it is not varied very much between different phoneme with same speaker).

And thus unlike the models based MFCC which relative large data is required to learn the vocal tract information, the models based on phase feature including rich vocal source information [10] can be learned well using relative little data. The performance was almost kept even only 10 seconds was used for model training. By combining MFCC with the phase information, the speaker identification performance was remarkably improved from 51.2% of MFCC to 90.5%.

3) Speaker identification results for noisy speech: The speaker identification results obtained from the individual methods and the combination methods are shown in Table IV. Speaker recognition performance is degraded significantly in noisy environments. To address this problem, we cut the 40% of frames with the lowest speech power, because of adopting only higher Signal to Noise Ratio (SNR) sections. This means we use 60% of the frames in the speech. This method obtains the power from all frames in the speech, sorts the values, and then leaves the lower 40% unused. We apply this method only to test data, not to the training data. Comparing with the results using all frames, the results using the 40% of frames cut in the speech were improved in all conditions. The reason was that the unreliable likelihood of frames with low power was removed. Next, we focus only on the results using the 40% of frames cut. The proposed pseudo-pitch-synchronized phase information outperformed the conventional phase information in almost all cases. This means that the proposed method can catch the pseudo-pitch accurately by searching amplitude peaks in all frames, and more effective phase information was extracted even in the noisy environments. On average, comparing with the conventional phase information, the speaker identification rate improved from 55.3% to 63.3% ("Syn. phase). That is, the average error reduction rate is 17.9%. It is interesting that MFCC is more robust for non-stationary noise than for stationary noise, but that the opposite is true with phase information. Therefore, these features are complementary to each other. When the MFCC-based method is compared with the combination of MFCC and pseudo-pitch-synchronized phase, the speaker recognition rate is improved from 55.0% to 75.9% (a relative error reduction of 46.4%). The optimal weight defined in Eq. 6 was determined empirically.

4) Speaker verification results for telephone speech with channel variability: The Eual Error Rates (EERs) of speaker verification based on MFCC, phase and the combination systems are shown in Table V. The EER of pesudo-pitch-

TABLE IV
SPEAKER IDENTIFICATION RESULTS IN NOISY ENVIRONMENTS (%) ("STAT" MEANS STATIONARY NOISE IN A COMPUTER ROOM; "NON-STAT" MEANS NON-STATIONARY NOISE IN A EXHIBITION HALL; ♯ SPEAKERS: 270).

	using 40% cut frames all in the speech			all frames
	Noise conditions		Ave.	Ave.
	10 dB stat/non-stat	20 dB stat/non-stat		
MFCC	35.0 / 42.3	61.5 / 81.3	55.0	31.8
Conv. phase	52.4 / 24.5	76.4 / 67.7	55.3	47.2
Syn. phase	65.4 / 23.4	86.4 / 78.0	63.3	55.4
MFCC+Conv. phase	60.3 / 53.1	87.9 / 92.1	73.4	55.1
MFCC+Syn. phase	67.0 / 52.3	90.1 / 94.2	75.9	61.6

978-1-5090-4045-2/17/$31.00 ©2017 IEEE

TABLE V
EQUAL ERROR RATE OF SPEAKER VERIFICATION BASED ON MFCC, PHASE AND THE COMBINED SYSTEMS EVALUATED ON NIST SRE 2003 (%).

MFCC	16.1
Syn. phase	20.1
MFCC+Syn. phase	13.6
Residual phase[2]	22.0

synchronized phase is 20.1%, and it is better than that of residual phase (22.0%) proposed by Murty et al. [2]. By combining our proposed phase information with MFCC, the EER (13.6%) is much better than that of MFCC (16.1%). The results indicated that our proposed phase information worked well even for telephone speech with channel variability.

V. CONCLUSIONS

In this paper, we summarized the effective pseudo-pitchsynchronized phase information with maximum amplitude synchronization for speaker identification for different speaking rate speech and noisy speech, and added the evaluation of speaker identification for short duration speech and speaker verification for telephone speech with channel variability. The proposed pseudo-pitch-synchronized phase information outperformed than our proposed conventional pitch not-synchronized phase in almost all cases. The combination of MFCC and the pseudo-pitch-synchronized phase was robust for various speaking styles, environments, speech duration and channel variability. and the related speaker identification error reduction rates were about 50% compared with MFCC for almost cases. The results indicated that MFCC and the pahse information are complementary to each other. For noisy speech and speaker models training by short duration speech, the performance of individual phase information was even significantly better than that of MFCC. We consider that the phase information is very useful for speech processing.

VI. ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (No. 61771333), JSPS KAKENHI Grant Number 16K12461, and the National Basic Research Program of China (No. 2013CB329301).

REFERENCES

[1] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech Audio Processing*, vol. 7, no. 5, pp. 569–586, 1997.

[2] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[3] T. Drugman and T. Dutoit, "On the potential of glottal signatures for speaker recognition," in *INTERSPEECH 2010, Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September*, 2010, pp. 2106–2109.

[4] N. Zheng, L. Tan, and P. C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Processing Letters*, vol. 14, no. 3, pp. 181–184, 2007.

[5] P. Rajan, S. H. K. Parthasarathi, and H. A. Murthy, "Robustness of phase based features for speaker recognition," 2009.

[6] M. K. K. Jia, J. Epps, E. Ambikairajah, and E. H. C. Choi, "Ls regularization of group delay features for speaker recognition," in *INTERSPEECH 2009, Conference of the International Speech Communication Association, Brighton, United Kingdom, September*, 2009, pp. 2887–2890.

[7] R. M. Hegde, H. A. Murthy, and G. V. R. Rao, "Application of the modified group delay function to speaker identification and discrimination," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings*, 2004, pp. I–517–20 vol.1.

[8] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker recognition by combining mfcc and phase information," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[9] L. Wang, S. Ohtsuka, and S. Nakagawa, "High improvement of speaker identification and verification by combining mfcc and phase information," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4529–4532.

[10] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE Transactions on Audio Speech Language Processing*, vol. 20, no. 4, pp. 1085–1095, 2012.

[11] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing speech detection using modified relative phase information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 660–670, 2017.

[12] K. Shimada, K. Yamamoto, and S. Nakagawa, "Speaker identification using pseudo pitch synchronized phase information in voiced sound," *Proc. on APSIPA ASC 2011*, 2011.

[13] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, and M. Iwahashi, "Dnn-based amplitude and phase feature enhancement for noise robust speaker identification," in *INTERSPEECH 2016, Conference of the International Speech Communication Association, September*, 2016, pp. 2204–2208.

[15] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining mfcc and phase information in noisy environments," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4502–4505.

[16] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models ," *Speech Communication*, vol. 17, no. 12, pp. 91–108, 1995.

[17] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan*, vol. 20, no. 3, pp. 199–206, 1999.

[18] "http://www.itl.nist.gov/iad/mig//tests/sre/2003/index.html."

[19] "http://research.nii.ac.jp/src/en/jeida-noise.html."