The Seventh Western Pacific Regional Acoustics Conference
WESTPRAC VII
Kumamoto, Japan, 3-5 October 2000

LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION PERFORMANCE IN CAR ENVIRONMENTS FOR VARIOUS PHONEME MODELS

Miichi YAMADA*, Katsumi NISHITANI**, Satoshi NAKAMURA*, Kiyohiro SHIKANO*

*Nara Institute of Science and Technology
8916-5 Takayama-cho, 630-0101 Japan
Fax: +81-743-72-5289
{miichi-y,nakamura,shikano}@is.aist-nara.ac.jp

**Harness System Technologies Research, Ltd.
1-1-3, Shimaya, 554-0024, Japan
Fax: +81-6-6465-0596
nishitani-katsumi@gr.sei.co.jp

ABSTRACT

This paper describes the continuous speech recognition performance in the car environments. Especially various kinds of phoneme models are evaluated. Since the speech recognition performance considerably degrades in the noisy environments, we must cope with this problem in the car environments. There are two primary factors which cause the degradation of the recognition performance. One is the additive noises such as the background noises, and the other is the multiplicative distortion such as the reverberation in the car cabin which is emphasized by the distance between a speaker and a microphone. In this paper, the phoneme models which take the additive noises and the multiplicative distortion into account are trained from the simulated speech data. When the car engine is off, the best word recognition rate is 98.8% for the multiplicative distortion phoneme model which is trained with the speech data generated by the multiplicative distortion simulation. When the car is in the running condition, the best recognition rate is 97.2% for the phoneme model which considers the multiplicative distortion and the additive noises. These results show the effectiveness of the phoneme models which are trained from the simulated speech database in the car environments.

KEYWORDS: Phoneme model, Continuous speech recognition, Car environment, Additive noise, Multiplicative distortion

INTRODUCTION

The speech recognition technology has been rapidly applied to the various fields. Especially, the speech recognition are required as the convenient input media for the portable digital assistances and the car navigation systems. In the *Eurospeech99*, many papers which describe the speech recognition for the car navigation system were presented. The speech data collection efforts in the car environments were also reported [1].

When the speech recognition is used in the real environments, noises are a primary factor for the degradation of the speech recognition performance. In the car environments, there are two types of noises. One is the additive noise caused by the engine and the outside road noises. The other is the multiplicative distortion such as the reverberation in the car cabin. The phoneme models which are robust against these noises are required.

This paper describes the design and the training of the robust phoneme models which are applied to the continuous speech recognition in the car environments. The phoneme models are trained based on the speech data which are generated by the car environment simulation with the background additive noise and the multiplicative distortion (the reverberation in the car cabin). These phoneme models are applied to the car navigation task together with the dictation system *JULIUS* developed by the IPA project [3]. The language model is redesigned for the car navigation task, which includes the address names and the command input.

PHONEME MODELS IN THE CAR ENVIRONMENTS

The robust phoneme models in the car environments have to take the following two types of noises into account. This section explains the two types of noises which degrade the speech recognition performance, and the design and the training of the phoneme models which are robust against those noises in the car environments.

Additive noises. Additive noises degrade the speech wave or the spectrogram additively. Additive noises can be classified into two categories, the stationary noises which are generated by an engine and an air-conditioner, and the non-stationary noises such as the jolting road noises and the wiper sounds.

Multiplicative distortion. The other type of noises is the multiplicative distortion (noises). The multiplicative distortion is represented by the convolution in the time domain (a speech wave) and the multiplication in the spectrum domain. The multiplicative distortion includes the reverberation in the car cabin, and the microphone characteristics. The effect of the multiplicative distortion is not negligible in the car environments, because the several tens meter distance between a speaker and a microphone results in the significant reverberation phenomena. To measure the multiplicative car cabin distortion, we adopt the impulse response which is measured by the TSP (Time Stretched Pulse) method[2] as follows; First, a dummy-head sets up on a passenger seat. Next, a microphone sets up to a sun-visor in front of the passenger seat. Finally, an impulse response from the dummy head to the microphone is measured by the TSP method.

Phoneme models applied the noises. Phoneme HMM models are adopted in this large vocabulary continuous speech recognition experiment. These HMM phoneme models have to be applied to the speech recognition in the car environments, where the additive noises and the multiplicative distortion are contained. A phoneme HMM model, which is trained from the clean speech data, is used as an initial model, and then the phoneme models are trained using the speech data which contain the additive noises and/or the multiplicative distortion. The EM (Expectation-Maximization) algorithm is applied to the phoneme HMM training.

Noisy speech data (Y(t)) which is contained the noises is given by

$$Y(t) = X(t) \otimes A(t) + E(t),$$

where X(t) is clean speech data, A(t) is multiplicative distortion, and E(t) is an additive noise.

LANGUAGE MODEL IN THE CAR ENVIRONMENTS

The original language model from *JULIUS* has to be adapted to the car navigation task. To make a language model for the car navigation task, word dictionary and the language model were created from the address names and the operational commands in the car navigation system. 19800 place name words and 156 operation command words are included. The language models of the word trigram/bigram and the word dictionary (17120 words) are designed from these vocabulary. Table 1 shows the example of a place name and an operation command.

Table 1: Example of Place and Operation Command.

Place	NARAKEN IKOMASHI TAKAYAMACHO
Operation Command	RUTOSHOUKYO (reset the route)

RECOGNITION EXPERIMENTS

JULIUS, which is a large vocabulary continuous speech recognition (LVCSR) engine developed in the IPA (Information-technology Promotion Agency) dictation free software project in Japan, is utilized for the recognition experiments. The recognition rates are calculated by a correct answer rate calculation tool which is also developed in the same IPA project.

Preliminary dictation experiments. The phoneme HMM models are trained from the speech data mixed with the additive noises in the various SNR (Signal-to-Noise Ratio) conditions. The word recognition rates between these phoneme models and the test data under the various SNR conditions are experimentally investigated. The initial clean speech phoneme model (we call *clean model*) is a gender-independent monophone HMM where the number of the mixtures is 16 per HMM state. This initial clean model is supplied by the IPA project. The clean speech data sets are the ASJ Phonetically Ballance Sentences (ASJ-PB) and Japanese Newspaper Article Sentences(JNAS). The total number of utterances is about 45000. The additive noise data are obtained from the car environment sound database by NTT-AT, which is recorded in the driving conditions with the closed windows. The car type is "ESTIMA". In this preliminary experiment, the multiplicative distortion is not mixed in the clean speech data (ASJ-PB and JNAS). To simulate the noisy speech data, we mixed the clean speech data with the NTT-AT additive noise data by controlling the SNR level. Various additive noise phoneme models are trained from the simulated speech data.

The IPA language model used in this dictation experiment is the word trigram/bigram based on the 20000 vocabulary words. The testing speech data are 200 utterances (100 utterances per gender, and each gender includes 23 speakers) selected from JNAS, which are not included in the training set. The noisy testing speech data are also simulated by mixing with the NTT-AT additive noise data. The SNR levels for the experiment are 10dB, 5dB, and 0dB. Fig.1 shows the results of the 20k dictation experiment. The HMM phoneme model trained by the training set with the same SNR as that of the test set shows the best word recognition rate among all phoneme models.

Recognition experiment in the car environments. The IPA dictation system *JULIUS* is used in the command and address name recognition in the real car environments. The language model includes 19800 address names and 156 commands. The four kinds of the phoneme HMM models are used. The first phoneme model is the IPA clean one. The second model is the *additive model* which is trained from the speech data mixed with the NTT-AT additive noise. The third model is the *impulse response(IR) model* which is trained from the speech data generated by the convolution with the impulse response between a speaker and a microphone (multiplicative distortion). The last model is the *additive-IR model* which is trained from the speech data simulated with the additive noise and the impulse response. The additive noise SNR is 10dB in order to adjust the SNR level to the real car environment SNR condition.

The testing speech data is recorded in the real car environments. The testing speech data specification is described in Table 2. These data are uttered in the car "ESTIMA" in two conditions.

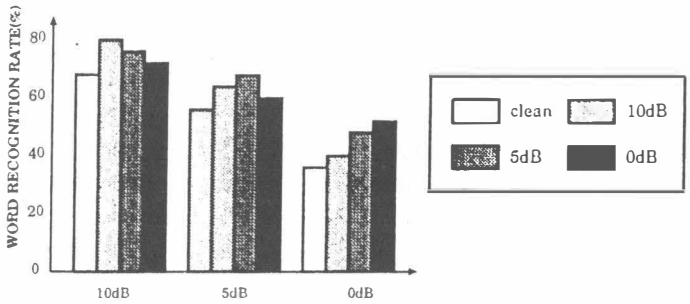

Fig.1 Dictation Experiment in Noisy Conditions for Various Phoneme Models. (JNAS, 20000 words)

One is the condition that the car engine is off (STOP). The other is the running condition (RUN). The SNR is about 10dB in the car running condition. The first decoding pass by the dictation system *JULIUS* is enough for the car navigation task. So only the bigram language model is used. Table 3 shows the experiment result for the car navigation task. The word recognition rate is over 95% for every phoneme model. In the STOP condition, the word recognition rate by the IR model was higher than that by the clean model. In the RUN condition, the additive-IR model shows the highest word recognition rate among the four phoneme models.

Table 2: Testing Speech Data for Car Environments.

Place	Kinki area (6 prefectures)
Command	Control of Car-navigation System
Speaker	3 males and 2 females
Test set	100 utterances (60 addresses, and 40 commands)

Table 3: Word Recognition Rates for Phoneme Models. (

condition \ model	clean	additive	IR	additive-IR
STOP	98.6	98.2	98.8	96.8
RUN	95.5	95.7	94.9	97.2

CONCLUSIONS

In the continuous speech recognition for the car environments, the various kinds of the HMM phoneme models are evaluated. These phoneme models are trained from the speech data simulated with the additive noises and the multiplicative distortion in the car environments.

The highest word recognition rate 98.8 % was attained for the IR model in the car stopping condition. In the car running condition, the additive-IR model, which is simulated the additive noise and the multiplicative distortion by the impulse response, showed the highest word recognition rate 97.2 %. These results showed the effectiveness of the simulated car environment phoneme models, which were trained from the simulated speech data.

REFERENCES

1. "Proc. of Eurospeech99"
2. "On the simulation as a transfer function of an acoustic system (Part 2)," Y. Suzuki, F. Asano, T. Sone, *J. of Acoust. Soc.*, **45**, 44-50 (1986) (in Japanese)
3. "JULIUS - a Japanese Large Vocabulary Continuous Speech Recognition Engine Based on Word N-gram and Multi-Pass Search Strategy," L. Lee, T. Kawahara, S. Doshita, *ASJ Lec. Pap. Ser.*, 51-52 (1998) (in Japanese)
4. "A word correct rate calculation tool based on Japanese morphological characteristics in dictation systems," S. Yamamoto, K. Ito, K. Shikano, S. Nakamura, *ASJ Lec. Pap. Ser.*, 155-156 (1999) (in Japanese)