INTERSPEECH 2009 BRIGHTON

Observation of empirical cumulative distribution of vowel spectral distances and its application to vowel based voice conversion

Hideki Kawahara[1], *Masanori Morise*[2], *Toru Takahashi*[3], *Hideki Banno*[4], *Ryuichi Nisimura*[1] *and Toshio Irino*[1]

[1]Department of Design Information Sciences, Wakayama University, Wakayama, Japan
[2]Department of Media Technology, Ritsumeikan University, Japan
[3]Graduate School of Informatics, Kyoto University, Japan
[4]Department of Information Engineering, Meijo University, Japan

kawahara@sys.wakayama-u.ac.jp

Abstract

A simple and fast voice conversion method based only on vowel information is proposed. The proposed method relies on empirical distribution of perceptual spectral distances between representative examples of each vowel segment extracted using TANDEM-STRAIGHT spectral envelope estimation procedure [1]. Mapping functions of vowel spectra are designed to preserve vowel space structure defined by the observed empirical distribution while transforming position and orientation of the structure in an abstract vowel spectral space. By introducing physiological constraints in vocal tract shapes and vocal tract length normalization, difficulties in careful frequency alignment between vowel template spectra of the source and the target speakers can be alleviated without significant degradations in converted speech. The proposed method is a frame-based instantaneous method and is relevant for real-time processing. Applications of the proposed method in-cross language voice conversion are also discussed.

Index Terms: voice conversion, STRAIGHT, vowel structure, line spectral pair, vocal tract length

1. Introduction

A simple and fast voice conversion method is introduced, inspired by recent findings in human speech perception. Humans can adapt to speakers very quickly even when the speaker is unknown. A comprehensive test of monosyllabic speech perception indicated that humans can adapt to any unknown speakers only with exposure of up to five monosyllables [2], as shown in Fig. 1. It should also be noted that the baseline performance of vowel identification for a wide variety of vowel deformations in terms of vocal tract length and glottal pulse rate well exceeds currently available automatic speech recognition (ASR) systems [3].

These may suggest that humans are extracting underlying universal structure of speech sounds spoken by a specific speaker with a rather short exposure to his/her utterances [4]. The most likely scenario is that important clues for such adaptation are extracted from vowels and employed to deform and relocate the universal structure [2, 5].

The proposed method tries to recapitulate this function by designing the mapping function for voice conversion as a deviation-preserving deformation function. Empirical cumulative distribution of spectral distances between vowel segments is used to design this mapping function and its constituent mapping functions.

The other motivation of the proposed method is based on demands in game applications. Using a game player's voice as a game character's voice can be very effective for enhancing the reality of scenes and a player's immersion. This second source of motivation inevitably introduces large variability in recording/digitizing conditions and potential asymmetry between the source and the target speech for voice conversion.

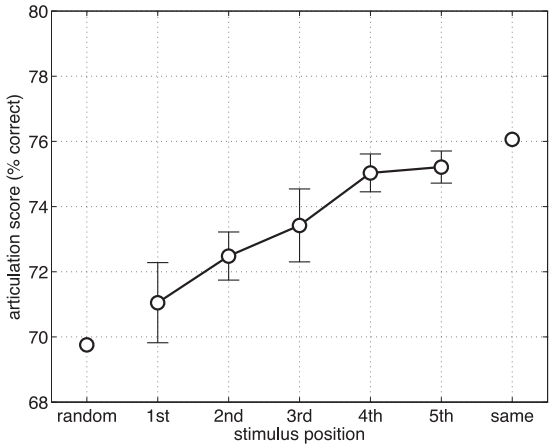

Figure 1: *Perceptual adaptation to speakers in Japanese isolated syllable identification task by telephone intelligibility testing crew members. In the "random" condition, all monosyllables are randomly selected from different speakers. In the "same" condition, all monosyllables are spoken by one speaker. In the center results are for "blocked" condition, where up to five monosyllables spoken by one speaker are presented in one contiguous test block followed by the next block spoken by a different speaker. (Figure 6 of reference [2] is arranged and redrawn. Please refer to [2] for details.)*

For recording/digitizing conditions, difference in sampling rates introduces difference in frequency range, and difference in type of microphones introduces spectral differences in the low frequency ends due to proximity effects of directional microphones and low-cut filters for compensation.

The potential asymmetry is typically found in the case of changing a character's voice. In this case, the target voice is the user's voice and the source is the character's voice. Heavy manual labeling and editing are already involved in the authoring process of the game character. Consequently, the source voice tends to be rich in annotation. On the other hand, it is not practical to ask users to edit and provide label information for their voice before starting a game. Therefore, annotation of the target voice tends to be poor.

2. Background

The proposed method was designed taking these considerations into account. This section introduces the background of the proposed method.

2.1. TANDEM-STRAIGHT

The proposed voice conversion procedure is implemented using the TANDEM-STRAIGHT system [1] as a building block of the procedure. This section briefly introduces the basic structure of the system and describes the essential features in its spectral estimation.

TANDEM-STRAIGHT and STRAIGHT [6], its predecessor, are basically channel VOCODERs [7]. They decompose input speech into three components, fundamental frequency (F0), aperiodicity spectrogram and F0-adaptively smoothed spectral envelope (STRAIGHT spectrum). These parameters are (if necessary, manipulated and then) fed into a synthesis subsystem to reproduce synthesized speech sound. The F0 information is used to set the repetition rate of the excitation pulse and the STRAIGHT spectrum is used to design the filter to be excited by the pulse.

Spectral estimation in TANDEM-STRAIGHT consists of two stages to eliminate the effects of periodic excitation on the estimated spectrographic representation. The first stage (TANDEM-stage) provides a stable power spectral representation that doesn't show temporal variations irrespective of positioning of the analysis time window [8]. It is made possible by averaging two power spectra calculated using two time windows temporally separated one half pitch period. The second stage (STRAIGHT-stage) eliminates periodic spectral variations due to periodicity by using a spectral smoother adaptively designed based on F0 information. A compensating digital filtering on the frequency domain is also applied to make this smoothing process meet the condition for consistent sampling [9]. This interference-free spectral representation plays an important role in the proposed method.

2.2. Perception of speaker identity

Vowels are rich carriers of a speaker's identity. Allocation of formant frequencies of vowels provides clues to the speaker's size and vocal tract shape [10]. F0s of vowel segments provide information about the gender of the speaker. F0s also provide age information.

3. Outline of proposed method

Figure 2 shows a schematic diagram of the proposed method. The method consists of two stages, the design phase and the conversion phase.

3.1. Design phase

The goal of this stage is to design constituent mapping functions based on vowel prototypes of the source and the target speakers. It is crucial to construct good vowel prototypes from the given examples. It is also important to find a relevant strategy to design a mapping function for each vowel prototype pair.

3.1.1. Selection of representative frames

Instead of using an averaged spectrum or whole spectral frames in a labeled vowel segment, one representative spectral frame is selected as the spectrum that represents the vowel segment. Selection of the frame is based on the cumulative distribution of spectral distance between the frame in question and the other frames in the same vowel segment. The frame that has the largest number of close neighbors is selected as the representative frame of each vowel segment. This process is applied to all vowel segments.

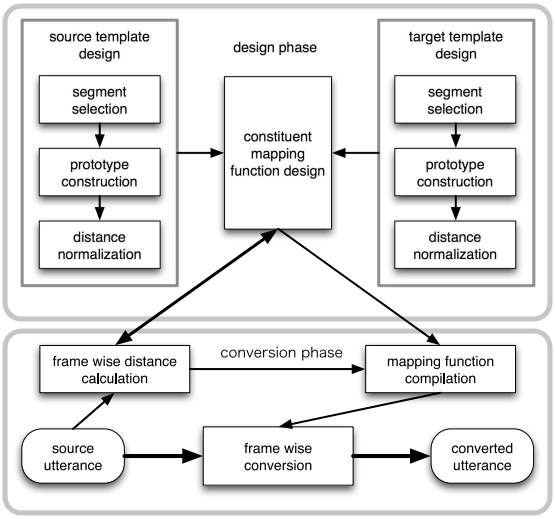

Figure 2: *Schematic diagram of proposed procedure.*

3.1.2. Compensation of static biases

Due to different acquisition processes, speech materials from different sources tend to have different frequency responses. Relatively low-order cosine expansion of averaged power spectra represented in terms of dB scale on the frequency axis is used to estimate static bias due to this factor. Voice register and gender also contribute to the global shape of vowel spectra. These biases are memorized (recorded) and removed (cancelled) before calculating vowel prototypes.

3.1.3. Vowel prototypes

Vowel prototypes are averaged STRAIGHT spectra of representative frames of representative segments. The representative segments are selected based on cumulative distribution of spectral distances between the representative frame of a vowel segment and that of the other vowel segment belonging to the same vowel category (intra-category distribution), as well as the cumulative distribution of spectral distance between the segment and the vowel segment belonging to different vowel categories (inter-category distribution). Top N (N depends on available amount of data) segments are selected for each category. They are selected based on a criterion to maximize differences between intra-category and inter-category distributions.

3.1.4. Constituent mapping function design

The goal of this subsystem is to design a mapping function that can deform the vowel spectral template of the source speaker to that of the target speaker for each vowel category. The mapping function consists of two sub-functions: a frequency axis mapping function and a spectral framework mapping function. In short, the frequency axis mapping function deforms the frequency axis of the source speaker to align corresponding formant peaks of the source and the target spectra. The spectral framework mapping function adjusts spectral differences after this formant alignment using a spectrally smooth shape. To make the frequency axis alignment, two optional procedures are implemented, manual anchor point assignment and automatic alignment. Details of the latter method are described in another paper [11].

In addition to this constituent mapping function, spectral distance to proximity conversion function is also designed because distribution of spectral distances of intra- and inter- category segments and their boundaries differ depending on vowel categories. This conversion function is designed for spectral distances at the boundaries between intra- and inter-category segments to have the same converted proximity values using the sigmoidal approximation of each cumulative distribution. Proximity values are bounded by 0 and 1.

3.2. Conversion phase

The goal of this phase is to convert each incoming STRAIGHT-spectrum of the source speaker into that of the target speaker in a frame-wise fashion. This is done using composite mapping functions. The functions are compiled based on proximity between each input frame and the source speaker's vowel templates.

3.2.1. Proximity calculation

In this subsystem, first, spectral distances between an input STRAIGHT-spectrum and the source speaker's vowel templates are calculated at each frame. Secondly, these distances are converted to proximity values using the conversion functions that were designed in the design phase.

3.2.2. Mapping function compilation

In this subsystem, proximity values are used to weight the deviation of each constituent mapping function from the average of the mapping functions. This ensures that even in the worst case, when an input spectrum does not match well with any templates, the average mapping can be applied.

3.2.3. Frame-wise conversion of parameters

This subsystem performs actual conversion of STRAIGHT-spectrum, F0 and aperiodicity spectrogram. Spectral conversion is performed using the compiled mapping functions. After this conversion, static biases in spectral shapes are added back. The F0 conversion ratio is determined based on the average of each log-F0 frequency. The frequency axis of the aperiodicity spectrogram is also converted by the frequency axis mapping sub-function of the compiled spectral mapping function. Finally, the averaged difference between cepstral representations of the source and the target templates is compensated using a smoothed cepstram lifter. The underlying idea of this liftering is similar to "global variance compensation" proposed in the literature [12]. These converted parameters are fed into the synthesis subsystem of the TANDEM-STRAIGHT system.

4. Conversion examples

Several conversion patterns were tested and yielded natural sounding converted speech in informal listening tests. This section illustrates operation of the proposed method using a typical pattern of usage, where a user makes a recording of his/her isolated vowel samples for changing prerecorded (and heavily annotated) utterances. To simulate this scenario, we excerpted four sentences from a database with simultaneous EGG recording [13] and with annotations were used as the source utterances. The speaker is a Japanese male university student. Total length of the four utterances is 11.024 s. The original recording sampling rate is 48,000 Hz and down sampled to 16,000 Hz. A directional microphone was used in recording. The target was excerpted from the other database. A sequence of five isolated Japanese vowels was excerpted and used as the target. The speaker is a Japanese male adult. The original sampling rate is

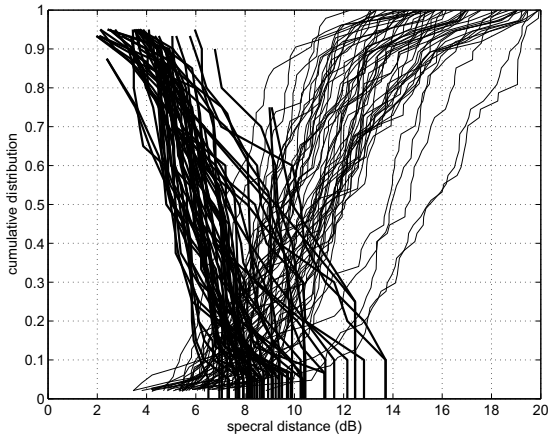

Figure 3: *Empirical cumulative distribution of MFCC distances in source utterances. (thick lines: intra-category distribution, thin lines: inter-class distribution)*

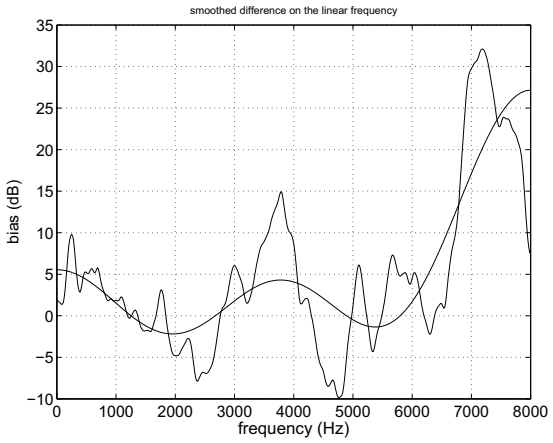

Figure 4: *Static bias on linear frequency axis.*

44,100 Hz and resampled at 16,000 Hz. An omnidirectional microphone was used in recording. Total length of the voiced region is 0.545 s.

Figure 3 shows the observed cumulative distribution of the spectral distances of intra- and inter-category segments of the source utterances. The spectral distance used here is dB distance of MFCC filter bank outputs with mean value alignment. Please note that overlap of cumulative distributions of intra- and inter-category distance is significant.

Figure 4 shows the static bias on the linear frequency axis. The smooth line in the plot represents the estimated bias component. The component is calculated from the difference between the averaged spectrum of the voiced segments. Differences in recording microphones and resampling methods as well as speaker differences are contributing factors to this bias component.

Figure 5 shows cumulative distribution of the distances from the target templates. The blue line, denoted as WO, represents the distances between the source and the target templates without bias compensation. The crossing point of intra- and inter-category distribution is about 0.4, indicating that vowel structure difference is masked by the bias term. The green line, denoted as SS, represents the distances with static bias compensation. The red line, denoted as CV, represents the distance with static bias compensation and frame-wise conversion using the compiled composite mapping function. Finally, the black line,

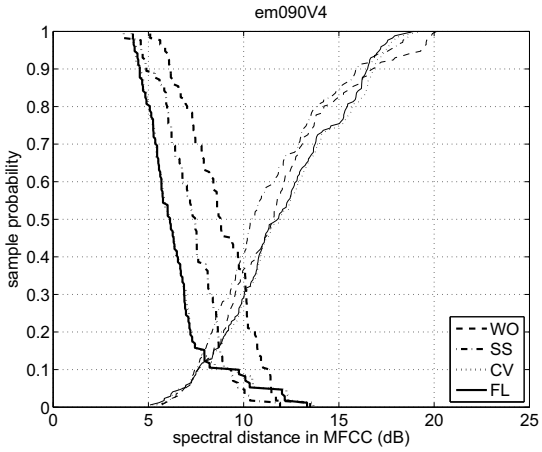

Figure 5: *Cumulative distributions of distances from target templates. Color represents test conditions. (thick lines: intra-category, thick lines:inter-category)*

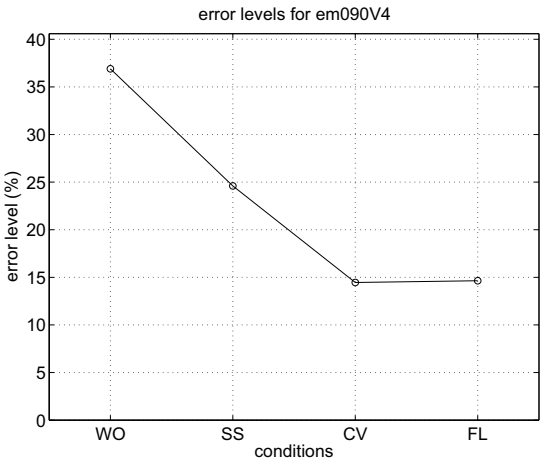

Figure 6: *Classification error at each processing stage.*

denoted as FL, represents the distance after the additional cepstral liftering. Please note that distributions of CV and FL are similar to the average distributions shown in Fig. 3, indicating that the vowel structure is successfully converted.

Figure 6 is a summary of Fig. 5. Vowel identification scores are calculated using the distance distributions in Fig. 5 for setting the best decision boundaries.

The original speech materials and converted sounds are provided as media files in this proceeding. Other conversion examples are also linked to the web page for our Interspeech 2009 presentations [14]. It can be easily seen that the original naturalness is preserved in the converted utterances and the speaker's identity is clearly converted, even though the samples are in Japanese. A series of formal subjective evaluation tests are currently conducted.

5. Discussion

The proposed method is general enough to be applicable to other languages. The more challenging question is on cross-linguistic voice conversion because vowel prototypes are generally different in other languages. However, we are optimistic. Taking into account the fact that the coordinate system to represent vowel prototypes can be defined independently from the represented prototypes, the vocal identity of a specific speaker represented as a deviation from the average voice in one language system can be transformed to the corresponding deviation in the target language. It is an interesting topic for further study.

6. Conclusions

The proposed method makes it possible to convert the voice of the source utterances relying only on isolated vowel samples of the target speaker, while preserving the naturalness of the original utterances. The method is also applicable to real-time processing because the conversion is frame wise.

7. Acknowledgements

This work is partly supported by Grant-in-Aid for Scientific Research (A) 19200017 by JSPS and the CrestMuse project by JST.

8. References

[1] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H., "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation", Proc. ICASSP, 3933–3936, 2008.

[2] Kato, K. and Kakehi, K., "Listener adaptability to individual speaker differences in monosyllabic speech perception", J. Acoust. Soc. Jpn., 44(3):180–186, 1998. [in Japanese]

[3] Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H. and Irino, T., "The processing and perception of size information in speech sounds", J. Acoust. Soc. Am., 117(1):305-318, 2005.

[4] Minematsu, N., "Mathematical evidence of the acoustic universal structure in speech", Proc. ICASSP, I:889–892, 2005.

[5] Minematsu, N., Asakawa, S. and Hirose, K., "Para-linguistic information represented as distortion of the acoustic universal structure in speech", Proc. ICASSP, I:261–264, 2006.

[6] Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction", *Speech Communication*, 27(3–4):187–207, 1999.

[7] Dudley, H., "Remaking speech", *J. Acoust. Soc. Am.*, 11(2):169–177, 1939.

[8] Morise, M., Takahashi, T., Kawahara, H. and Irino, T., "Power spectrum estimation method for periodic signals virtually irrespective to time window position", Trans. IEICE, J90-D(12):3265–3267, 2007. [in Japanese]

[9] Unser, M., "Sampling – 50 years after Shannon", Proceedings of the IEEE, 88(4):569–587, 2000.

[10] Turner, R.E., Walters, T.C. , Monaghan, J.J.M. and Patterson, R.D., "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data", J. Acoust. Soc. Am., 125(4):2374–2386, 2009.

[11] Kawahara, H., Morise, M., Takahashi, T, Banno, H., Nisimura, R. and Irino, T., "Automatic frequency axis alignment based on vocal tract length normalization and phase of vocal tract transfer functions", 2009. [submitted]

[12] Toda, T. and Tokuda, K., "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis", IEICE Trans. Inf. & Syst., E90-D(5): 816–824, 2007.

[13] Atake, Y., Irino, T., Kawahara, H., Lu, J., Nakamura, S. and Shikano, K., "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components", Proc. ICSLP, 907–910, 2000.

[14] http://www.wakayama-u.ac.jp/˜kawahara/IS09demos/