Modeling of contextual effects and its application to word spotting

Yuji YONEZAWA and Masato AKAGI

Japan Advanced Institute of Science and Technology, Hokuriku
1-1 Asahidai Tatsunokuchi, Nomi, Ishikawa 923-12, Japan
yonezawa@jaist.ac.jp, akagi@jaist.ac.jp

ABSTRACT

We propose a model of spectral contextual effects to simulate the superior recognition ability of humans and apply it to a front-end processor for word spotting. This model assumes that perceived spectra are influenced by adjacent spectral peaks and that the magnitude of the influence can be estimated by the minimum classification error criterion. Three experiments were carried out to evaluate the performance of the model. The results show that the model can compensate for neutralized spectra and bring them to their typical patterns. This improves word spotting accuracy.

1. INTRODUCTION

Continuous speech contains incomplete articulated phonemes due to the limitations of the acoustic apparatus. This is one of the most serious problems in automatic continuous speech recognition systems. Humans, on the other hand, can correctly perceive such phonemes as if they were uttered clearly. The explanation for this is that humans perceive phonemes via a compensation mechanism, and contextual effects can contribute to this compensation. Therefore, modeling contextual effects and applying them to speech recognition could improve continuous speech recognition accuracy.

One of authers has already proposed a functional model of contextual effects based on psychoacoustic experiments[1]. This functional model assumed that perceived spectral peaks are influenced by adjacent spectral peaks and are shifted along the frequency axis by the sum of the interactions between spectral peaks.

In this paper, however, assuming that perceived spectra are influenced by adjacent spectral peaks and that the magnitude of this influence can be estimated by the minimum classification error criterion (MCE)[2], we refine the model to make it easy to apply speech recognition as a front-end processor, especially to cope with coarticulation problems in a word spotting task. Through the use of the model, spectra are modified so that they incorporate contextual effects, and approach their typical aspects. This performance of the model can reduce the difference between reference data and input data in a word spotting task, and thus improve word spotting accuracy.

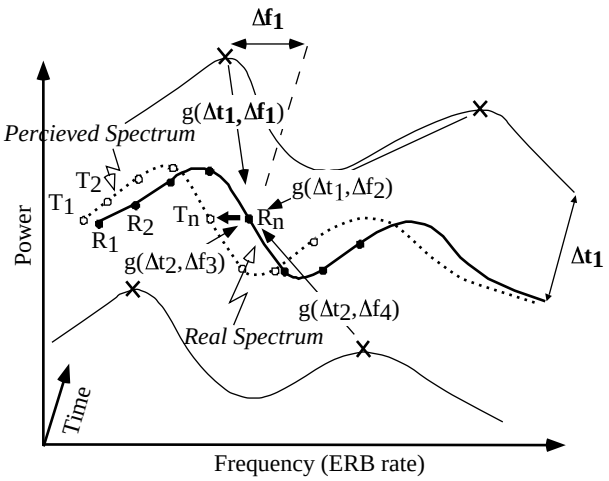

Figure 1: Concepts of the model. Sampling points(R_n) of real spectrum shift to T_n along the frequency axis by adjacent spectral peak influences.

Section 2 describes a model of spectral contextual effects. Section 3 describes three experiments for evaluating the performance of the model. In the first experiment, we investigated how the model modifies formant trajectories. In the second, we measured the performance of the model as a front-end processor for vowel recognition. And in the third, we measured its performance as a front-end processor for word spotting.

2. MODEL OF SPECTRAL CONTEXTUAL EFFECTS

2.1. Concepts of the Model

The model assumes that perceived spectra are influenced by adjacent spectral peaks and are shifted along the frequency

axis. This assumption originated in the previous model[1]. Figure 1 shows the concepts of the model. When humans perceive a spectrum, each adjacent spectral peak ($\times$) influences the perception of the spectrum by $g(\Delta t, \Delta f)$, where Δt is the difference in time (ms) between a sampling point (R_n; $\bullet$) of the real spectrum and the spectral peak, and Δf is the difference in frequency (ERB[3]) between them. $\Delta t > 0$ indicates that the spectral peaks influence perception of the spectrum backwards and $\Delta f > 0$ indicates that they influence perception of the spectrum from higher frequencies. Sampling points (T_n; $\circ$) of the perceived spectrum are shifted by the sum of $g(\Delta t, \Delta f)$ along the frequency axis from those of real spectra. Thus, the sampling points of the perceived spectrum are represented as follows:

$$T_n = R_n + \sum_{m}^{M} \sum_{n}^{N} g(\Delta t_m, \Delta f_n) \qquad (1)$$

where M is the number of frames in which contextual effects are considered and N is the number of spectral peaks in the m-th frame. When $g(\Delta t_m, \Delta f_n)$ is positive, a sampling point R_n of a real spectrum is perceptually shifted to higher frequency by the influence of a spectral peak.

2.2. Estimation of $g(\Delta t, \Delta f)$

Since the contextual effects model is intended to simulate the human compensation mechanism, the function $g(\Delta t, \Delta f)$ must closely match the results of psychoacoustic experiments. In addition, the model with $g(\Delta t, \Delta f)$ should help improve speech recognition accuracy. Therefore, $g(\Delta t, \Delta f)$ was roughly estimated on the basis of the results of psychoacoustic experiments[1], and values of it were estimated using MCE.

Approximation of the function $g(\Delta t, \Delta f)$. Results of previous experiments[1] suggest that when Δt or Δf is either very large or almost zero, spectral peaks do not influence spectra. Shigeno also suggests that contextual effects could be constructed from assimilation and contrast effects[4]. Considering these results, we approximated function $g(\Delta t, \Delta f)$ as follows:

$$g(\Delta t, \Delta f) = \begin{cases} e^{A|\Delta f|} \sin(B\Delta f)(Ce^{-\frac{\Delta t^2}{D}} \\ \qquad -Ee^{-\frac{\Delta t^2}{F}})|\Delta t|^G, & \text{if } |\Delta f| \leq \frac{\pi}{B} \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

where $e^{(A|\Delta f|)} \sin(B\Delta f)$ expresses that $g(\Delta t, \Delta f)$ is almost zero when Δf is very large or almost zero, and A and B are coefficients describing the range of Δf in which the peaks are influenced. $Ce^{-\frac{\Delta t^2}{D}}$ and $Ee^{-\frac{\Delta t^2}{F}}$ illustrate the distribution of assimilation and contrast effects respectively, and G is a coefficient describing the degree of decay around $\Delta t = 0$.

Estimation of coefficients A to G. Coefficients A to G were estimated by the MCE criterion. The aim of this was to construct a model that would improve speech recognition accuracy. The ATR speech database was used as speech

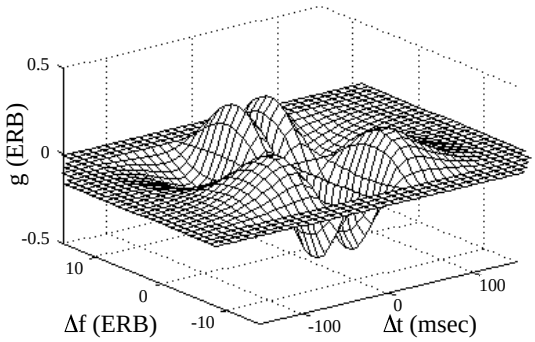

Figure 2: The function $g(\Delta t, \Delta f)$ represented by Eq. (3).

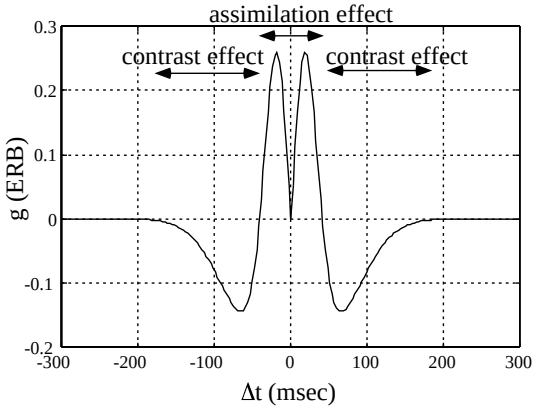

Figure 3: g versus Δt for $\Delta f = 4$ ERB of Fig. 4. $g > 0$ indicates an assimilation effect, and $g < 0$ indicates a contrast effect.

data. In this database, 25 Japanese sentences (SC1) from the "conference registration task", which were uttered continuously by a male, were used for MCE learning. The sampling frequency was 20 kHz.

First, spectral peaks were extracted from 16-th order LPC spectra, provided their frequency was under the 24 ERB rate (2800 Hz). Next, the center segment of each vowel in the speech data was extracted by a 30-ms Hamming window. 834 samples were chosen for learning. The extracted speech data were converted into 40-th order FFT-cepstrum smoothed spectra on the ERB-rate scale. Influences of the spectral peaks were calculated using the function $g(\Delta t, \Delta f)$ and these spectra were modified by Eq. (1). Then the modified spectra were divided into each vowel category by the discriminant function defined as the Euclidean distance between the modified spectrum and each isolated uttered vowel spectrum. After that, the coefficients A to G in Eq. (2) were estimated based on the MCE criterion. The estimated function $g(\Delta t, \Delta f)$ is shown in Figs. 2 and 3 and described as

follows

$$g(\Delta t, \Delta f) = \begin{cases} e^{-0.031|\Delta f|}\sin(0.27\Delta f) \\ (0.023e^{-\frac{\Delta t^2}{775}} - 0.0036e^{-\frac{\Delta t^2}{5075}}) \\ |\Delta t|^{1.16}, & \text{if } |\Delta f| \leq \frac{\pi}{0.27} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

The results show that contextual effects were most evident when $\Delta f = 5$ ERB, and that assimilation effects were evident when $\Delta t < 50$ ms and contrast effects were evident when 50 ms$< \Delta t <$170 ms. These results are consistent with psychoacoustic experiments[1].

3. EVALUATION OF THE MODEL

3.1. EXPERIMENT 1: Extrapolation of reduced formant trajectories

The aim of experiment 1 was to investigate how the model modifies formant trajectories. Formant trajectories in continuous utterances are reduced due to coarticulation and do not get to their typical position, i.e., the isolated uttered formant position. If the model can extrapolate reduced formant trajectories, recognition accuracy in continuous speech could be improved.

An example of spectral peak trajectories is shown in Fig. 4. The formant trajectories do not get to their typical position without the models and the model brings the formant trajectories to almost their typical position.

Figure 5 illustrates the spectra at 1500 ms in Fig. 4. It also shows that the model brings the neutralized spectra to the typical spectrum. The Euclidean distance between them was reduced by 8% by applying the model.

These results show that model can reduce coarticulation effects.

3.2. EXPERIMENT 2: Recognition accuracy of vowels in continuous speech

The aim of experiment 2 was to investigate whether the model as a front-end processor can improve recognition accuracy of vowels in continuous speech. Spectra in continuous speech are neutralized due to coarticulation, and this neutralization reduces recognition accuracy. Since the model can compensate for neutralization of the spectrum, we can expect some improvement in recognition accuracy by using the model.

Three sets of speech data (SC1, SC2, SC3) in the ATR speech database were used for the test. SC1 is the same data as used for estimating $g(\Delta t, \Delta f)$. The center segment of each vowel in the speech data was extracted and converted into the 40-th order FFT-cepstrum smoothed spectrum. There were 2335

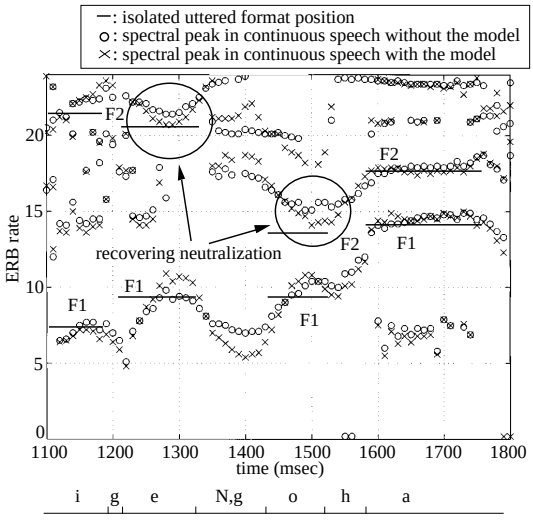

Figure 4: spectral peak trajectories in part of a sentence (/gengoha/) and isolated utterd formant position of each vowel.

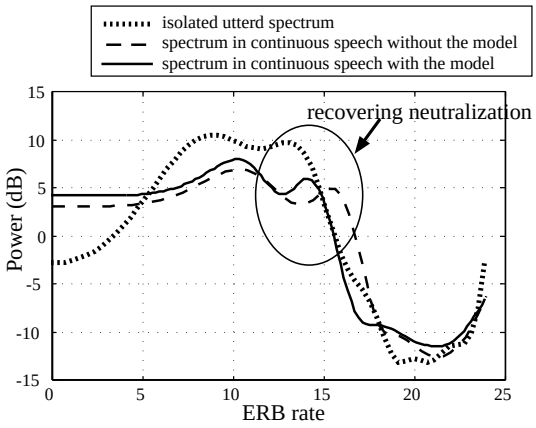

Figure 5: Spectra of /o/ at 1500ms in Fig. 4. The model shifts the neighborhood of F2 to lower frequency and recovers neutralization.

samples in total. They were modified by the model and were divided into each vowel category by the discriminant function defined as the Euclidean distance between the modified spectrum and each isolated uttered vowel spectrum.

The results are shown in Fig. 6, and results when not using the model are shown for comparison. The results of the χ^2 test ($\chi^2 = 5.05$) suggest that the improvement achieved by the model is significant (significance probability is 5%). These results show that spectrum modification by the model improves vowel recognition accuracy and that the model can work well as a front-end processor in vowel recognition.

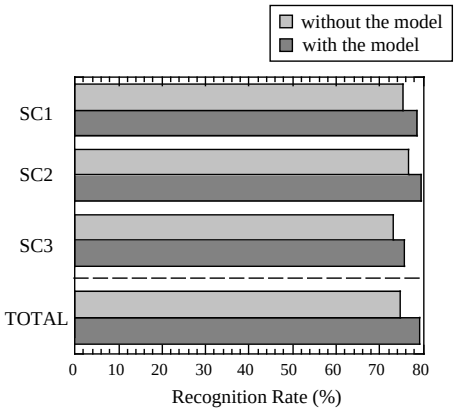

Figure 6: Vowel recognition accuracy.

3.3. EXPERIMENT 3: Word Spotting

Coarticulation reduces word spotting accuracy in continuous speech when isolated uttered words are used as reference patterns. The model was used as a front-end processor for word spotting to cope with coarticulation effects. The aim of experiment 3 was to investigate whether using the model can improve word spotting accuracy.

As in the first two experiments, the ATR speech database uttered by the male was used as speech data. 75 continuously uttered Japanese sentences from the database were used as input data. 13 Japanese words uttered in isolation were used as reference data, i.e., as keywords.

Figure 7 shows a block diagram of the experiment. Input data and reference data were converted into 40-th order FFT-cepstrum smoothed spectra, and the spectra of input data were modified by the model. Then, continuous DP matching[5] was applied to detect sections in which the score was below the threshold value and the minimum. The results of the experiments (Fig. 8) show that the model can improve word spotting accuracy.

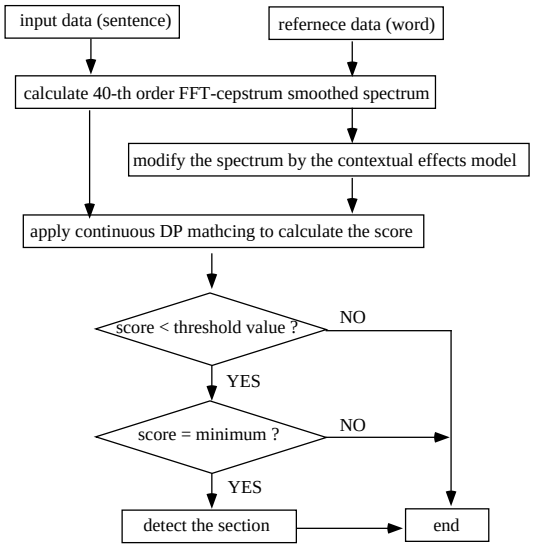

Figure 7: Block diagram of the word spotting with the model.

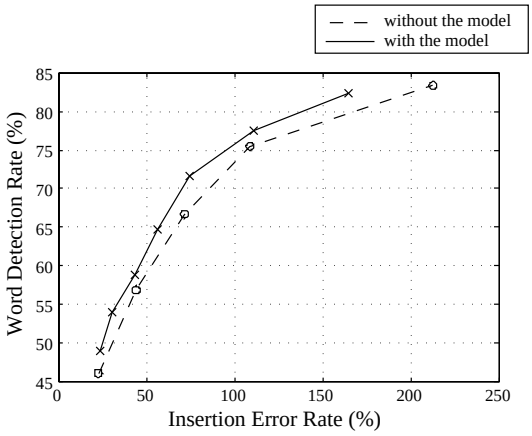

Figure 8: Comparison of word spotting performance. Insertion error rate is defined as "number of insertion errors"/"number of subject words to be detected in input data".

4. CONCLUSION

Our model of spectral contextual effects, constructed using results of previous psychoacoustic experiments and MCE, can reduce coarticulation effects. This improves vowel recognition accuracy and word spotting accuracy. However, the model can modify spectra only along the frequency axis and can not reduce difference of spectra along power axis as in Fig. 5. Furter study on this is required.

5. REFERENCES

1. Masato Akagi. Modeling of contextual effects based on spectral peak interaction. *J. Acoust. Soc. Am.*, 93(2):1076–1086, February 1993.
2. Biing-Hwang Jung and Shigeru Katagiri. Discriminative learing for minimum error classification. *IEEE Trans. SP*, 40(12):3043–3054, December 1988.
3. B.R.Glasberg and B.C.J.Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, pages 103–138, 1990.
4. Sumi Shigeno. Assimilation and contrast in the phonetic perception ov vowel. *J. Acoust. Soc. Am.*, 90(1):103–111, July 1991.
5. Ryu-ichi OKA. Continuous words recognition by use of continuous dynamic programming for pattern matching. *Trans. Inst. Elec. Comm. Eng. Jpn.*, J67-D(6):677–684, June 1984.