Acoustic Model Adaptation for Codec Speech based on Learning-by-Doing Concept

Shingo Kuroiwa, Satoru Tsuge, Koji Tanaka, Kazuma Hara, and Fuji Ren

Faculty of Engineering, The University of Tokushima,
Tokushimashi 770-8506, Japan
{kuroiwa,tsuge,ren}@is.tokushima-u.ac.jp,
WWW home page: http://a1-www.is.tokushima-u.ac.jp/

Abstract. 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. In this paper, we propse a novel acoustic model adaptation technique that generates "speaker-independent" HMM for the target environment based on the learning-by-doing concept. 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. By using the generated data after coding, the acoustic model is adapted to codec speech. Experimental results on G.723.1 codec speech recognition show that the proposed method improves speech recognition performance. A relative word error rate reduction of approximately 12% was observed.

Keywords: Speech Recognition, Model Adaptation, Codec Speech, Speech Synthesis, Learning-by-Doing

1 Introduction

In recent years, telephone speech recognition systems encompassing thousands of vocabularies have become practical and widely used [1,2]. 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 [3–5].

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[6]. 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

© A. Gelbukh (Editor)
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 105-114

of speech information). These distortions cause a mismatch between the feature vectors of input speech and acoustic models[7, 4].

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) [8] or MAP (Maximum A Posterior probability)[9] also require a large amount of adaptation data to estimate "speaker-independent" models[3]. Actually, in [3] they used at least 1,000 utterances from 30 speakers to estimate codec-dependent HMMs.

We propose in this paper novel adaptation methods based on a learning-by-doing concept, in which a speech recognition system utters sentences in a target environment and adapts acoustic models by listening to them. This method does not need codec 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 codec speech. Consequently, this method can adapt the acoustic model to various codec 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 future work.

2 Infuluence of the speech codec on speech recognition

2.1 Baseline method

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

For the speech coder, we selected the G.723.1 Annex A (6.3 and 5.3 kbps)[10] which has the lowest bitrate in the ITU-T H.323 recommendation[11].

2.2 ITU-T G.723.1 speech codec

The G.723.1 standard[10] 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.

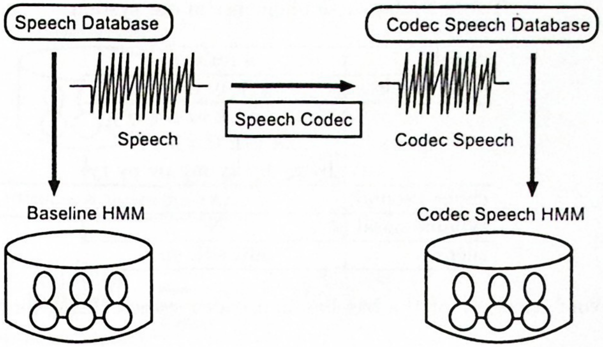

Fig. 1. A diagram of training method

Table 1. Acoustic analysis conditions

sampling rate	8 kHz
window	Hamming
FFT point	256
frame length	25 ms
frame shift	10 ms
feature vector	0-12 mel-cepstral coefficients[13](CMS) + delta + delta-delta (total 39)

2.3 Experimental conditions

The baseline acoustic models (*baseline model*) were trained with ASJ speech databases of phonetically balanced sentences[15]. Training data consist of 5,168 utterances (sampled 8kHz and 16bit) from 103 male speakers. The codec speech acoustic models (*codec speech HMM*) were trained with the same training data but they were coded by G.723.1 Annex A (6.3kbps, 5.3kbps). For the open test set, 100 utterances (1,578 words) from 23 male speakers, which are the utterences of Japanese standard dictation task in Japanese newspaper article sentence speech corpus (JNAS)[15], were used. The acoustic analysis conditions are shown in Table 1. The speech signals were windowed by a 25ms Hamming window with a 10ms shift, the mel-cepstral coefficients were obtained by mel-cepstral analysis[12, 13]. The 39-dimensional feature vector was comprised of 13 mel-cepstral coefficients (0-12th with CMS) including their delta and delta-delta coefficients. We utilized a SPTK[13] for acoustic analysis and HTK[14] for HMM training.

Table 2 shows Japanese phoneme set in our system. The "silB" and "silE" denote the silence at the beginning/end of the speech. 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. We used

Table 2. Japanese phonemes in our system

vowels	a i u e o
long vowels	a: i: u: e: o:
consonants	b d g p t k m n r w y ch j sh ts f h s z by gy hy ky my ny py ry
choked sound	q
syllabic nasal	N
silence	silB silE sp

Table 3. Word accuracy of the baseline and codec speech HMM for G.723.1 codec speech

Acoustic Model	bitrate of codec speech	
	6.3kbps	5.3kbps
baseline model	80.4 %	76.1 %
codec speech HMM	83.0 %	80.7 %

a Julius[16, 17] for the recognizer with a 20,000-word lexicon and the *tri*-gram language model.

The recognition performance is evaluated by word accuracy using the follwing equation :

$$Accuracy = \frac{N - D - S - I}{N} \cdot 100\ (\%), \tag{1}$$

where N is the total number of words, D is the number of deletions, S is the number of substitutions, and I is the number of insertions.

2.4 Evaluation of codec speech model

Table 3 shows evaluation results of the *baseline model* and *codec speech HMM*. From the figure, the word accuracy of the codec speech is lower than that of uncoded speech. Also, *codec speech HMM* achieved higher performance than that of uncoded speech HMM (*baseline model*).

These results indicate that HMM trained with codec speech data can improve the recognition performance for codec speech. However, it is difficult to obtain a large quantity of codec speech data for every coding method. Therefore, HMM adaptation methods that do not require large quantities of codec 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 2 shows

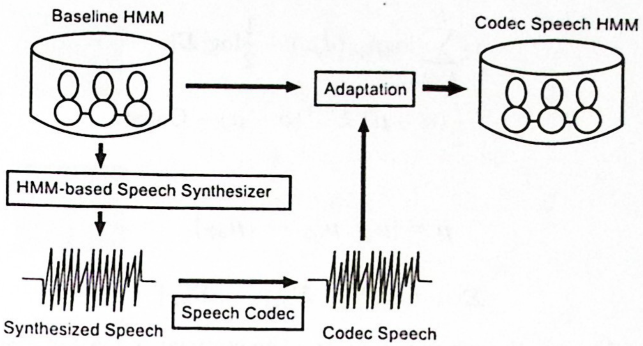

Fig. 2. Adaptaion method using speech synthesis based on HMM

a diagram of the proposed method. This method consists of HMM-based speech synthesis and HMM adaptation by the codec synthetic speech. The proposed adaptation process is : First, the HMM-based speech synthesizer generates (1) 503 phonetically-ballanced sentences[18], 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 [12].

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 2. 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 :

$$\begin{aligned}&\log P(\boldsymbol{q}, \boldsymbol{o}|\lambda, T)\\&= a_d \log(\boldsymbol{q}|\lambda, T) + \log P(\boldsymbol{o}|\boldsymbol{q}, \lambda, T)\end{aligned}$$

$$= a_d \sum_{k=1}^{K} \log p_{qk}(d_{qk}) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu) - \text{Const.} \quad (2)$$

where

$$\mu = [\mu_{q1}, \mu_{q2}, \cdots, \mu_{qT}] \quad (3)$$

$$\Sigma = \mathrm{diag}[\Sigma_{q1}, \Sigma_{q2}, \cdots, \Sigma_{qT}] \quad (4)$$

and μ_{qt} and Σ_{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[19] speech is synthesized from the generated sequence 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[18]. 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 codec-speech HMM is estimated with MLLR using these codec utterances. One full matrix for a global regression class is used as a transformation matrix of MLLR[8].

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:

$$b_i(o) = \sum_{k=1}^{M} c_{im}\mathcal{N}(o, \mu_{im}, \Sigma_{im}) \quad 1 \le i \le N \quad (5)$$

$$\mathcal{N}(o, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp(-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)) \quad (6)$$

where M is the number of Gaussian mixtures. N is the number of states. μ_{im} the mean vector and Σ_{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:

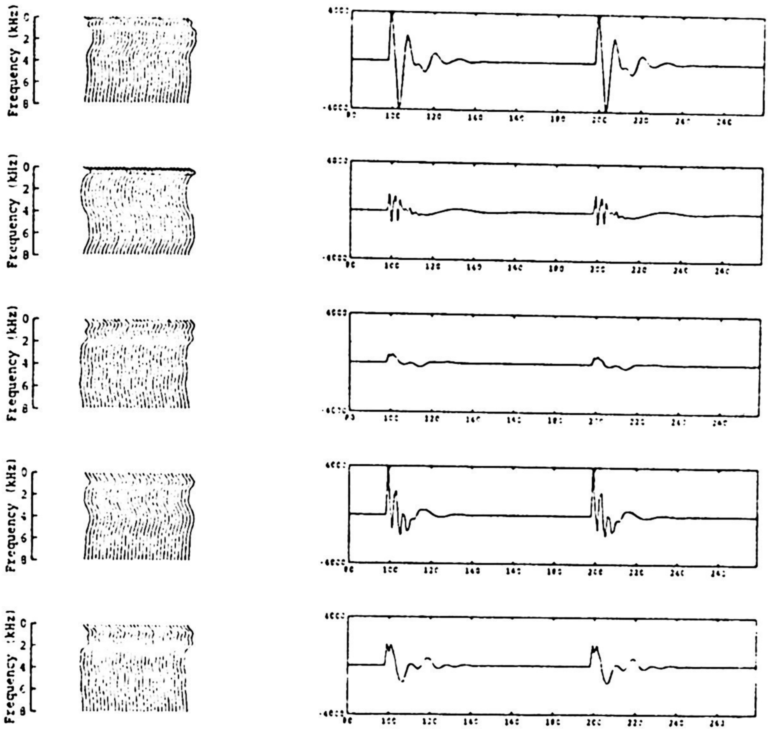

Fig. 3. Japanese vowel (/a/,/i/,/u/,/e/,/o/) spectrums and waveforms generated from the mel-cepstral coefficients

1. Generate L frame speech from the mean vector μ_{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 $(c'_1, c'_2, \cdots, c'_L)$ from the codec speech.
4. Replace the mean mel-cepstral coefficients,

$$\bar{c}'_L = \frac{\sum_{l=1}^{L} c'_l}{L}, \tag{7}$$

with the mean vector μ_{im} of the HMM. The delta and delta-delta parameters are not adapted in this paper.

Figure 3 shows an example of the Japanese vowel (/a/,/i/,/u/,/e/,/o/) spectrums and waveforms generated from the mel-cepstral coefficients of the HMM mean vectors.

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 codec speech training data (*codec speech HMM*).

We consider that *codec speech HMM* shows the upper limit of the proposed method.

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 4. 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 *codec 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.

Table 4. Word accuracy of proposed method for G.723.1 codec speech

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

5 Summary

In this paper, we propose novel acoustic model adaptation methods based on a learning-by-doing concept, in which a speech recognition system utters sentences in a target environment and adapts acoustic models by listening to them. The proposed methods generate adaptation data from the acoustic models (HMMs) by using HMM-based speech synthesis. By using the generated data after coding, the system adapt the acoustic models using these data. Experimental results of G.723.1 codec speech recognition indicated that the proposed mean-vector based adaptation method improved the recognition accuracy of codec speech when compared to the non-adaptation HMM. Additionally the word accuracy of the proposed method approaches that of the codec 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.

Acknowledgments

This research has been partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 15700163, Grant-in-Aid for Exploratory Research, 17656128, 2005, Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (B), 14380166, 17300065, 2005, and International Communications Foundation (ICF).

References

1. Dahl, D. (ed.): Pracrical Spoken Dialog Systems, Kluwer Academic Publishers, Dordecht (2004)
2. Cohen, M., Giangola, J., Balogh, J.: Voice User Interface Design, Addison Wesley, Boston (2004)
3. Naito, M., Kuroiwa, S., Kato, T., Shimizu, T. and Higuchi, N.: "Rapid CODEC Adaptation for Cellular Phone Speech Recognition" Proc. European Conf. Speech Communication and Technology (2001) 857–860
4. TanakaK., Kuroiwa, S., Tsuge, S. and Ren, F.: "The Influence of Speech Coders for IP Telephone on Speech Recognition" Proc. Int. Conf. Information-2002, **3**, (2002) 44–48

5. Kato, T., Naito, M., and Shimizu, T.: "Noise-Robust Cellular Phone Speech Recognition Using Codec-Adapted Speech and Noise Models", IEEE Int. Conf. Acoustics, Speech, and Signal Processing (2002) 1315–1318
6. LillyB.T. and PaliwalK.K.: "Effect of Speech Coders on Speech Recognition Performance", Proc. Int. Conf. Spoken Language Processing (1996), 877–880
7. 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" Proc. European Conf. Speech Communication and Technology (2001) 1691–1694
8. Leggetter, C.J. and Woodland, P.C.: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models" Computer Speech and Language, **9** (1995), 171–185
9. Gauvian, J.L. and Lee, C.H.: "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains" IEEE Trans. Speech and Audio Processing, **2** (1994) 291–298
10. ITU-T Recommendations: "ITU-T G.723.1 Annex A – Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s," International Telecommunication Union (1996)
11. ITU-T Recommendations: "ITU-T H.323 – Packet-based multimedia communications systems," International Telecommunication Union (2003)
12. Tokuda, K., Yoshimura, T., Masuko,T., Kobayashi, T. and Kitamura, T.: "Speech parameter generation algorithms for HMM-based speech synthesis," IEEE Int. Conf. Acoustics, Speech, and Signal Processing (2000) 1315–1318
13. "Speech Signal Processing Toolkit"
http://kt-lab.ics.nitech.ac.jp/tokuda/SPTK/
14. "HTK Hidden Markov Model Toolkit"
http://htk.eng.cam.ac.uk/
15. Ito, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K., and Itahashi, S.: "JNAS:Japanese speech corpus for large vocabulary continuous speech recognition reseach," Journal of the Acoustical Society of Japan E, **20** (1999) 199–207
16. Lee, A., Kawahara, T., and Shikano, K.: "Julius — an open source real-time large vocabulary recognition engine," Proc. European Conf. Speech Communication and Technology (2001) 1691–1694
17. "Julius – Open-Source Large Vocabulary CSR Engine,"
http://julius.sourceforge.jp/en/julius.html
18. 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, **9** (1990) 357–363
19. Imai, H.: "Cepstral analysis synthesis on the mel frequency scale" IEEE Int. Conf. Acoustics, Speech, and Signal Processing (1983) 93–96