Performance Prediction of Speech Recognition Using Average-Voice-Based Speech Synthesis

Tatsuhiko Saito¹, Takao Nose¹, Takao Kobayashi¹, Yohei Okato², Akio Horii²

¹Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
²Information Technology R&D Center, Mitsubishi Electric Corporation

Abstract

This paper describes a performance prediction technique of a speech recognition system using a small amount of target speakers’ data. In the conventional HMM-based technique, a speaker-dependent model was used and thus a considerable amount of training data was needed. To reduce the amount of training data, we introduce an average voice model as a prior knowledge for the target speakers’ acoustic models, and adapt it to the target speakers’ ones using speaker adaptation. Experimental results show that the use of average voice model effectively save the amount of training data of the target speakers, and the prediction accuracy is significantly improved compared to the conventional technique especially when a smaller amount of training data was needed. To reduce the amount of training data, a speaker-dependent model was used and thus a considerable amount of target speakers’ data. In the conventional HMM-based technique, a speaker-dependent model was used and thus a considerable amount of speech data was needed. To reduce the amount of speech data, a promising approach is to use average-voice-based speech synthesis [6]. Average voice model is an acoustic model trained using multiple speakers’ speech data. In [2], a word recognition rate was predicted from phoneme recognition rates using transition information between phonemes. A more direct predicting technique was proposed in [2] where between-word distance (BWD) was defined using sub-phonemic units. The criteria calculated from the distribution of BWDs were shown to be well corresponding to word recognition rates. It was also reported that the estimation of the upper limit of the word recognition rate from recognition rates of word pairs [3] would be an effective technique.

Index Terms: speech recognition, performance prediction, average-voice-based speech synthesis, speaker adaptation

1. Introduction

In the development of speech recognition systems for practical applications, it is essential to verify whether the system performance is sufficient for users’ demand. This means that we must evaluate the system using a wide variety of speech samples with various speakers and vocabularies through many experiments with different conditions. However, the cost for preparing such speech data is very expensive and time-consuming. If we can predict the speech recognition rate for arbitrary words or sentences using a small amount of speech data, the evaluation cost would be greatly reduced, which enables us to develop recognition systems more efficiently.

Several attempts related to the prediction of average recognition performance have been made [1–3]. In these techniques, words or sentences are divided into sub-word units such as phonemes, and the recognition performance is predicted as a function of distance or probabilistic measure determined for each sub-word unit. In [1], a word recognition rate was predicted from phoneme recognition rates using transition information between phonemes. A more direct predicting technique was proposed in [2] where between-word distance (BWD) was defined using sub-phonemic units. The criteria calculated from the distribution of BWDs were shown to be well corresponding to word recognition rates. It was also reported that the estimation of the upper limit of the word recognition rate from recognition rates of word pairs [3] would be an effective technique. However, these techniques do not provide intrinsic performance characteristics of the acoustic model of a recognition engine.

Recently, an alternative technique has been proposed where the test speech samples are generated using speech synthesis [4]. This technique utilized HMM-based speech synthesis [5] which can generate natural-sounding synthetic speech for arbitrary sentences using only several tens minutes speech data of a target speaker. One of the advantages of this technique is that it is easy to evaluate the recognition performance in a real acoustic environment by contaminating the generated speech waveform by additive noise and/or reverberation. On the other hand, it is difficult to generate speech without degrading the speaker individuality using only a few minutes or less amount of speech data of the target speaker. Therefore, the cost for preparing speech data is still high when the number of test speakers increases in the development of speech recognition system.

In this paper, we employ the average-voice-based speech synthesis [6] in the recognition performance prediction to reduce the required amount of speech data of the test speakers. The average voice model is trained using multiple speakers’ speech data and can be used as a prior knowledge of the acoustic model. By combining with a speaker adaptation technique, we can obtain the model for synthesis using a less amount of speech data of the target speaker. We examine the prediction accuracy of the proposed average-voice-based technique by comparing to the conventional speaker-dependent technique through phoneme and isolated word recognition experiments.

2. Recognition performance prediction based on speech synthesis

2.1. Performance prediction using HMM-based speech synthesis

Instead of preparing real speech data of a target speaker, an HMM-based speech synthesis framework was used to generate arbitrary test patterns for the assessment of speech recognition system in [4]. In this technique, first, a speaker-dependent model is trained for each target speaker using a sufficient amount of his/her speech data. Secondly, synthetic speech waveforms are generated for arbitrarily given texts of test words or sentences using the speaker-dependent models. Next, all the synthetic speech samples are fed into the target speech recognition system. Finally, the performance of the target system is assessed by analyzing the recognition results of the target speakers.

2.2. Average-voice-based approach to performance prediction

A problem of the conventional technique of [4] using a speaker-dependent model is that the prediction capability directly depends on the amount of available training data of the target speaker. To maintain the prediction accuracy with a less amount of speech data, a promising approach is to use average-voice-based speech synthesis [6]. Average voice model is an acoustic model trained using multiple speakers’ speech data. In the average voice model, the speakers’ characteristics are averaged by the statistical expectation process of model training, and most
of the speech property of a certain speaker is lost. This implies that the average voice model has native information of the acoustic characteristics of the target language. Using the average voice model as a prior information, we can efficiently obtain the target speaker’s model with a smaller amount of training data than the speaker-dependent case by applying a speaker adaptation technique such as maximum likelihood linear regression (MLLR) [7]. In the model training and adaptation of this study, we use hidden semi-Markov model (HSMM) with explicit state-duration distributions [8]. A block diagram of the proposed technique is shown in Fig. 1.

2.3. Speaker adaptation based on linear transforms

As for the speaker adaptation of this study, we use the HSMM-based constrained structural maximum a posteriori linear regression (CSMAPLR) algorithm [9] as a linear transformation algorithm which gives better performance and robustness than conventional MLLR-based ones. In the CSMAPLR adaptation, the mean vector of the average voice model and the covariance matrix of the output pdf, \( \mu_i \) and \( \Sigma_i \), respectively, and the mean and variance of the state-duration pdf, \( \tilde{m}_i \) and \( \tilde{\sigma}_i^2 \), respectively, are linearly transformed as follows:

\[
\hat{\mu}_i = \zeta \mu_i - \epsilon \quad (1)
\]
\[
\hat{\Sigma}_i = \zeta \Sigma_i \zeta^T \quad (2)
\]
\[
\hat{m}_i = \chi m_i - \nu \quad (3)
\]
\[
\hat{\sigma}_i^2 = \chi \sigma_i^2 \chi^T \quad (4)
\]

where \( \zeta \) and \( \epsilon \) are the transformation matrix and the bias vector of the output pdf, respectively, and \( \chi \) and \( \nu \) are the transformation coefficient and the bias term of the state-duration pdf, respectively. These transformation parameters are estimated based on structural MAP criterion [10] in which the structure of the decision tree is utilized for robust estimation. In the estimation process, the transformation parameters of the parent node are used as the parameters of the prior distribution of the transformation parameters to be estimated [11].

2.4. Canonical model training with STC and SAT

In the decision-tree-based context clustering for average voice model, the nodes of the decision tree do not always have training data of all speakers, and some nodes could have data from only a single speaker. This speaker-biased node causes degradation of quality of average voice and synthetic speech after speaker adaptation. To alleviate this problem, shared-decision-tree-based context clustering (STC) [12] is known to be effective. In STC, a common tree structure is constructed for all speakers, and each node has training data of all speakers.

When training data sets of respective speakers differ widely, we encounter another problem that the distributions of average voice model have bias depending on speaker and/or gender. This would also lead to the degradation of the quality of synthetic speech. For this problem, speaker adaptive training (SAT) is a well-known technique to train a canonical model. In this study, we examine whether these speaker normalization techniques improve the prediction accuracy of the recognition performance.

3. Database and speech synthesis conditions

We describe here the experimental conditions for speech synthesis used both in phoneme and isolated word recognition tasks. We used a Japanese speech database which consisted of short utterances including addresses, numbers, and facilities names called the point of interests (POIs). There are 43 male and the same number of female non-professional speakers, and each of them uttered about 500 words and sentences. For an average voice model training, we used the ATR Japanese speech database set B. It includes six male and four female professional narrators’ speech data. The number of training sentences for the average voice model were 450 sentences per speaker, 4,500 sentences in total. The number of POI words was about 250 (7 minutes in total) per speaker, and we used these words as the test utterances. The data sets for training and adaptation were chosen so as not to be included in the test data.

Speech signals were sampled at a rate of 16 kHz and windowed by a 25-ms Blackman window with 5-ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis. The feature vector consisted of 25 mel-cepstral coefficients including the zeroth coefficient and log \( F_0 \), and their delta and delta-delta coefficients, and total dimensionality of the feature vector became 78. We used five-state left-to-right model with diagonal covariance matrices. In the HMM-based speech synthesis a context-dependent model is generally used, where phonetic and prosodic contextual factors are taken into account. However, we used only triphone context because it was found in preliminary experiments that the difference of prediction performance between cases with and without using other contexts was not significant. Furthermore, this leads to being cost-effective because of nonnecessity of label creation using a variety of contextual factors. In parameter generation, we did not use any post processing method for spectral enhancement such as [13] since preliminary experimental results have not shown its effectiveness.

In the following experiments, we evaluated three types of synthetic speech generated from 1) speaker-dependent model, 2) speaker-adapted model, and 3) speaker-adapted model with canonical model training described in Sect. 2.4. We evaluated the proposed technique by phoneme and isolated word recognition tasks.

4. Evaluation on phoneme recognition

4.1. Recognition conditions

For the training of the speaker-independent model in the phoneme recognition, we used 50 male and 50 female speech data taken from the JNAS (Japanese newspaper article sentence) database [14]. The number of training sentences was 100 per
Figure 2: PERs of synthetic and natural speech samples for respective speakers in case of 50 utterances training or adapting. Each cross mark corresponds to the result for one speaker.

4.2. Results

Figure 2 shows PERs of synthetic and natural speech samples for respective speakers. “SD” is the result of speaker-dependent model trained using 50 utterances. “SA” and “SA(STC+SAT)” are the results of speaker-adapted model using same 50 utterances for adaptation without and with canonical model training. To give objective measures of prediction accuracy, we also show correlation coefficients and RMS errors of PERs between synthetic and natural speech samples. From the results, when using 50 utterances for adapting or training, we can see the RMS error of SD became large, while that of SA was much smaller. Moreover, the use of canonical model training also slightly improved the prediction accuracy.

Figure 3 shows correlation coefficients and RMS errors when the number of utterances for training and adapting is changed. It is noted that we did not evaluate the SD with less than 50 utterances since the amount of training data was not sufficient. From the viewpoint of correlation, the speaker-dependent model gave comparable performance to the speaker-adapted model even if the amount of training data is not sufficient. On the other hand, the RMS errors of phoneme recognition rates of the speaker-dependent model substantially increased when the amount of training data became smaller. This means that the conventional speaker-dependent technique gives only the relative performance of speech recognition between speakers. To obtain the more accurate prediction performance of the recognition systems, the proposed average-voice-based technique seems to be suitable when the amount of training data of the target speaker is limited.

5. Evaluation on isolated word recognition

5.1. Recognition conditions

For the isolated word recognition, we used Julius [15] as the decoder. We used an acoustic model included in Julius dictation kit v4.0. To increase the number of vocabularies in the dictionary, we added the words included in the database of Japanese geographic information system (GIS) [16] to POI words. The number of words included in the GIS database is about 190,000. Moreover, we decomposed GIS entries into morphemes, and added part of these morphemes into the dictionary. As a result, the number of words included in word dictionary became 946,373. The pronunciation data was automatically created by Japanese text analysis tool, MeCab [17]. In the decoding process, we set the beam width to 1,000. Speech signals were sampled at a rate of 16 kHz and windowed by a 25-ms Hamming window with 10-ms shift. The feature vector consisted of 1–12th MFCCs, log energy, and their delta, and the total dimensionality of the vector became 26. We used three-state left-to-right model with diagonal covariance matrices. The number of mixtures was set to 16 based on the result of a preliminary experiment. We used phonetic networks based on Japanese phonetic concatenation rules. The phoneme error rate (PER) is given by

$$\text{PER}(\%) = \frac{D + S}{N} \times 100,$$  

(5)

where $D$ and $S$ are the numbers of deletion and substitution errors, respectively, and $N$ is the total number of phonemes in the reference transcriptions.

5.2. Results

Figure 4 shows WERs of synthetic and natural speech samples for respective speakers. Each cross mark corresponds to the result for one speaker. Compared to the results of phoneme recognition, the recognition rates of synthetic speech became slightly higher, and this caused the degradation of the prediction accuracy.
In this paper, we examined the use of average-voice-based speech synthesis for predicting speech recognition performance. Experimental results showed that the proposed average-voice-based technique significantly improved the prediction accuracy of the conventional speaker-dependent technique when the amount of target speaker’s data is limited. It was also shown that valid correlation of recognition error rates between synthetic and natural speech in phoneme and isolated word recognition. However, it is also clear that the prediction accuracy was significantly improved by speaker-adapted model when we used a small amount of target speaker’s speech. Figure 5 shows correlation coefficients and RMS errors when the number of utterances for training and adapting is changed. The results show the similar tendency to those of the phoneme recognition, and the proposed technique works better than the conventional one. It is also seen that the performance of the speaker-adapted model does not change significantly when the number of adaptation utterances is more than 10.

6. Conclusions

In this paper, we examined the use of average-voice-based speech synthesis for predicting speech recognition performance. Experimental results showed that the proposed average-voice-based technique significantly improved the prediction accuracy of the conventional speaker-dependent technique when the amount of target speaker’s data is limited. It was also shown that valid correlation of recognition error rates between synthetic and natural speech in phoneme and isolated word recognition. In the future work, we will examine the prediction accuracy in other environmental conditions such as LVCSR system and in English corpus.

7. References