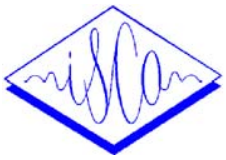
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Acoustic model adapatation for coded speech using synthetic speech

Shingo Kuroiwa, Koji Tanaka, Satoru Tsuge, Fuji Ren

Department of Information Science & Intelligent Systems
Faculty of Engineering, The University of Tokushima, Japan
kuroiwa@is.tokushima-u.ac.jp

10.21437/Interspeech.2004-17

Abstract

In this paper, we describe a novel acoustic model adaptation technique which generates "speaker-independent" HMM for the target environment. Recently, personal digital assistants like cellular phones are shifting to IP terminals. The encoding-decoding process utilized for transmitting over IP networks deteriorates the quality of speech data. This deterioration causes degradation in speech recognition performance. Acoustic model adaptations can improve recognition performance. However, the conventional adaptation methods usually require a large amount of adaptation data. The proposed method uses HMM-based speech synthesis to generate adaptation data from the acoustic model of HMM-based speech recognizer, and consequently does not require any speech data for adaptation. Experimental results on G.723.1 coded speech recognition show that the proposed method improves speech recognition performance. A relative word error rate reduction of approximately 12% was observed.

1. Introduction

In recent years, telephone speech recognition systems encompassing thousands of vocabularies have become practical and widely used. These systems are generally utilized by automatic telephone services for booking an airline ticket, inquiring about stock, receiving traffic information, and so on. However, the recognition accuracy of cellular phones is still inadequate due to compression coding or ambient noise.

Recently, personal digital assistants like cellular phones are shifting to IP terminals. For transmission over IP networks, speech data must be encoded at the sending end and subsequently decoded at the receiving end. This coding process deteriorates the quality of the voice data. Although most people can not notice this deterioration, it seriously affects the performance of those speech recognizers not designed for low-quality voice data[1]. The major causes of speech recognition performance degradation are : distortion in the transmission environment (transmission error and packet loss), and low bitrate speech coding (loss of speech information). These distortions cause a mismatch between the feature vectors of input speech and acoustic models[2, 3].

The best way to overcome this degradation is by collecting a large amount of data in the target enviroment and training acoustic models using them. However, this method requires huge costs. Adaptation methods, such as MLLR (Maximum Likelihood Linear Regression) [4] or MAP (Maximum A Posterior probability)[5] also require a large amount of adaptation data to estimate "speaker-independent" models[6]. Actually, in [6] they used at least 1,000 utterances from 30 speakers to estimate codec-dependent HMMs.

We propose in this paper a novel adaptation method utilizing HMM-based speech synthesis. This method does not need coded speech data for adaptation because these data are generated by speech synthesis from the acoustic model. By using the generated data after coding, the acoustic model is adapted to coded speech. Consequently, this method can adapt the acoustic model to various coded speech without any speech data if the coding method is specified.

Presented in section **2** is the effect of the speech coder for use with IP telephones on speech recognition. Section **3** presents our approach and section **4** presents our experiments, followed by conclusions and an outline for further work.

2. Infuluence of the speech corder on speech recognition

2.1. Baseline method

In coded speech recognition, the easiest and ideal method is to use an acoustic model that is trained with coded speech. A diagram of this training method is shown in Figure **1**. This method requires a large quantity of coded speech for training an accurate acoustic model.

In order to verify the effect of the speech coder on speech recognition, we evaluated recognition performance using an acoustic model which was trained with coded speech.

For the speech coder, we selected the G.723.1 Annex A (6.3 and 5.3 kbps) which has the lowest bitrate in the ITU-T H.323 recommendation.

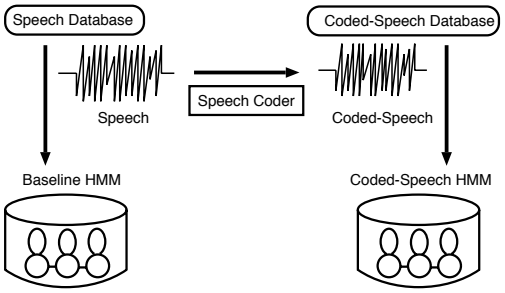


Figure 1: A diagram of training method

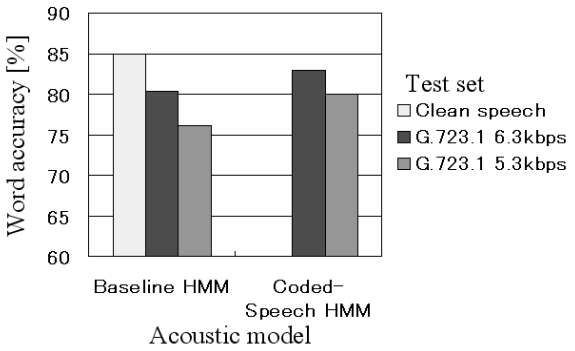


Figure 2: Evaluation results of the baseline and coded speech HMM

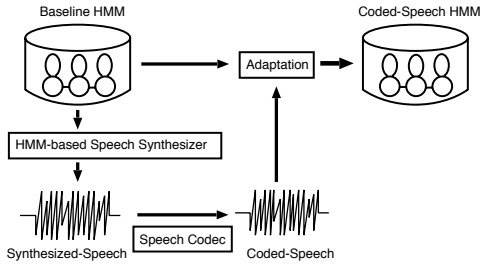


Figure 3: Adaptaion method using speech sunthesis based on HMM

2.2. ITU-T G.723.1 speech corder

The G.723.1 standard is an analysis-by-synthesis linear predictive coder and it provides a dual coding rate at 6.3 and 5.3 kbps. For the higher rate of 6.3kbps, the encoder uses multipulse maximum likelihood quantization (MP-MLQ). For the lower late of 5.3kbps, the encoder employs an algebraic code excited linear prediction (ACELP) scheme. An option for variable rate operation is available using voice activity detection (VAD), which compresses the silent portions.

2.3. Experimental conditions

The baseline acoustic models were trained with ASJ speech databases of phonetically balanced sentences (ASJ-JNAS). Training data consist of 5,168 utterances (sampled 8kHz and 16bit) from 103 male speakers. The coded speech acoustic models were trained with the same training data but they were coded by G.723.1 Annex A (6.3kbps, 5.3kbps). For the test set, 100 utterances (1,578 words) from 23 male speakers were used. The speech signals were windowed by a 25ms Hamming window with a 10ms shift, the mel-cepstral coefficients were obtained by mel-cepstral analysis[7, 8]. The 39-dimensional feature vector was comprised of 13 mel-cepstral coefficients (0-12th with CMS) including their delta and delta-delta coefficients.

For the acoustic model, shared state triphone HMMs with sixteen Gaussian mixture components per state were trained. The number of states was approximately 1,000.

2.4. Evaluation of coded speech model

Figure **2** shows evaluation results of the baseline and coded speech HMM. From the figure, the word accuracy of the coded speech is lower than that of uncoded speech. Also, coded speech HMM achieved higher performance than that of uncoded speech HMM.

These results indicate that HMM trained with coded speech data can improve the recognition performance for coded speech. However, it is difficult to obtain a large quantity of coded speech data for every coding method. Therefore, HMM adaptation methods that do not require large quantities of coded speech data are desired.

3. Adaptation using synthetic speech

A problem of conventional HMM adaptation methods is that they require a large quantity of training data for adaptation. The proposed method generates adaptation data from the HMM using a HMM-based speech synthesizer. Figure **3** shows a diagram of the proposed method. This method consists of HMM-based speech synthesis and HMM adaptation by the coded synthetic speech. The proposed adaptation process is : First, the HMM-based speech synthesizer generates (1) 503 phonetically-ballanced sentences[9], or (2) speech waveforms corresponding to all output distributions of all states. Next, a speech coder encodes and decodes these waveforms. Finally, baseline HMM is adapted using the encode-and-decode waveforms. This method does not require speech data for adaptation and it is applicable to any coder if the input and output of a waveform is known.

3.1. HMM-based speech synthesis

In this section, we describe the the HMM-based speech parameter generation algorithm according to [7].

Let $\boldsymbol{q} = \{q_1, q_2, \cdots, q_T\}$ be the state sequence and $\boldsymbol{o}$

$= [\boldsymbol{o}'_1, \boldsymbol{o}'_2, \cdots, \boldsymbol{o}'_T]'$ be the vector of the output parameter sequence generated along with a single path $\boldsymbol{q}$ in the same manner as the Viterbi algorithm. The output distribution of each state is assumed to be a single Gaussian distribution for convenience of explanation.

For a given continuous HMM λ, the output speech parameter sequence $\boldsymbol{o}$ is obtained by maximizing $P(\boldsymbol{q}, \boldsymbol{o}|\lambda, T)$ with respect to $\boldsymbol{q}$ and $\boldsymbol{o}$. Since all HMMs used in the system were left-to-right models with no skipping, the probability of state sequence $\boldsymbol{q}$ is determined only by explicit state duration densities $p_q(d_q)$, i.e., the probability of d_q consecutive observations in state q. This HMM λ is the baseline HMM in Figure **3**. Let a_d be a scaling factor on state duration scores and Const. be the normalization factor of Gaussian distributions, then the logarithm of $P(\boldsymbol{q}, \boldsymbol{o}|\lambda, T)$ can be written as :

$$\begin{aligned}
&\log P(\boldsymbol{q}, \boldsymbol{o}|\lambda, T) \\
&= a_d \log(\boldsymbol{q}|\lambda, T) + \log P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T) \\
&= a_d \sum_{k=1}^{K} \log p_{qk}(d_{qk}) - \frac{1}{2}\log|\boldsymbol{\Sigma}| \\
&\quad -\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{o}-\boldsymbol{\mu}) - \text{Const.} \qquad (1)
\end{aligned}$$

where

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_{q1}, \boldsymbol{\mu}_{q2}, \cdots, \boldsymbol{\mu}_{qT}] \qquad (2)$$

$$\boldsymbol{\Sigma} = \text{diag}[\boldsymbol{\Sigma}_{q1}, \boldsymbol{\Sigma}_{q2}, \cdots, \boldsymbol{\Sigma}_{qT}] \qquad (3)$$

and $\boldsymbol{\mu}_{qt}$ and $\boldsymbol{\Sigma}_{qt}$ are the mean vector and the covariance matrix associated with state q_t, respectively. We assume that the total number of states which have been visited during T frames is K ($\sum_{k=1}^{K} d_{qk} = T$).

By using a MLSA (Mel Log Spectral Approximation) filter[10] speech is synthesized from the generated sequence $\boldsymbol{o}$.

3.2. Acoustic model adaptation using synthesized 503 sentences

The first adaptation method is simple. First, the HMM-based speech synthesizer generates 503 phonetically-ballanced sentences. We consider that the synthsized speech was uttered by one "speaker-independent" speaker because it is synthesized from a "speaker-independent" HMM. Next, a speech coder encodes and decodes these utterances. Finally, a coded-speech HMM is estimated with MLLR using these coded utterances. One full matrix for a global regression class is used as a transformation matrix of MLLR.

3.3. Acoustic model adaptation using waveforms corresponding to mean vectors

The second proposed method uses synthetic speech segments corresponding to the mean vector of each output distribution.

The output distribution of state i is defined as follows:

$$b_i(\boldsymbol{o}) = \sum_{k=1}^{M} c_{im}\mathcal{N}(\boldsymbol{o}, \boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}) \quad 1 \le i \le N \qquad (4)$$

$$\begin{aligned}
&\mathcal{N}(\boldsymbol{o}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\boldsymbol{o}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{o}-\boldsymbol{\mu})\right) \qquad (5)
\end{aligned}$$

where M is the number of Gaussian mixtures, N is the number of states, $\boldsymbol{\mu}_{im}$ is the mean vector and $\boldsymbol{\Sigma}_{im}$ is the covariance for the output probability functions of mixture m at state i. c_{im} is the mixture weight of mixture m at state i.

The following process is performed on all output distributions:

1. Generate L frame speech from the mean vector $\boldsymbol{\mu}_{im}$, using the speech synthesis algorithm described in Section **3.1**.
2. Encode and decode the synthetic speech by a speech coder and decoder.
3. Analyze the mel-cepstral coefficients ($\boldsymbol{c}'_1$,$\boldsymbol{c}'_2$, $\cdots$, $\boldsymbol{c}'_L$) from the coded speech.
4. Replace the mean mel-cepstral coefficients $\overline{\boldsymbol{c}}'_L = \frac{\sum_{l=1}^{L}\boldsymbol{c}'_l}{L}$ with the mean vector $\boldsymbol{\mu}_{im}$ of the HMM. The delta and delta-delta parameters are not adapted.

4. Experiments

To evaluate the proposed methods, we compared the recognition performance of the following HMMs :

(1) the HMM trained by uncoded speech (*baseline model*),

(2) the HMM adapted using 503 synthetic speech described in Section **3.2** (*503 sentences*),

(3) the mean-vector adapted HMM described in Section **3.3** (*mean-vector based*),

(4) the HMM trained with coded training data (*coded speech HMM*).

Coded speech HMM shows the upper limit of the proposed method.

Table 1: Word Accuracy of proposed method for G.723.1 coded speech (%)

HMM or adaptation method	bitrate of coded speech 6.3kbps	5.3kbps
baseline model	80.4	76.1
503 sentences	76.3	73.5
mean-vector based	**82.0**	**79.0**
coded speech HMM	83.0	80.7

4.1. Conditions

Acoustic analysis and HMM training conditions are the same as detailed in Section **2.3**. The feature vector is comprised of 13 mel-cepstral coefficients (0-12th), and their delta and delta-delta coefficients. For the acoustic model, shared state triphone HMMs with sixteen Gaussian mixture components per state were trained.

For the *mean-vector based* adaptation method, the adaptation data corresponding to each distribution of each state, were generated at a 150 Hz pitch frequency by the MLSA filter. From a preliminary experiment, the adaptation data for unvoiced sounds were also excited with a 150 Hz pitch. For this experiment the length of each data was 0.3 seconds.

4.2. Experimental results

The experimental results are provided in Table **1**. As shown by this table, the *mean-vector based* adaptation method improves the word accuracy of the *baseline model*. The proposed method was effective in both G.723.1 Annex A coders of 6.3kbps and 5.3kbps. We observed an improvement in word accuracy of approximately 1.5 points (8% relative error reduction) at 6.3kbps and about 3 points (12% relative error reduction) at 5.3kbps. The proposed method slightly degrades the recognition performance of the *coded speech HMM*. However, the proposed method did improve the recognition performance of the *baseline* without any training speech.

On the other hand, the HMM adapted using 503 synthetic speech did not improve the accuracy. One reson for this discrepancy may be that we did not study the structure of MLLR transformation matrix sufficiently. We also expect an essential problem is that the synthetic speech was one average speaker's voice.

5. Summary

Acoustic model adaptation methods using syntetic speech segments for coded speech recognition have been presented. The proposed methods generate adaptation data from the HMM for recognition and adapt the HMM using these data. Experimental results of G.723.1 coded speech recognition indicate that the proposed mean-vector based adaptation method improved the recognition accuracy of coded speech when compared to the non-adaptation HMM. Additionally the word accuracy of the proposed method approaches that of the coded speech HMM.

For this study, only the MFCC mean vectors of HMM were adapted. Currently we are trying to adapt the covariance matrix. In the future, the proposed method will be adapted to other coding and environments.

6. Acknowledgements

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 14350204, 14380166, Grantin-Aid for Young Scientists (B), 15700163, 2004, Hoso-Bunka Foundation (HBF) and International Communications Foundation (ICF).

7. References

[1] Lilly !$B.T. and Paliwal !$K.K., "Effect of Speech Coders on Speech Recognition Performance", ICSLP-96, pp.877–880, 1996.

[2] Gallardo-Antolin, A., Pelaez-Moreno, C., and Diaz-de-Maria, F., "A robust front-end for ASR over IP and GSM networks: an integrated sinario" Eurospeech-2001, pp. 1099-1102, Sep. 2001.

[3] Tanaka !$K., Kuroiwa, S., Tsuge, S. and Ren, F., "The Influence of Speech Coders for IP Telephone on Speech Recognition" !8Information-2002, Vol.3, pp.44–48, July 2002.

[4] Leggetter, C.J. and Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models" !% Computer Speech and Language, Vol.9, pp.171-185, 1995

[5] Gauvian, J.L. and Lee, C.H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains" !8IEEE Trans. Speech Audio Process., Vol.2, pp.291-298, April 1994.

[6] Naito, M., Kuroiwa, S., Kato, T., Shimizu, T. and Higuchi, N., "Rapid CODEC Adaptation for Cellular Phone Speech Recognition" Eurospeech-2001, pp.857–860, Sep. 2001.

[7] Tokuda, K., Yoshimura, T., Masuko,T., Kobayashi, T. and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis," ICASSP-2000, pp.1315–1318, June 2000.

[8] "Speech Signal Processing Toolkit" !% http://kt-lab.ics.nitech.ac.jp/tokuda/SPTK/

[9] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H. and Shikano, K., "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis", Speech Communication, Vol.9, pp.357–363, Sep. 1990.

[10] Imai, H., "Cepstral analysis synthesis on the mel frequency scale" ICASSP-83, pp.93–96, 1983.