A study on analysis and synthesis of speech sounds uttered in car under noisy environments

Yoshinari Takeyama (410075)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 9, 2006

Keywords: Speech Recognition, Speech Synthesis, Lombard Effect.

1 Introduction

Features of speech sounds uttered in noisy environments, (e.g., in car environments), are considerably different from that of sounds uttered in noise-free environment. These differences are caused by Lombard effect and these bring rapid reduction of performances in automatic speech recognition (ASR) systems. To solve this problem, huge speech sound databases collected in car are required for training acoustic models in the ASR systems. However, there are few speech sound databases collected in car. This is because it is too difficult to collect huge speech sound data in any car environments. In order to resolve this issue without collecting speech data, we have been trying to synthesize a large amount of various mimic speech sounds in car from a few speech sounds uttered in car. In this paper, we propose a method for converting clean speech sounds into mimic speech sounds in car as a function of velocity. That is one factor characterizing car environment. The converted speech sounds could be used for training acoustic models in the ASR systems.

2 Proposed method

The proposed conversion system consists of three parts: analysis part (feature decomposition), conversion part, and synthesis part (feature composition). In this system, input is clean speech sound, and output is mimic speech sound in car. Analysis part decomposes acoustical features of clean speech sound. These acoustical features are converted into corresponding acoustical features of mimic speech sound in car by converting function $H(v)$ as a function of velocity v. Synthesis part synthesizes a mimic speech sound in car using converted acoustical features. In this manner, mimic speech sounds in car are synthesized from clean speech sounds. In this study, we adopted STRAIGHT which enables high-quality analysis and synthesis for speech sounds. In this system, acoustical features of fundamental frequency (F_0), power, formant frequencies (F_1, F_2, and F_3), spectrum tilt, and vowel duration were converted into that of mimic speech sounds in car. These are representative features distorted by Lombard effect.

3 Analysis and Results

In order to convert acoustical features of clean speech sounds into that of sounds in car environments, it is necessary to examine relationships between velocity and acoustical features of speech sounds uttered in car. Speech sounds uttered in car were recorded as target of analysis. Deviations of acoustical features based on that of clean speech were analyzed.

Deviations in terms of fundamental frequency and power have no large difference between each vowel and all of them increase monotonically as velocity increases. This is because Lombard effect occurred remarkably. These features can be represented as one function, independing upon vowels.

Although deviations in terms of formant frequencies increase roughly as velocity increases, trends of increasing are different between vowels. Means of formant frequencies might not be estimated accurately because of fluctuations of acoustical features about some subjects and vowels. Generally, Lombard effect causes formant shift, and it depends on frequency. Low

frequency shifts up and high frequency shifts down. This suggests that it is necessary to compose converting functions separately according to vowel and band of frequency.

Deviations in terms of spectrum tilt increase roughly as velocity increases. This shows that spectrum tilts are gradually flat as power of high frequency bands increases because of Lombard effect. There are some differences in terms of trend of changing between vowels.

Deviations in terms of vowel duration depended on order of mora. First mora and last mora tend to be longer than others. Especially the last morae tend to be long strongly. It is necessary to compose converting functions with regard to duration separately upon order of mora.

4 Synthesis of mimic speech sounds uttered in car environments

Mimic speech sounds in car are synthesized using the converting functions. Converting functions are composed as the second-order polynomials by using the least mean square adaptation method from deviations of acoustical features. The converting functions with regard to fundamental frequency and power are represented, independing upon vowels. The converting functions with regard to formant frequencies and spectrum tilt are represented, according to the vowels. The converting functions with regard to duration is represented, according to the order of mora.

Evaluation results of converted speech sounds with subjective auditory impression by comparing with that of uttered speech sounds are shown as follows. The pitch of both sounds raised and the volume of both grew as velocity increased. Thus, converting functions for fundamental frequency and power are reasonable. However, because of the fluctuations of formant frequencies and spectrum tilt, some converted speech sounded strange against uttered speech. These values may be still different from uttered speech.

5 Evaluation results of converted speech sounds

Evaluation results of converted speech sounds whether these sounds useful for training acoustic models or not are shown as follows. ASR systems

recognize speech sounds using acoustical features (e.g., MFCC). Thus, if these features of converted speech sounds are close to that of recorded speech sounds, converted speech sounds are useful for training acoustic models. MFCC and Δ MFCC and Δ power are measured from converted speech sounds and recorded speech sounds. And distributions of these features are analyzed. As a results, distributions in terms of converted speech sounds are closer to that of recorded speech sounds than that of clean speech sounds. This shows converted speech sounds are more useful for training acoustic models for car environments than clean speech sounds.

6 Conclusion

In this paper, we proposed a method for converting clean speech sounds into mimic speech sounds in car. In this method, we analyzed fundamental frequency (F_0), power, formant frequencies (F_1, F_2, and F_3), spectrum tilt, and vowel duration as acoustical features, and we constructed a system for synthesizing a mimic speech sounds uttered in car environments by controlling these features as a function of velocity.

As a results, it was shown that converted speech sounds are more useful than clean speech sounds for training acoustic models for car environments. It is expected to raise ASR accuracy in car environments for training acoustic models using a large amount of speech sounds synthesized by this system.