MODELING OF VARIATIONS IN CEPSTRAL COEFFICIENTS CAUSED BY F_0 CHANGES AND ITS APPLICATION TO SPEECH PROCESSING

Nobuaki MINEMATSU *Seiichi NAKAGAWA*
mine@tutics.tut.ac.jp *nakagawa@tutics.tut.ac.jp*

Department of Information and Computer Sciences, Toyohashi University of Technology,
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, Aichi-ken, 441-8580 JAPAN

ABSTRACT

In this paper, the correlation between spectral variations and F_0 changes in a vowel sound is firstly analyzed, where the variations are also compared to VQ distortions calculated in a five-vowel space. It is shown that the F_0 change approximately by a half octave produces the spectral variation comparable to the averaged VQ distortion when the codebook size is the number of the vowels. Next, a model to predict the cepstral coefficients' variations caused by the F_0 changes is built based on the multivariate regression analysis. Experiments show that the generated frame by the model has a remarkably small distance to the target frame and that the distance is almost the same as the VQ distortion with the codebook size being 10 to 20. Furthermore, the model is evaluated separately in terms of a spectral envelope predictor with a given F_0 and a mapping function of feature sub-spaces. It is indicated that, while the models should be built dependently on phonemes and speakers as a spectrum predictor, adequate selection of parameters can enable the speaker/phoneme-independent models to work effectively as a mapping function.

1. INTRODUCTION

In most of the conventional studies in speech processing, the variations in spectral envelopes were assumed to be representing differences among phonemes, phonemic environments, speakers, channels, languages, and so forth. The variations caused by the F_0 changes were already reported using formant-based features in several papers[1]-[3]; nevertheless, they were still considered small enough to ignore in the speech processing techniques.

In recent studies, modeling the variations by the F_0 changes was effectively introduced into speech recognition[4] and speech synthesis[5]. Although these studies could improve the system performance, they didn't show any answer to questions "*How* large are the variations ?" and "Can they be comparable to the distances between different phonemes ?" In this paper, comparing the variations observed in each of Japanese five vowels to the VQ distortions calculated in the five-vowel space, the questions will be answered.

After the analysis, a model to predict the cepstral coefficients' variations is built based on the multivariate regression analysis. Here, F_0 and its derivative (henceforth, ΔF_0) are used as a part of predicting factors. And the model is evaluated separately in terms of a spectral envelope predictor with a given F_0 and a mapping function of feature sub-spaces. In the former evaluation, the prediction performance is calculated using the cepstral distance as a measure of the prediction errors. In the latter, the model, which maps a feature vector from an n-dimensional space into another, is considered to modify the distribution of acoustic features of a phoneme. Hence, the model is expected to increase the averaged distance between the distributions of different phonemes by the modification. Furthermore in this paper, the dependence on speakers and phonemes in both evaluation schemes is also investigated.

2. ANALYSIS OF SPECTRAL VARIATION

2.1. Speech Material

Since this study examines the spectral variations caused by the F_0 changes, speech samples corresponding to *source* (before the change) and to *target* (after the change) are required. Here, we assume that the above two kinds of samples, or frames, can be approximately obtained from two speech segments which are temporally close to one another and satisfy an F_0-related condition as shown in **Figure 1**. In the figure, the local patterns of F_0, **a** and **b**, are similar to those of **a**′ and **b**′ respectively.

In recording the speech samples, since the prediction model is supposed to use ΔF_0 as one of the predicting factors, 10 speakers (5 male and 5 female adults) were asked to utter each vowel sound so that its F_0 contour would curve as in **Figure 2**. Before recording the samples of each vowel, pure tones with their F_0 contours drawing the required curves were presented five times through headphones. For a variety of ΔF_0 in the vowel sounds, an initial part of the samples were uttered with β fixed and α changed at n ($\geqq 7$) levels. And each speaker was asked to keep the lowest tone, which is indicated as **L** in **Figure 2**, being fixed within the

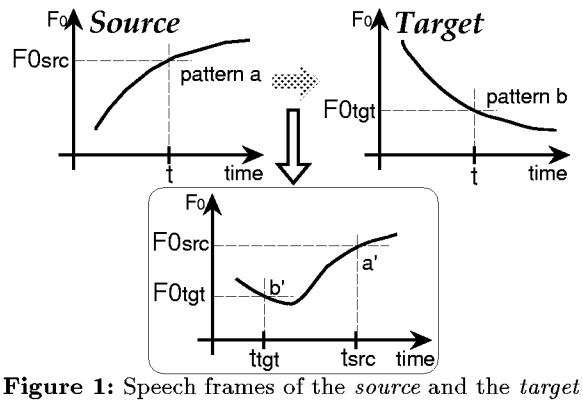

Figure 1: Speech frames of the *source* and the *target*

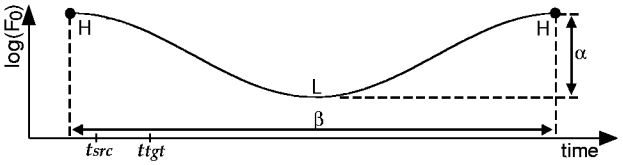

Figure 2: A required $F0$ curve for the recording

speaker over all the utterances. Next, setting β to another value, the speaker was asked to repeat the vowel in the same manner, that is, at n levels of α. And 0.6, 1.2, and 1.8 [sec] were assigned to β in this order. Throughout the recording, two utterances were requested for each $F0$ configuration of each vowel. Thus, $n \times 3 \times 2$ utterances were recorded for each vowel. A set of the first utterances in each $F0$ configuration will be called set-**1** for training and that of the second ones will be set-**2** for testing in this paper.

From the above material, all the pairs of frames satisfying either of the following conditions were extracted. A frame pair is represented as (fr(t_{src}), fr(t_{tgt})) below.

- $|t_{tgt} - t_{src}| < 100$ [msec] and $|F0_{tgt}/F0_{src}| > 2^{2/12}$
- 100 [msec] $\leqq |t_{tgt} - t_{src}| \leqq 300$ [msec]

And it should be noted that both of t_{src} and t_{tgt} must be located in a central part of a vowel sound. The extracted frames were used in the following analyses on the assumption that fr(t_{src}) and fr(t_{tgt}) were frames before and after the $F0$ change respectively as shown in **Figure 1**. Through all the experiments, speech samples were digitized with 10 kHz and 16 bit sampling. And the acoustic analysis was performed using 25.6 msec frame length and 5.0 msec frame rate. $F0$ was extracted with the same rate. As the cepstrum coefficients, 1 to 16 dimensions of LPC cepstrum coefficients were used after the 16-th order LPC analysis.

Although both of the male and the female speech samples were used through the analyses and experiments, space of paper allowed us to describe only the results of the male material, which resembled those of the female material.

2.2. VQ Distortion in a Five-vowel Space

The VQ distortions in a five-vowel space were calculated as a function of the codebook size, which will be compared to the spectral variations in Section **2.3.**. The VQ codebooks were obtained by applying the LBG algorithm to set-**1** and, using set-**2**, the distortion $D(M)$ was calculated as

$$D(M) = \frac{1}{N_2} \sum_{x \in \text{set-2}} \min_{1 \leqq k \leqq M} d(x, c_k^1) \ ,$$

where M and c_k^1 are the codebook size and the k-th centroid in the codebook for set-**1** respectively. x is a parameter vector, which is comprised of 16 cepstrum coefficients. And N_2 is the number of x in set-**2**. In this paper, the distance between two vectors, $d(y, z)$, is defined as the square of the Euclid distance between them. In the analyses, the codebooks were made for each speaker and the VQ distortions were also calculated for each speaker as shown in **Figure 3**. In the figure, it can be seen that the VQ distortion with the codebook size being the number of the

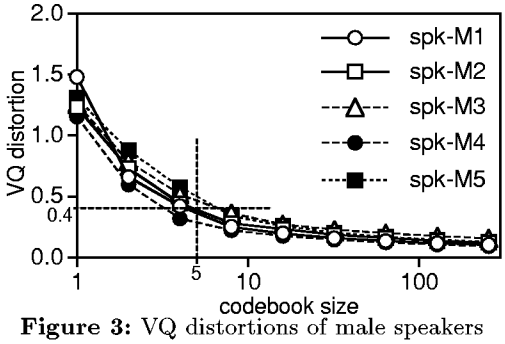

Figure 3: VQ distortions of male speakers

vowels is approximately 0.4, which will be called ρ in this work. If the variations by the $F0$ changes can exceed ρ, it means that the $F0$ changes can generate the larger spectral variations than the distortion produced by dividing the distributions of the 5 vowels' features into 5 sub-spaces.

2.3. Analysis of the Variation

Using set-**2**, the spectral variations by the $F0$ changes were calculated separately for each vowel over the speakers and they are plotted in **Figure 4** as a function of level of the $F0$ changes. In this calculation, an additional condition $t_{src} < t_{tgt}$ was introduced, without which the graphs will be completely symmetric. In the figure, the spectral variations are represented by the cepstral distance d between two frames extracted in Section **2.1.**. And the level of the $F0$ change is indicated by an integer n using which a value of $2^{(i/12)}$ can approximate $F0_{tgt}/F0_{src}$ the best. Here, the increase of n by 12 corresponds to that of $F0$ by an octave.

Several findings are obtained from the figure. Even with no $F0$ change, the cepstral distance is observed to be around 0.2. The rough symmetry in the graphs indicates that the $F0$ change of a given magnitude will produce approximately the same variation irrespective of the direction of the change. And it can be seen that the $F0$ changes by more than half an octave can generate the larger spectral variations than ρ, which is represented by a dash line. This fact implies that the $F0$ changes can produce the variations comparable to the distances between different phonemes.

3. MODELING OF THE VARIATION

A model for predicting the variations by the $F0$ changes was built based on the multivariate regression analysis, where a part of the followings were used as the predicting factors.

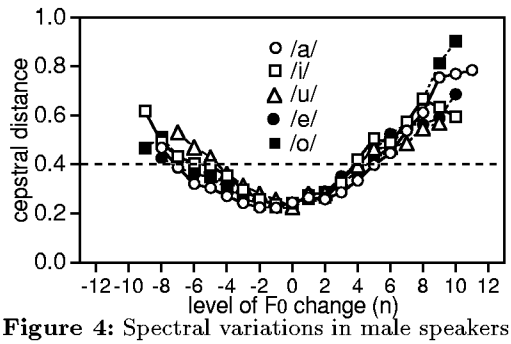

Figure 4: Spectral variations in male speakers

- $F0_{src}$, $\Delta F0_{src}$
- $F0_{tgt}$, $\Delta F0_{tgt}$
- c_j (j=0,...,16)

Here, c_j is the j-th dimension of the cepstrum of the *source*. And the i-th dimension of the cepstrum of the *target*, C_i (i=1,...,16), is treated as the predicted factor as in

$$\mathbf{C} = \mathbf{UV} + \mathbf{W} + \mathbf{E} \qquad (\hat{\mathbf{C}} = \mathbf{UV} + \mathbf{W}), \qquad (1)$$

where **C** and **V** denote a vector comprised of 1 to 16 dimensions of the *target* cepstrum and that of the predicting factors respectively. **U** and **W** are a regression matrix and a constant vector calculated by the multivariate regression analysis. And **E** is a vector representing an error term.

Since it is easily assumed that the relation between the variations and the $F0$ changes is naturally non-linear and that a matrix **U** and a vector **W** can characterize the relation only roughly. Thus, the multiple matrices and vectors were examined by dividing the training data into m subgroups. In preliminary experiments, the division of the data with m being 4 by the scalar quantization of $F0_{tgt}$ effectively reduced the prediction error. Then in the following sections, using set-**1**, 4 matrices and 4 vectors were obtained for each category, such as phonemes or speakers.

4. EVALUATION OF THE MODEL

4.1. Decreasing Prediction Error

The prediction errors were calculated using each of the following 9 combinations of the factors in **Table 1**. Here, using set-**1**, the models were built separately for each phoneme of each speaker. The error was defined as the averaged cepstral distance $\bar{d}$ between the target frame and the predicted frame by Equation (**1**) over the vowels and the speakers.

Figure 5 shows the prediction errors calculated by using each set. Comparisons between **B**&**C**, **D**&**E**, **F**&**G**, and **H**&**I** show that the use of $\Delta F0$ helps reducing the error. And the minimum error in the open experiments is observed in **B**. Then, the cepstral distances in **B** are plotted in **Figure 6** in the same format as **Figure 4**, indicating the remarkable effectiveness to decrease the spectral variations. And the decreased variations are almost the same as the VQ distortions at the codebook size of 10 to 20.

In order to examine the dependence of modeling capability on phonemes and speakers, using **B** and **D**, the following 4 configurations of modeling were investigated.

Table 1: 9 combinations of the predicting factors

A:	$F0_{src}$,	$\Delta F0_{src}$,	$F0_{tgt}$, $\Delta F0_{tgt}$,	c_j (j=0~16)
B:	$F0_{src}$,	$\Delta F0_{src}$,	$F0_{tgt}$, $\Delta F0_{tgt}$,	c_j (j=1~16)
C:	$F0_{src}$,		$F0_{tgt}$,	c_j (j=1~16)
D:	$F0_{src}$,	$\Delta F0_{src}$,	$F0_{tgt}$, $\Delta F0_{tgt}$,	c_j (j=i)
E:	$F0_{src}$,		$F0_{tgt}$,	c_j (j=i)
F:	$F0_{src}$,	$\Delta F0_{src}$,	$F0_{tgt}$, $\Delta F0_{tgt}$	
G:	$F0_{src}$,		$F0_{tgt}$	
H:			$F0_{tgt}$, $\Delta F0_{tgt}$	
I:			$F0_{tgt}$	

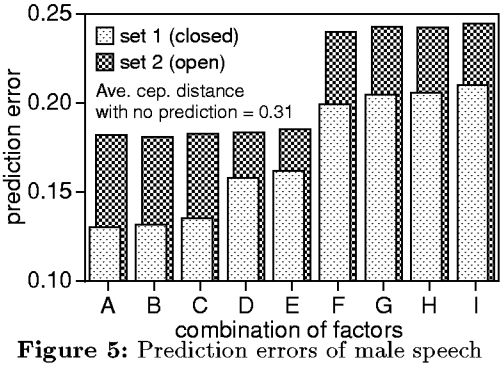

Figure 5: Prediction errors of male speech

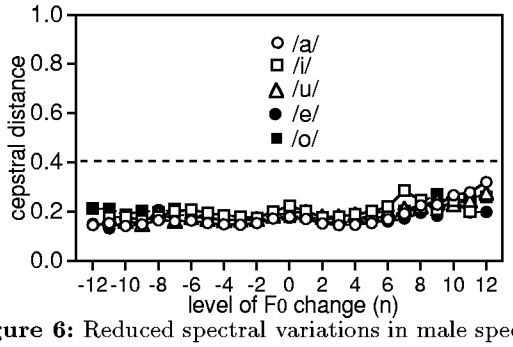

Figure 6: Reduced spectral variations in male speech

1) models separately for each phoneme using samples of all the speakers.
2) models separately for each phoneme using samples of all the speakers but s, are used as the model for speaker s.
3) models separately for each speaker using samples of all the vowels.
4) models separately for each speaker using samples of all the vowels but v, are used for the model for vowel v.

Models 1) and 2) differ in that, while the former use speech samples of s when testing those of s, the latter do not. Difference between models 3) and 4) exists in the use of speech samples of v when testing those of v. And it should be noted that the number of speakers and that of vowels are both 5 in the experiments. Using set-**2**, the averaged prediction errors over the speakers and the vowels for each configuration are shown in **Table 2**. In the table, the errors in the case of speaker and phoneme dependent models and those in the case of no prediction are also listed. When speech samples of a testing vowel or speaker are not used, i.e. 2) or 4), the errors are not reduced compared to those without prediction, indicating that the information on speakers or vowels should be given to the prediction. In comparison of difference between 1) and 2) and that between 3) and 4) in **B**, the stronger dependence is observed on phonemes than on speakers. These results point out that the analysis and modeling of the other voiced phonemes will be necessary especially for applications requiring less prediction errors, such as speech synthesis.

Table 2: Averaged prediction errors for each configuration

	no pred.	spk & vowel dependent	1)	2)	3)	4)
B	0.31	0.18	0.21	0.39	0.20	0.68
D	0.31	0.18	0.24	0.31	0.24	0.31

4.2. Increasing Separation among Classes

As mentioned in Section **1.**, the modification by the proposed model is expected to increase the averaged distance between the distributions of different phonemes. Applications such as speech recognition are considered to require the increased separation among classes rather than the decreased prediction error as discussed in Section **4.1.**. Furthermore, since an input speech frame is *unknown* in speech recognition, the following two measures should be prepared for evaluating the proposed models.

a) change of the distance between two vowels before and after the modification using the models of the two.
b) change of the distance between two vowels before and after the modification using the model of one vowel.

In measure a), the focus is placed upon difference between $\delta(p(v_i), p(v_j))$ and $\delta(q(v_i, M_i), q(v_j, M_j))$, where $p(v_i)$ is the distribution of vowel v_i's features and $q(v_i, M_i)$ is the modified distribution of the v_i's features by model M_i. $\delta(y, z)$ means the Bhattacharyya distance between classes y and z. On the other hand, in b), difference between $\delta(p(v_i), p(v_j))$ and $\delta(q(v_i, M_i), q(v_j, M_i))$ is observed. Even with larger separation found in a), reduced separation in b) indicates that M_i may work as a transformer of vowel v_j into v_i not just as a modifier of spectral envelopes of v_j.

In this section, another prediction manner is introduced. For example, Equation (**1**) and the factors **H** give us

$$\hat{C}_i = \phi_{i0} + \phi_{i1} F0_{tgt} + \phi_{i2} \Delta F0_{tgt},$$

which predicts C_i by adding $F0$ related variations to a constant value ϕ_{i0}, which can be interpreted as a neutralized value of C_i. Then an operation of

$$C_i^{observed} - \phi_{i1} F0_{tgt}^{observed} - \phi_{i2} \Delta F0_{tgt}^{observed}$$

can be considered to predict the neutralized value. And this operation can also be utilized as a mapping function and will be called as *backward* prediction. Equation (**1**)-based prediction will be denoted as *forward* prediction.

Figure 7 shows the results of a) and b) with forward(+) /backward(−) predictions using speaker and phoneme dependent models. In the figure, using set-**2**, the minimum, maximum and mean distances obtained for each speaker are averaged over the speakers and plotted. Here, the modification was done with $F0$ and $\Delta F0$ set to the averaged value and zero respectively. It is found that, although the modification always makes the separation larger in a), only the backward prediction can enlarge the separation in b).

Dependence on speakers and phonemes was also examined. **Table 3** shows the error reduction rates obtained in recognition experiments, where an input frame was firstly modified assuming that it belonged to vowel x, then matched with a multivariate Gaussian model (GM) for x. Here, the models were built using set-**1** which were modified with $F0$ and $\Delta F0$ being the averaged value and zero. It is interesting that, while $\mathbf{I}^-$ shows the highest rate with the speaker and phoneme dependent models[4], $\mathbf{B}^+$ gives us even higher rates in 1), 2), and 3). These results are also supported by a preliminary analysis using measures a) and b).

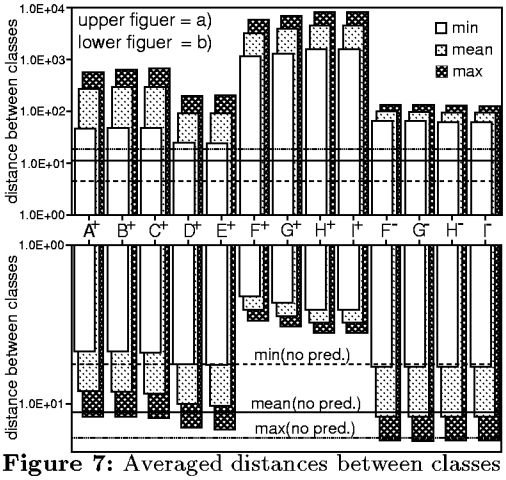

Figure 7: Averaged distances between classes

Table 3: Recognition error reduction rates [%]

	spk & vowel dependent	1)	2)	3)	4)
$\mathbf{B}^+$	2.97	25.8	30.4	18.5	4.56
$\mathbf{D}^+$	12.1	23.1	25.1	15.2	10.4
$\mathbf{I}^-$	13.9	20.0	12.4	−9.1	−4.9

5. CONCLUSIONS

In this paper, the correlation between spectral variations and $F0$ changes was firstly analyzed, where the variations were also compared to VQ distortions calculated in a five-vowel space. As a result, it was shown that the $F0$ change can produce the comparable variations to the cepstral distances between two phonemes. Next, a prediction model for cepstral coefficients' variations was built using the multivariate regression analysis. This model was evaluated separately in terms of a spectral envelope predictor with a given $F0$ and a mapping function of feature sub-spaces. It was indicated that, while the models should be built dependently on phonemes and speakers as the former, with adequate selection of parameters, the speaker/phoneme-independent models can work effectively as the latter. Several works, however, remained to be done, such as adequate treatment of other voiced phonemes and evaluation of the proposed models in speech recognition/synthesis.

REFERENCES

1. H. Kasuya *et al.*, "Changes in pitch and first three formant frequencies of five Japanese vowels with age and sex of speakers," J. Acoust. Soc. Jpn. 24, 6, pp.355-364 (1968, in Japanese).
2. D. O'Shaughnessy, "Speech Communication," Addison Wesley (1987).
3. H. Mizuno *et al.*, "A formant frequency modification algorithm dealing with the pole interaction," Trans. IEICE, Vol.J78-A, No.3, pp.287-294 (1995, in Japanese).
4. H. Singer *et al.*, "Pitch dependent phone modelling for HMM-based speech recognition," J. Acoust. Soc. Jpn. (E) 15, 2, pp.77-86 (1994).
5. K. Tanaka *et al.*, "A text-to-speech system with transformation of spectrum envelope according to fundamental frequency." *Report of Spring Meet. Acoust. Soc. Jpn.*, 2-7-1, pp.217-218 (1997, in Japanese).