LYRIC RECOGNITION IN MONOPHONIC SINGING USING PITCH-DEPENDENT DNN

Dairoku Kawai, Kazumasa Yamamoto and Seiichi Nakagawa

Department of Computer Science and Engineering, Toyohashi University of Technology, Japan
{kawai, kyama, nakagawa}@slp.cs.tut.ac.jp

ABSTRACT

One of the difficulties in sung speech recognition is the small distance in an acoustic space between phonemes in sung speech. Therefore we considered clustering the speech based on a pitch (fundamental frequency F0) and creating a larger distance between the phonemes. In addition, we considered a two-stage training method of DNN-HMM: the first stage is trained by using conventional acoustic features like MFCCs, and the second stage is re-trained by augmented features with a pitch feature. We expected to train pitch information more explicitly in the second stage of the training and obtained a relative improvement of 9% as expected.

Index Terms— lyrics recognition, pitch information, stepwise approach, DNN-HMM

1. INTRODUCTION

There are two problems facing lyric recognition in singing; accompaniment music removal and speech recognition. To simplify the problem, we focus on speech recognition. Previous researches [1, 2, 3, 4, 5] for lyric recognition were mainly conducted on closed language condition. Wang et al. [1] attempted lyric recognition using a read-speech database and achieved high accuracy (93.1%) by imposing a strong language constraint, including the testing texts of lyrics, that is, very low perplexity.

On the other hand, the study of large vocabulary sung-speech recognition under an open condition is difficult because of the lack of sung-speech database and differences between sung- and spoken-speech in acoustic and language properties. Mesaros et al. [6] used lyrics language models (LMs) and MLLR-adapted GMM-HMM-based acoustic models (AMs) and showed word accuracy of 12.4% in male and female monophonic English singing (it corresponds to a word accuracy of 35.2% replicating in our database [7]). McVicar et al. [8] leveraged repeated lyric phrases and formed a consensus transcription by integrating the repeated portion of the lyrics. For the male and female monophonic English singing, their system showed a word accuracy of 9.5%. In our previous work [7], to deal with phoneme lengthening, we added variational pronunciations to a pronunciation lexicon. In addition, to deal with a lack of sung-speech, we generated pseudo sung-speech from spoken-speech using a neural network based voice transformation. Our system showed a word accuracy of 59.0% in male monophonic Japanese singing.

The quality of the estimated vocal tract transfer function depends on a pitch (fundamental frequency F0) [9]. Pitch information is complementary to energy information in spoken-speech recognition and effective in discriminating between voiced and unvoiced speech [10].

Mesaros et al. examined MFCCs calculated from the phoneme /m/ sung by a male singer with a descending scale of fundamental frequency from 415Hz to 208 Hz [6]. They showed that 3rd MFCC was affected by the variation in pitch. Ozeki et al. showed that sung-speech recognition becomes more difficult as pitch becomes higher [11]. Sung-speech has a peak in the spectrum envelope around 2.8kHz, which is referred to as singing formant [12]. Tatsumi et al. compared the sung and spoken spectrum envelopes and showed that sung-speech neutralized the first two formant frequencies [13].

The rest of this paper is organized as follows: In Section 2, we describe our acoustic and language database. In Section 3, we present our previous approach. In Section 4, we compare the Gaussian distributions of three types of speaking styles: read-speech, spontaneous-speech, and sung-speech, then we describe the reason why sung-speech recognition is difficult. In Section 5, we present our augmented approach, and in Section 6, we show its recognition results. We conclude the paper in Section 7.

2. DATABASE

Our constructed database list is shown in Table 1. We collected 130k pieces of Japanese lyrics texts uploaded by users in Piapro, a lyrics database [7].

We collected 40 pieces of music sung by 40 males uploaded by users in Piapro for acoustic analyses on sung speech and training/adapting AMs [7]. We collected seven pieces of commercial Japanese popular music sung by seven male singers for the test set of lyrics recognition [7]. These music vocal tracks were extracted by taking the difference between the original sound and the accompaniment track using Utagoe Rip [14]. We used the ASJ+JNAS [15, 16] for acoustic analyses on read speech. We used the Corpus of Spontaneous Japanese (CSJ) [17] for acoustic analyses on spontaneous speech and training initial AMs. We recorded pairs of 15 pieces of the read-speech of lyrics and 15 pieces of the sung-speech of the lyrics from seven people with voice training experience for acoustic analyses.

Table 1. Constructed database

(a) Language database

Title	Number of words
Piapro (130k lyrics)	28.6M

(b) Speech database

Title	Num. of speakers	Time length
Sung speech for testing [7]	7	19:01
Spontaneous speech [7]	797	122 hours
Male read speech [15, 16]	133	33 hours
Female read speech [15, 16]	164	44 hours
Sung speech [7]	40	1:39:28
Read speech of parallel data [7]	7	8:59
Sung speech of parallel data [7]	7	25:12

3. LYRIC RECOGNITION SYSTEM

3.1. N-gram Language Model [7]

We used Palmkit (http://palmkit.sourceforge.net) to make word-based n-gram LMs and Witten-Bell smoothing for insufficient n-grams. We trained word-based 3-gram language model (LM) based on lyrics corpora. The vocabulary of the LM is restricted to the top 20k most frequently appearance words. The LM has a 2% OOV rate and perplexity of 113 on the lyrics of the test set.

3.2. Pronunciation dictionary [7]

Insertion errors caused by phoneme lengthening often appear in the form of consecutive vowels. We added automatically modified pronunciation to the pronunciation lexicon in order to capture longer uttered vowel. For instance, the word "cho cho (butterfly)" can be extended as shown in Table 2. In this case, the number of extended pronunciations was eight ($= 2^3$). We obtained the recognition performance gain with the augmented pronunciation dictionary [7].

Table 2. Example of pronunciation extension

Literalization (P. butterfly)	chō	cho
Pronunciation (original)	cho u	cho
Variational Pronunciation (extension)	cho u u cho u ⋮ cho o u u	cho cho o cho o

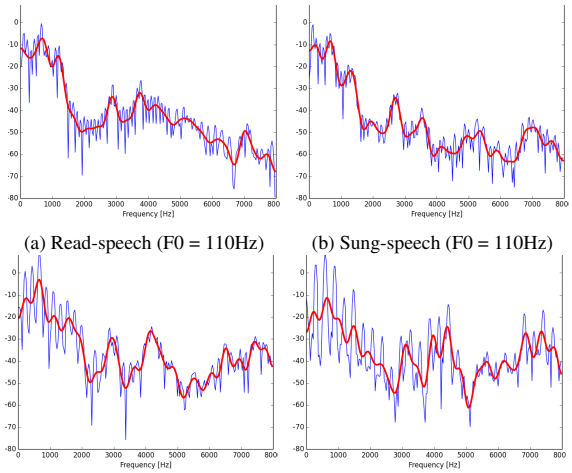
3.3. DNN-HMM

The DNN consists of five layers; the input layer has 429 units, the three hidden layers have 1024 rectified linear units each, and the output layer has 580 units. The inputs are 11 frames of 39 dimensional features: 12 MFCCs, 12 delta MFCCs, 12 delta-delta MFCCs, log energy, delta log energy, and delta-delta log energy. We also used two or four-dimensional features with an additional pitch feature. This leads to 431 units or 433 units for the input layer. The number of output layer units corresponds to the context-independent acoustic states of the HMMs: five states $\times$ 116 syllables. Each Japanese syllable corresponds to a mora, which is a suitable unit for sung speech recognition. The DNN was trained by fine-tuning without pre-training. To make a singing-adapted DNN, we considered two types of training. One is training simply using a large amount of spontaneous-speech data and a small amount of sung-speech data. The other is stepwise training, which is explained in Section 5.

4. SUNG SPEECH ANALYSIS

We compared spectra of the vowel /a/ of read-speech and sung-speech uttered by the same speaker in Figure 1. These spectra are obtained from read-sung speech parallel data which we collected. We found that higher-order frequencies become larger as the pitch becomes higher. In addition, the spectral envelope appears more clearly as the pitch becomes lower and strongly depends on a pitch. Therefore, we expect to model the vocal tract characteristics more accurately as we classify the speech by pitch.

Table 3 shows Bhattacharyya distance between Gaussians for vowels (see [7] for more details) where the Bhattacharyya distance given Gaussian distributions $\mathcal{N}_1(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$ and $\mathcal{N}_2(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$ is defined as follows:



(a) Read-speech (F0 = 110Hz) (b) Sung-speech (F0 = 110Hz)

(c) Sung-speech (F0 = 220Hz) (d) Sung-speech (F0 = 296Hz)

Fig. 1. Spectrum of vowel /a/ of read-speech and sung-speech uttered by the same speaker

$$BD = \frac{1}{8}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) + \frac{1}{2} ln(\frac{|\boldsymbol{\Sigma}|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}),$$
$$\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}. \quad (1)$$

The largest distance between vowels is in read speech, followed by spontaneous and sung speech. The bigger the distance is, the easier their recognition is. This result shows that sung speech is the most difficult to recognize. The reason why the distance between vowels in sung-speech becomes smaller is that the first two formant frequencies in sung speech are neutralized [13] and that MFCCs emphasize lower frequencies. As shown in Figure 1, the spectrum of sung-speech differs depending on its pitch. In addition, sung-speech has a wide pitch range as shown in Figure 2. Therefore, by splitting sung speech into three pitch ranges, the distance between vowels become larger than before and also become similar to or larger than that of spontaneous-speech. This indicates that the pitch information has effective information for sung speech recognition. As shown in Table 4, the determinants of pitch dependent Gaussians of sung speech are smaller than the pitch independent ones, where the value of determinant corresponds to the size of the variance of the distribution. This result also indicates that clustering by pitch works well. In read speech recognition, the use of the pitch feature resulted in limited improvement [10]. The pitch feature is effective in discriminating voiced and unvoiced speech but competes with energy feature. This might be because spoken speech has a narrow pitch range as shown in Figure 2 and there is little spectral change.

For reference, Figure 3 shows a visualization of 2 dimensional vector space which is dimensionality reduced from the 12 dimensional MFCCs by using t-SNE [18]. The t-SNE is a dimensionality reduction method that operates by minimizing the divergence between two distributions: a distribution that expresses pairwise similarities of the given high-dimensional data points and a distribution that expresses pairwise similarities of the corresponding unknown low-dimensional data points. As shown in Figure 3, the variance of sung speech is the smallest (see the top of the figure) and vowels in sung speech are neutralized. Additionally, we found that several sets

of data points in sung, which were very different from each other, were separated by pitch-based clustering (see the black circle in Figure 3). These results are consistent with the result of analyses of Gaussian distributions.

Table 3. Bhattacyaryya distance between vowels (the average for 5 vowels)

(a) Three speaking styles : read, spontaneous, and sung speech

	Read	Spontaneous	Sung
Ave.	1.46	1.01	0.82

(b) Sung speeches clustered by three pitch ranges

	Sung		
	∼173Hz	174∼260Hz	261Hz∼
Ave.	1.09	1.19	1.30

Table 4. Determinants of full covariance matrix (the average for 5 vowels)

(a) Three speaking styles : read, spontaneous, and sung speech

	Read	Spontaneous	Sung
Ave.	3.08E-03	7.09E-03	2.54E-03

(b) Sung speeches clustered by three pitch ranges

	Sung		
	∼173Hz	174∼260Hz	261Hz∼
Ave.	2.12E-03	1.59E-03	8.96E-04

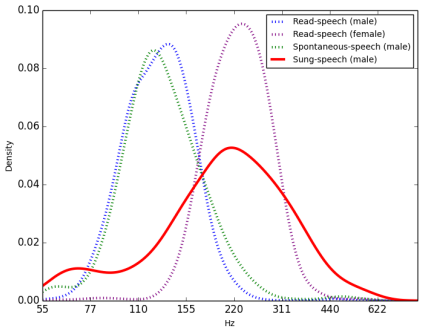


Fig. 2. Pitch histograms of three speaking styles

5. AUGMENTED METHOD

5.1. Pitch feature

Conventional speech recognition systems have generally used the acoustic features of MFCCs as the vocal tract characteristic and energy as the source characteristic. In our study, in addition to these features, we consider a pitch feature. First of all, we extract fundamental frequency F0 from the speech waveform with a technique using autocorrelation function [19]. By using the estimated F0, we define three types of pitch features as bellow:

$$\mathrm{Vo}(F0) = \begin{cases} 1 & \text{Detectable } F0 \\ 0 & \text{Undetectable } F0 \end{cases} \quad (2)$$

$$\mathrm{LogF0}(F0) = \begin{cases} log(F0) & F0 > 0Hz \\ 0 & \text{Undetectable } F0 \end{cases} \quad (3)$$

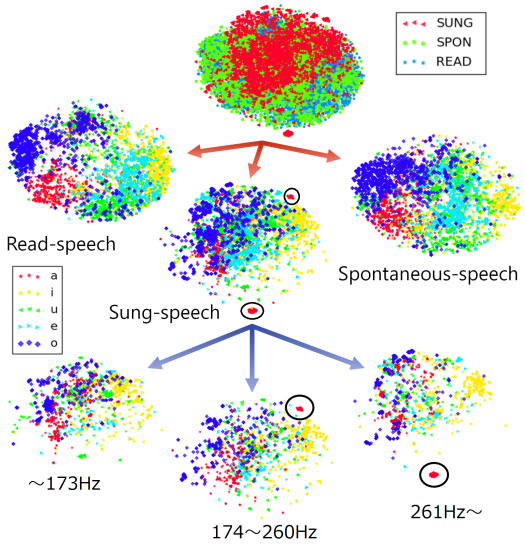


Fig. 3. Visualization of 12 dimensional MFCCs by t-SNE

$$\mathrm{V4}(F0) = \begin{cases} (1,0,0,0) & \text{Undetectable } F0 \\ (0,1,0,0) & 55Hz <= F0 < 173Hz \\ (0,0,1,0) & 174Hz <= F0 < 260Hz \\ (0,0,0,1) & 261Hz <= F0 \end{cases} \quad (4)$$

Equation (2) is binary which distinguishes between voiced and unvoiced speech. Equation (3) is the logarithm of F0, and Equation (4) is a one-hot vector which classifies into unvoiced and voiced speeches of three pitch ranges. We determined the boundaries by using 1/3 area sections of the pitch histogram of sung speech and the unvoiced section (see SUNG in Figure 2). Simply training of a DNN-HMM acoustic model by using features which are the concatenation of conventional acoustic features and the pitch feature might be insufficient because of a lack of sung-speech data and less pitch information compared with MFCC.

5.2. Stepwise training

Therefore, as shown in Figure 4, we consider stepwise training in which we train the model in two stages. In the first stage, we use only conventional acoustic features for spontaneous and sung speech, and get parameters which are closer to the optimal weight. At this time, we use a 0 filled dummy pitch instead of actual pitch. In the second stage, we update the parameters by using features which are the concatenation of conventional acoustic features and pitch feature. It might affect the effect of using the pitch information of spontaneous speech because spontaneous speech data is greater in size than sung speech. Therefore, we also use the actual pitch of sung speech and the dummy pitch of spontaneous speech on the second stage of stepwise training.

6. EXPERIMENT

6.1. Setup

The word-based n-gram LMs was trained using a lyrics corpus as explained in Section 2. The vocabulary of the LM was restricted

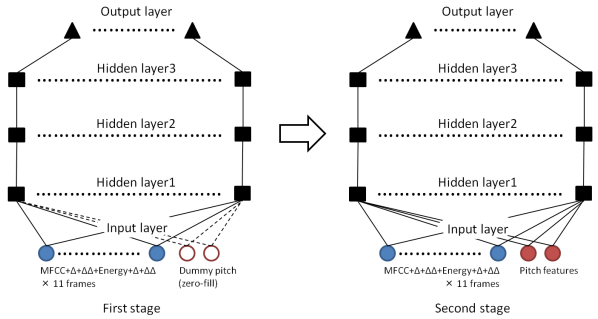


Fig. 4. Stepwise training of pitch information

Table 5. Test set classification (Reverberation/Chorus)

	noR/noC	noR/C	R/noC	R/C
Time length	2:38	6:31	4:33	4:01

to the top 20k most frequent appearance words. The pronunciation dictionary contained original and modified pronunciation in order to capture longer uttered vowels. We used 11 frames of 39-dimensional conventional acoustic feature (13 MFCCs, 13 ΔMFCCs, 13 ΔΔMFCCs; Conv) and a central frame of a two (Vo+logF0) or four (V4) dimensional feature of the proposed pitch feature as described in Section 5. The DNN-HMM acoustic models were trained by using CSJ [17] (797 speakers, 121 hours) and sung speech [7] (40 speakers, 1 hour 39 minutes). The DNN-HMM baseline model is trained by using spontaneous and sung speeches with only a conventional acoustic feature [7]. In comparison, we also used the baseline GMM-HMM-based model [7]. In this paper, we evaluated the DNN-HMM-based acoustic model integrated with pitch feature as follows:

Model A

The model A is trained by using spontaneous and sung speech with a conventional acoustic feature and an actual pitch feature.

Model B

The model B is trained by using spontaneous speech with a conventional acoustic feature and a dummy pitch feature and sung speech with a conventional acoustic feature and an actual pitch feature.

Stepwise

Superscript "stepwise" denotes the model trained in two stages and the subscript number denotes the updated hidden layer index on the second stage where "All" means the all hidden layers. The hidden layer index is counted from the side of the input layer.

We evaluated our test set that contained seven pieces of commercial music sung by seven males, as explained in Section 2. Pitch contours were not normalized to singers or songs. The test set is manually classified into four classes on the basis of whether it contains reverberation (R) or not (noR) and whether it contains back chorus (C) or not (noC) (see Table 5). These effects are present in many types of popular music.

The LM weight on the decoder was chosen from 1, 10, 15, 20, 25, and 30. An insertion penalty was chosen from -30, -20, -10, and 0. We report the recognition results by using the best weight and penalty.

6.2. Result

Table 6 shows the results of LVCSR on the test set. We only discuss the results of the test set noR/noC because the influence of reverberation and chorus tends to cause a false detection of pitch on the test set noR/C, R/noC, and R/C.

Table 6. Word accuracy of LVCSR [%]

AM	Test Set (Reverb./Chorus) noR /noC	noR /C	R /noC	R /C	All
(a) Baseline					
$\text{GMM-HMM}_{\text{SPON}}$	33.6	10.3	9.7	10.5	11.9
$\text{GMM-HMM}_{\text{SPON+SUNG}}$	42.6	26.0	16.4	27.2	26.6
$\text{DNN-HMM}_{\text{SPON}}$	51.2	18.9	20.1	19.7	21.7
$\text{DNN-HMM}_{\text{SPON+SUNG}}$	55.7	31.2	19.8	34.8	32.2
(b) Features:Conv+LogF0+Vo (DNN-HMM)					
A	54.9	31.9	**26.4**	31.8	32.8
B	54.1	31.7	16.4	**38.7**	31.2
$\text{A}^{\text{Stepwise}}_{\text{All}}$	53.7	33.7	21.4	33.8	32.9
$\text{B}^{\text{Stepwise}}_{\text{All}}$	53.7	34.1	21.4	31.8	32.7
(c) Features:Conv+V4 (DNN-HMM)					
A	57.0	31.9	24.8	36.1	33.0
B	52.5	32.0	18.2	33.8	30.7
$\text{A}^{\text{Stepwise}}_{\text{All}}$	**60.7**	33.8	20.4	37.7	**35.6**
$\text{B}^{\text{Stepwise}}_{1}$	56.6	32.6	21.1	33.8	32.7
$\text{B}^{\text{Stepwise}}_{2}$	57.8	32.4	19.2	35.7	34.2
$\text{B}^{\text{Stepwise}}_{1,2}$	57.8	33.5	18.6	34.1	33.6
$\text{B}^{\text{Stepwise}}_{\text{All}}$	**60.7**	**35.2**	21.4	34.4	35.0

The DNN-HMM baseline model showed the accuracy of 55.7% [7]. The model with features of Conv+LogF0+Vo did not improve in performance compared to the baseline model because it is difficult to express pitch class as a 1-dimensional feature such as *log(F0)*. Model A, with features of Conv+V4, outperformed the baseline model (55.7% → 57.0%).

Model B, with features of Conv+V4, did not improve in performance compared to the baseline model. This is because the model B is trained by using augmented features with dummy and actual pitch before the model parameters are not close to the optimal value. The model $\text{A}^{Stepwise}_{All}$ with features of Conv+V4 showed the best accuracy of 60.7% (a 9% relative improvement). We were concerned that the pitch feature of spontaneous speech had a bad affect because spontaneous speech data is of a greater size than sung speech, but the model $\text{A}^{Stepwise}_{All}$ was trained well. The model $\text{B}^{Stepwise}_{All}$, with features of Conv+V4, also showed the best accuracy of 60.7%. This might be because the augmented features were trained after the model parameters were trained sufficiently in the first stage. We compared retraining layers by using the specific hidden layers. The model in which more layers were retrained showed better performance in comparison with the model $\text{B}^{\text{Stepwise}}_{\text{1or2or1,2orAll}}$.

7. CONCLUSION

We incorporated pitch information into conventional acoustic features to create a larger distance between vowels in sung-speech. Therefore we considered three types of pitch features: voiced flag, $log\ F0$, and one-hot vector of pitch class with a four-dimensional vector. We also considered the two-stage training of DNN-HMM to train pitch information more explicitly. As a result of an experiment, our system produced a word accuracy of 60.7% without the case of reverberation and chorus. To the best of our knowledge, this accuracy is the best among all published papers except for ours (the best method was based on $\text{GMM-HMM}_{SPON+SUNG}$, see [6, 7]). In the future work, we will consider the use of repeated lyrics phrases, that is, online language adaptation.

8. REFERENCES

[1] C. Kai Wang, R.-Y. Lyu, and Y.-C. Chiang, "An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker." pp. 1197–1200, 2003.

[2] A. Sasou, M. Goto, S. Hayamizu, and K. Tanaka, "An ARHMM-based speech analysis method and an evaluation of a singing-voice recognition," *Report of IEICE. SP, Speech*, vol. 105, no. 199, pp. 19–24, Jul 2005 (in Japanese).

[3] A. Sasou and M. Goto, "Japan patent, jp4576612b," 2007 (in Japanese).

[4] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. Okuno, "Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals," pp. 257–264, Dec 2006.

[5] T. Hosoya, M. Suzuki, A. Ito, S. Makino, L. A. Smith, D. Bainbridge, and I. H. Witten, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval," pp. 532–535, 2005.

[6] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, 2010.

[7] D. Kawai, K. Yamamoto, and S. Nakagawa, "Speech analysis of sung-speech and lyric recognition in monophonic singing," pp. 271–275, 2016.

[8] M. McVicar *et al.*, "Leveraging repetition for improved automatic lyric transcription of popular music," pp. 3141–3145, 2014.

[9] A. de Cheveigné and H. Kawahara, "Missing data model of vowel identification," *J Acoust Soc Am*, vol. 105, pp. 3497–3508, 1999.

[10] N. Kitaoka, D. Yamada, and S. Nakagawa, "Speaker independent speech recognition using features based on glottal sound source," *Proc. Int. Conf. on Spoken Language Processing*, pp. 2125–2128, 2002.

[11] H. Ozeki, T. Kamata, M. Goto, and S. Hayamizu, "The influence of vocal pitch on lyrics recognition of sung melodies," *Proc. of Autumn Meeting of the ASJ*, 2003 (in Japanese).

[12] J. Sundberg *et al.*, *The Science of Musical Sounds*. Northern Illinois University Press, 1989.

[13] M. Tatsumi and H. Fujisaki, "Acoustic characteristics of sung vowels," *IEICE technical report. Speech*, pp. 55–60, 1978.4 (in Japanese).

[14] TODAKEN, "Utagoe lip." (accessed 24th Spt 2014).

[15] T. Kobayashi, S. Itabashi, S. Hayamizu, and T. Takezawa, "ASJ continuous speech corpus for research," *The Journal of the Acoustical Society of Japan*, vol. 48, no. 12, pp. 888–893, Dec 1992 (in Japanese).

[16] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," pp. 722–725, 1998.

[17] K. Maekawa *et al.*, "Spontaneous speech corpus of Japanese." pp. 947–952, 2000.

[18] L. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[19] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," pp. 97–110, 1993.