ICSLP, 3:774-777, 2000

ON THE IMPORTANT MODULATION-FREQUENCY BANDS OF SPEECH FOR HUMAN SPEAKER RECOGNITION

Takayuki Arai, Mahoro Takahashi, Noboru Kanedera†, Yukiko Takano†, and Yuji Murahara

Dept. of Electrical and Electronics Eng., Sophia University, Tokyo, JAPAN
† Ishikawa National College of Technology, Ishikawa, JAPAN

ABSTRACT

By means of perceptual experiments, we investigated what range of modulation frequency components of the mel-frequency cepstral coefficients (MFCC) contains the most important information for speaker identification.

In our study, we conducted two perceptual experiments using an MFCC-based re-synthesis scheme with two types of excitation. In Experiment I, speech sounds were re-synthesized from the extracted pitch and white noise. In Experiment II, speech sounds were re-synthesized only from white noise to avoid including pitch information. For each experiment the original speech sounds were uttered by two sets of five professors. A total of 44 students (16 for Exp. I and 28 for Exp. II) who attend the professors' classes participated in the experiments.

We analyzed the experimental results in order to estimate the relative importance of different modulation frequencies in speaker recognition. The results show that the most important speaker information was in modulation frequency components from 2 to 8 Hz for both Exp. I (pitch-excited) and Exp. II (noise-excited). These results also show that some contribution was derived from including modulation frequency components around 0 Hz. Hence, we concluded that dynamic features are important for human speaker identification as well as static features.

1. INTRODUCTION

The spectral representations of the time trajectory of a parameter, called the modulation spectrum, plays an important role in human and machine speech recognition [1]-[3]. Arai et al. [1] extended Drullman's experiments [2, 3] and reported that speech intelligibility is not greatly decreased as long as the filtered spectral components have a rate of change between 1 and 16 Hz.

An automatic speech recognition (ASR) experiment [4] also indicated that the most important linguistic information is in the modulation frequency components between 1 and 16 Hz, especially between 2 and 10 Hz. In some experimental environments, the use of components below 2 Hz or above 16 Hz can degrade recognition accuracy. Consequently, the important modulation frequency range for ASR is almost the same as that of human speech intelligibility, and this suggests that investigating human perception can help in the design of ASR systems.

For automatic speaker recognition, on the other hand, Van Vuuren et al. [5] concluded that the modulation frequency components between 0.1 and 10 Hz contain the most useful speaker information. Since humans are readily able to identify speakers, we ought to look at human perception for clues to increase the level of performance of the technology. However, no speaker identification experiments focusing on the modulation frequency domain have been reported yet.

In this study, we conduct two perceptual experiments using an MFCC-based re-synthesis scheme with two types of excitation. In the next section, we will describe the human speaker identification experiments. The results will be presented and discussed in Section 3.

2. HUMAN SPEAKER IDENTIFICATION EXPERIMENTS

2.1. Signal Processing

An overview of the signal-processing method that we used with a signal-processing tool [6] is illustrated in Fig. 1. The main purpose of this process is to pass the speech signal with various modulation frequency components.

The speech signals were first analyzed with a 25 ms Blackman window advanced in 5 ms steps. The 12th-order MFCC were obtained. Then, to bandpass certain modulation frequency components, the time trajectories of the resulting MFCC were filtered with 511-tap linear-phase FIR filters. Subsequently, from the temporally-filtered MFCC with pitch and white noise, the modified speech signal was reconstructed using the mel log spectrum approximation (MLSA) filter [7].

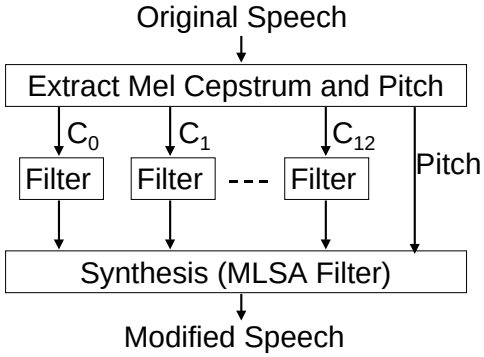

Figure 1: Block diagram of the speech processing.

2.2. EXPERIMENT 1

We processed utterances using signal processing based on the system in Section 2.1 using 36 different modulation filters. Thirty-six bandpass filters were used with modulation cutoff frequencies of 0, 0.25, 0.5, 1, 2, 4, 8, 16 Hz, and f_N (Nyquist frequency). When re-synthesizing speech signals, pitch and white noise were used.

The original speech samples were uttered by five professors of Ishikawa National College of Technology. A total of 16 students of the same college who listen to the lectures of those five professors participated in Experiment 1. We used a single sentence selected from the ATR speech corpus [8], "Arayuru genjitsu-o subete jibun-no hou-e nejimagetanoda."

Prior to the main session, we played another phrase uttered by the five professors, by way of reminder to the subjects of the voices of the speakers. Each phrase was played twice. In the main session, each subject listened to 180 stimuli (36 filtering conditions $\times$ 5 professors). We presented stimuli in the order of filtering conditions from difficult to easy to identify the speakers. In each condition, the stimuli were presented in random order. For each trial, the subjects were asked to give the speaker ID number, or zero if they could not identify the speaker.

2.3. EXPERIMENT 2

In Experiment 2, we used only white noise to re-synthesize the speech signal, because it was thought that pitch information might affect the experimental results. We processed different utterances using signal processing based on the system in Section 2.1 using 28 different modulation filters. Twenty-eight bandpass filters were used with modulation cutoff frequencies of 0, 0.5, 1, 2, 4, 8, 16 Hz, and f_N (Nyquist frequency). When re-synthesizing speech signals, only white noise was used.

The original speech sounds were uttered by five professors of Sophia University. A total of 28 students of the same university who listen to the lectures of those five professors participated in Experiment 2. We used 12 sentences selected from the ATR speech corpus [8], such as, "Terebi-gêmu-ya pasokon-de gêmu-o shite asobu."

Noise-excited speech sounds like whispering, and the speaker is often difficult to identify because there is no pitch information. We therefore trained the subjects. In the training session, the subjects heard five unprocessed speech sounds of the same sentence uttered by the five speakers. We tested whether the subjects could identify the speakers of five unprocessed speech sounds of a different sentence. Next, we trained the subjects until they could perfectly identify speakers of five processed but unfiltered speech sounds with another sentence. In the end, only the subjects who completed the training and passed all of these tests were allowed to participate in the main session.

In the main session, each subject listened to 60 stimuli. Each stimulus was a filtered sentence of one speaker. Each subject listened to a sentence of a speaker with no more than one filtering condition to avoid the possibility of learning. The stimuli were presented in random order. Subjects were forced to guess the identity of the speaker on each trial. During the main session, whenever they wished, the subjects could listen to the five unfiltered sounds as they were uttered by the five speakers in the training session.

3. RESULTS AND DISCUSSION

Figs. 1 and 2 show the recognition results for different bandpass filters applied in the modulation frequency band in Experiment 1 and 2, respectively. In each figure, the vertical axis shows the speaker identification rate, while the other axes indicate the lower cutoff frequency f_L and the upper cutoff frequency f_U of the bandpass filters.

The contribution to recognition performance was computed by a multiple regression based on [4] for each modulation frequency band as shown in Figs. 3 and 4. Each bar indicates the contribution for a 95% confidence interval of overall speaker recognition accuracy for the corresponding modulation frequency band. By including a modulation frequency band, the probability of error is the reciprocal of the corresponding contribution factor.

Both Figs. 3 and 4 show that most speaker information is contained between 2 and 8 Hz. In Experiment 1, a speech signal was synthesized from the MFCC and the pitch information. The pitch is, however, one of the most

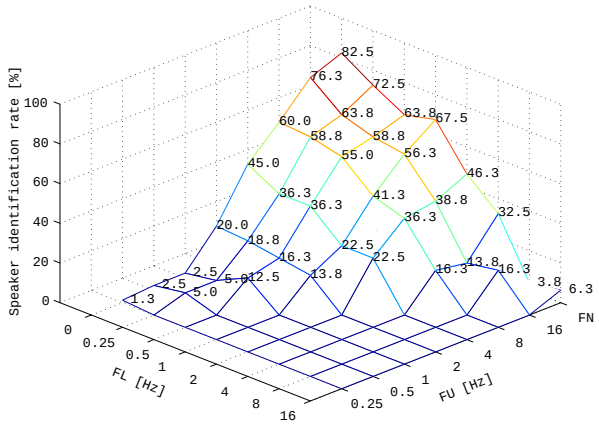

Figure 2: Recognition results for the bandpassed time trajectories (Experiment 1).

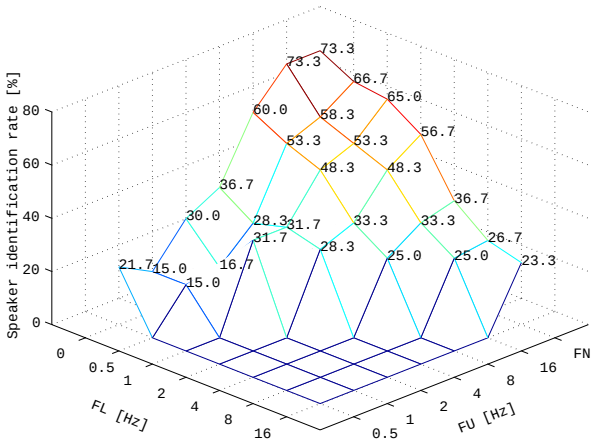

Figure 3: Recognition results for the bandpassed time trajectories (Experiment 2).

useful parameters for speaker recognition [9], and therefore the results might have been affected by pitch. In Experiment 2, we used only white noise without using pitch to re-synthesize a speech signal. It turned out that the results of two experiments were consistent. This indicates that the most important speaker information is in the modulation frequency components between 2 and 8 Hz, with or without pitch.

Thus, the modulation frequency band found to be most important is similar to that obtained in automatic speaker identification experiments [5], where the most important speaker information was in modulation frequency components between 0.5 and 8 Hz. This supports the conclusion that the time trajectories of MFCC include enough speaker information to identify speakers.

As can be seen in Figs. 3 and 4, including components of the modulation spectrum which are around 0 Hz increases contribution, while including components between 0.25 and 1 Hz leads to a decrease in contribution. This indicates that including components of modulation spectrum between 0 and 0.25 Hz increases the contribution, while including components between 0.25 and 0.5 Hz leads to a decrease in contribution. On the contrary, automatic speaker identification experiments [5] indicate that the modulation spectrum below about 0.125 Hz reduces the overall accuracy.

In perceptual experiments some benefit was derived from including the modulation frequency range around 0 Hz. Hence, it seems that static features such as formant structure are also important for human speaker identification.

4. CONCLUSIONS

We conducted perceptual experiments to determine what range of the modulation frequency components of MFCC is most important for speaker recognition. The results of our experiments suggest that the most important speaker information is in modulation frequency components ranging between 2 and 8 Hz. The static features also seem to be important for human speaker identification.

ACKNOWLEDGEMENTS

We acknowledge the effort of Prof. Tokuda of Nagoya Institute of Technology who developed the Speech Processing Toolkit. We would also like to thank the professors and students of Ishikawa National College of Technology and Sophia University who participated in the experiments. We would like to express our grateful thanks to Terri Lander of the University of Colorado on her helpful comments.

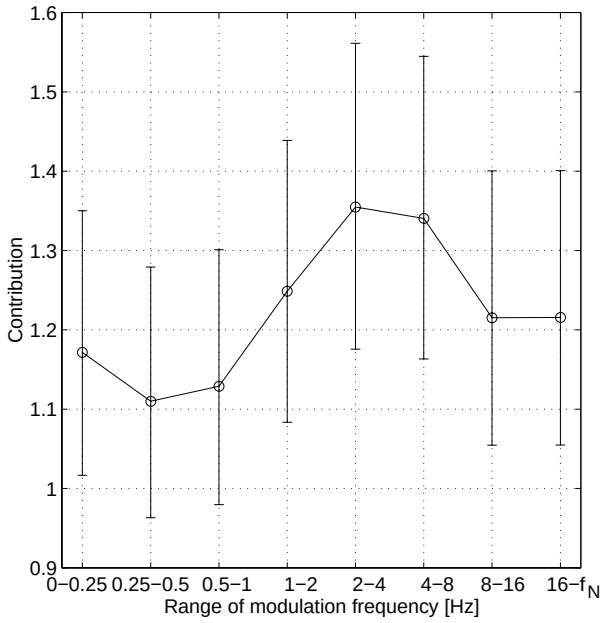

Figure 4: Contributions to recognition performance with 95% confidence intervals (Experiment 1).

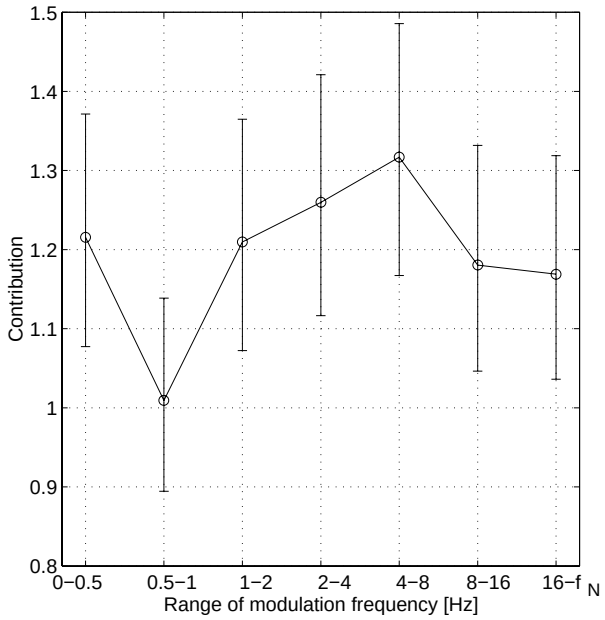

Figure 5: Contributions to recognition performance with 95% confidence intervals (Experiment 2).

5. REFERENCES

[1] T. Arai, M. Pavel, H. Hermansky and C. Avendano, “Syllable intelligibility for temporally filtered LPC cepstral trajectories,” *J. Acoust. Soc. Am.*, **105**, pp. 2783–2791, 1999.

[2] R. Drullman, J. M. Festen and R. Plomp, “Effect of temporal envelope smearing on speech reception,” *J. Acoustic. Soc. Am.*, **95**, pp. 1053–1064, 1994.

[3] R. Drullman, J. M. Festen, and R. Plomp, “Effect of reducing slow temporal modulations on speech reception,” *J. Acoustic. Soc. Am.*, **95**, pp. 2670–2680, 1994.

[4] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, **28**, pp. 43–55, 1999.

[5] S. van Vuuren and H. Hermansky, “On the importance of components of the modulation spectrum for speaker verification,” *Proc. ICSLP*, pp. 3205–3208, Sydney, Australia, 1998.

[6] K. Tokuda et al., “Signal-Processing Toolkit,” http://kt-lab.ics.nitech.ac.jp/˜tokuda/SPTK/.

[7] S. Imai, K. Sumita and C. Furuichi, “Mel Log Spectrum Approximation (MLSA) filter for speech synthesis,” *Trans. on IECE*, **J66-A**, pp. 122–129, 1983 (in Japanese).

[8] H. Kuwabara, Y. Sagisaka, K. Takeda and M. Abe, “Construction of ATR Japanese speech database as a research tool,” *ATR Technical Report*, **TR-I-0086**, 1989.

[9] L. R. Rabiner and R. W. Schafer *Digital Processing of Speech Signals*, Prentice Hall, New Jersey, 1978.