Rapid CODEC Adaptation for Cellular Phone Speech Recognition

Masaki Naito, Shingo Kuroiwa, Tsuneo Kato, Tohru Shimizu, and Norio Higuchi

KDDI R&D Laboratories
2–1–15 Ohara Kamifukuoka, 356–8502 Saitama, JAPAN
naito@kddlabs.co.jp

Abstract

Along with the ever increasing popularity of cellular phones, improving recognition accuracy in cellular phone speech has become an issue of growing concern. However, the distortion caused by current low-bit rate speech CODEC is nonlinear, so compensating for distortion by applying only a conventional CMN which assumes distortion is a stationary linear transfer on the cepstrum domain is difficult. In this paper, to improve speech recognition accuracy over cellular phone networks, we investigate the use of CODEC-dependent acoustic models and rapid CODEC adaptation using model selection based on maximum likelihood criterion. By using these methods we succeeded in reducing recognition errors in cellular phone speech by 33 %.

1. Introduction

Recognizing telephone speech is more difficult than recognizing high quality microphone speech because of channel distortion that occurs due to bandwidth limitation, impulse noise, frequency translation and so on. Cepstrum mean normalization(CMN) is known as an effective technique for removing convolutional linear channel distortion and is used as a standard technique in telephone speech recognition systems [1].

The number of cellular phone subscribers has increased rapidly in recent years and likely to exceed the number of fixed-line telephone subscribers in Japan. Along with the increasing popularity of cellular phones, the recognition accuracy of cellular phone speech has become an issue of growing concern. Cellular phone services however are offered by various carriers and these carriers employ various cellular phone systems that utilize different speech CODECs. The distortion caused by current low-bit rate speech CODECs is nonlinear and the characteristics of each CODEC are different. Therefore, removing distortion by applying only a conventional CMN which assumes distortion to be stationary linear transfer on the cepstrum domain is difficult.

On the other hand, by applying adaptation techniques which were originally developed for speaker adaptation, acoustic model for cellular phone speech can be trained from clean speech acoustic model with only a small amount of speech data corrected on the target cellular phone network[2]. However, In practical telephone speech recognition applications, models must be adapted rapidly, e.g. one utterance, because some service do require users to utter a few sentences during a call. The number of available target models is however limited such as to CODEC types on cellular phones. we therefore prepared these adapted models priori to performing recognition.

In this paper, we propose a new technique which selects the most suitable acoustic model for each utterance from a limited number of previously trained CODEC-dependent HMM(COD-HMM) for rapid adaptation. We first compared the performance of COD-HMM trained by conventional adaptation methods and examined their performance. We then compared the likelihood of matched COD-HMM and mismatched COD-HMM and evaluated effectiveness of a cellular-phone adaptation method which selects the most suitable COD-HMM based on the maximum likelihood criterion.

2. Cellular phone database

We collected the cellular phone speech database over an actual major cellular phone network in 1999. The speech CODEC used in this database is described in Table 1. All data were collected using an ISDN line. Note that the CODEC of the fixed-line(N) and PHS(P) have the same μ-law[3], and that the CODEC of the digital cellular(D) is selected from PSI-CELP and VSELP depending on the network load. The amount of speech data which used to train the COD-HMM is described in Table 1. Each speaker utters phonetically balanced sentences from a relatively quiet location such as the home or office.

3. Training and decoding of the CODEC-dependent HMM

3.1. Training method

We applied MAP and MLLR as the adaptation techniques, these were originally developed for speaker adaptation, to train CODEC-dependent HMM. As the baseline model, we trained 333 states shared-

Table 1: *Cellular phone speech database.*

			Number of speakers(utterances)	
Set	Type	CODEC	Male	Female
C	CDMA cellular[4]	EVRC(1.2–9.6kbps)	111(1553)	100(1786)
D	digital cellular[5]	PSI-CELP(5.6kbps)/VSELP(11.2kbps)	38(2198)	54(3663)
P	PHS[6]	μ-law(64kbps)	10(465)	20(923)
N	fixed-line	μ-law(64kbps)	1438(67609)	976(52018)

state context-dependent HMM using fixed-line telephone speech(N). We then adapted the baseline model to each CODEC by employing the following two types of adaptations.

1. adapt by MLLR. Every state treated as one regression class.(**MLLR**).
2. adapt by MAP after (1) (**MLLR+MAP**)

The digital cellular(D) includes two CODEC and we could not separate speech recorded through digital cellular(D) into PSI-CELP encoded speech and VSELP encoded speech. We therefore adapted a model for digital cellular (D) with mixed speech data encoded by both CODEC and prepared an identical model. The conditions for acoustic analysis are described in Table 2.

Table 2: *Acoustic analysis conditions.*

Sampling frequency	8 kHz
Frame shift	10 ms
Frame length	25 ms
Window type	Hamming
Feature parameters	MFCC 1–12 Δ MFCC 1–12, Δ logpow $\Delta\Delta$ MFCC 1–12, $\Delta\Delta$ logpow

3.2. Maximum likelihood criterion based COD-HMM selection

We performed the selection of the most suitable acoustic model using one utterance. In this method, selection of COD-HMM and speech recognition are performed simultaneously. These processes are performed by finding a phoneme sequence $\hat{\mathbf{p}}$ and COD-HMM $\hat{\mathbf{c}}$ for the given acoustic sequence $\mathbf{y}$ such that

$$P(\hat{\mathbf{p}}, \hat{\mathbf{c}}|\mathbf{y}) = \max_{\mathbf{p},\mathbf{c}} P(\mathbf{p}, \mathbf{c}|\mathbf{y}) \qquad (1)$$

from every possible phoneme sequence $\mathbf{p}$ and COD-HMM $\mathbf{c}$.

Under the assumption that events $\mathbf{p}$ and $\mathbf{c}$ are independent of each other, we modify the above equation as follows.

$$P(\hat{\mathbf{p}})P(\hat{\mathbf{c}})P(\mathbf{y}|\hat{\mathbf{p}}, \hat{\mathbf{c}}) = \max_{\mathbf{p},\mathbf{c}} P(\mathbf{p})P(\mathbf{c})P(\mathbf{y}|\mathbf{p}, \mathbf{c}) \quad (2)$$

where $P(\mathbf{p})$ is the *a priori* probability of the phoneme sequence $\mathbf{p}$, $P(\mathbf{c})$ is the *a priori* probability for use of COD-HMM $\mathbf{c}$, and $P(\mathbf{y}|\mathbf{p}, \mathbf{c})$ is the conditional probability of the acoustic sequence $\mathbf{y}$ given the phoneme sequence $\mathbf{p}$ and COD-HMM $\mathbf{c}$. In the following experiments, we equalize $P(\mathbf{c})$ for all COD-HMMs.

To reduce the computational cost we implement our speech recognition with model selection as follows[7]. First, we started with m different set of hypotheses corresponding to m COD-HMMs and started a frame-synchronous Viterbi search. For each frame, these hypotheses grow separately for each COD-HMM, and the hypotheses with a lower likelihood, among hypotheses for all COD-HMMs, are pruned by using the beam search technique. At the end of the input speech, the most suitable COD-HMM and the recognized result are obtained by selecting the hypotheses with the maximum likelihoods from among the hypotheses for all COD-HMMs. By using this method, those hypotheses for COD-HMMs that do not fit the target CODEC are pruned early during the recognition process. Therefore, the computational cost is reduced compared to running the recognition separately for all COD-HMMs and selecting the maximum likelihood result from the results separately recognized with each COD-HMM.

4. Experiments

We conducted a 3000 word vocabulary, which consists of the name of the company, word recognition experiments to evaluate obtained COD-HMM and the rapid CODEC adaptation method. Through CODECs described in Table 1, 15 male and 15 female speakers each utter 10 words

Table 3 shows the word error rate obtained by using baseline HMM for fixed-line(**N**), COD-HMMs trained by MLLR(**MLLR**) and COD-HMMs trained by MLLR and MAP (**MLLR+MAP**). The results show that COD-HMM adapted by MLLR reduce error in baseline HMM for fixed-line by 14-16%. By using MLLR and MAP, further improvements are obtained and result in a 33-37% reduction in errors in baseline HMM. These results clearly show that stationary linear transformation is not sufficient to reduce channel distortion caused by CODEC on cellular phone networks. Therefore, we used COD-HMM

Table 3: *Comparison of adaptation methods (word error rate %).*

Target set	N	MLLR	MLLR+MAP
N	2.7	-	-
C	7.0	5.9	4.4
D	7.9	6.8	5.3

Table 4: *Word accuracy(%) obtained by COD-HMMs.*

Target	COD-HMM			
set	N	P	C	D
N	2.7	2.4	3.0	7.9
P	2.5	2.2	4.1	9.8
C	7.0	10.3	4.4	9.9
D	7.9	9.8	9.9	5.3

adapted by MLLR and MAP in the following experiments.

Table 4 shows the word error rate obtained by recognizing test data of each CODEC by each COD-HMM. They show that best performance can be obtained by using an COD-HMM for each CODEC. In the case of fixed-line(N) and PHS(P), which use the same CODEC μ-law, the difference between accuracy obtained by using COD-HMMs for each set is small. But the other mismatch in CODEC of input speech and COD-HMM causes a large degradation in recognition accuracy.

Next, prior to attempting COD-HMM selection based on maximum likelihood criterion, we compare the likelihood of COD-HMM for each CODEC. For each utterance, the likelihood of the speech period normalized by the frame length of the period is calculated by using each COD-HMM. We then calculate the difference between likelihood of COD-HMMs for the target CODEC and the other CODEC.

Figure 1 shows histogram of the difference in likelihood of COD-HMM for a CDMA cellular(C) and another COD-HMM. This result shows the difference in likelihood of the appropriate COD-HMM and another COD-HMM are large enough to select models based on maximum likelihood criterion.

Table 5 shows statistics of differences in likelihood of COD-HMM for target CODEC and another CODEC during the speech period. The average difference in likelihood and standard derivation of difference is described in each cell of the table. For most combinations of speech data and COD-HMM, the difference in likelihood is large enough. But, fixed-line(N) and PHS(P) use same CODEC μ-law so that the difference in likelihood obtained by using each COD-HMM is small.

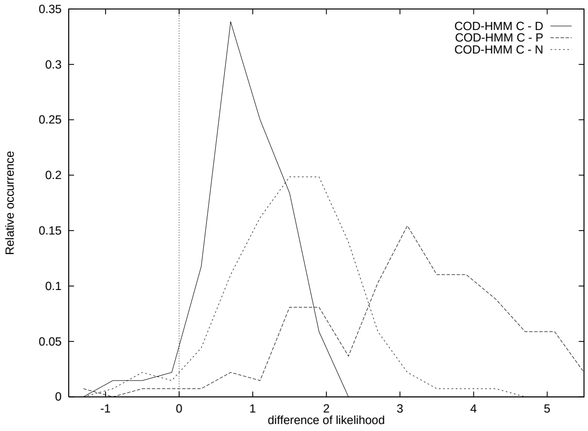

Figure 1: *Histogram of the difference in likelihood of COD-HMM for CDMA cellular(C) and another COD-HMM.*

Same statistics on the silence period described in Table 6. Differ from the results of speech period, these results show that the difference in likelihood between appropriate COD-HMM and other COD-HMM is small and the standard derivation is large in the silence period. Therefore, it is not advisable to selecting COD-HMM during the silence period.

Based on the above investigation, we conducted speech recognition experiments with COD-HMM selection based on maximum likelihood criterion. The characteristics of COD-HMM for fixed-line(N) and PHS(P) are similar. We therefore use three COD-HMM in parallel for fixed-line(N), CDMA cellular(C) and digital cellular(D). Furthermore, to avoid model selection failure during a silence period, we prepared a CODEC-independent silence model and used jointly it with each CODEC-dependent speech model. The CODEC-independent silence model is generated by composing a CODEC-dependent silence model based on the following equation.

$$b(\mathbf{y}) = \frac{1}{M}\sum_{m=1}^{M} b_m(\mathbf{y}) \qquad (3)$$

where M is the number of COD-HMMs and $b_m(\mathbf{y})$ is output probability of each COD-HMM.

The other approach to increase robustness to counter the distortion caused by the difference in CODECs is to train a CODEC-independent HMM. To compare the effectiveness of CODEC-independent modeling with the model selection approach, we generated a CODEC-independent HMM by composing each state of a COD-HMM based on equation 3.

The word error rate obtained by HMM for a fixed-line (**Fixed-line**), a manually selected model (**Manual**), a composed CODEC-independent HMM (**Compose**) and a likelihood based selected model (**Likelihood**) is described in Table 7. This result

Table 5: *Difference in likelihood between a matched COD-HMM and another COD-HMM during a speech period ([] standard deviation).*

Target	COD-HMM			
set	N	P	C	D
N	-	1.25(0.76)	1.68(0.72)	1.60(0.69)
P	0.35(0.62)	-	2.22(0.88)	1.48(0.75)
C	1.55(0.84)	3.22(1.41)	-	0.89(0.54)
D	1.63(0.77)	2.77(1.39)	0.69(0.57)	-

Table 6: *Difference in likelihood between a matched COD-HMM and another COD-HMM during a silence period ([] standard deviation).*

Target	COD-HMM			
set	N	P	C	D
N	-	0.51(1.49)	1.60(2.53)	1.56(2.05)
P	0.30(1.03)	-	2.24(2.33)	1.54(1.56)
C	1.22(2.74)	2.24(3.54)	-	0.97(2.28)
D	1.10(2.00)	2.10(2.53)	0.07(1.61)	-

Table 7: *Word error rate(%) obtained by using CODEC adaptation method.*

	Target set			
Method	N	P	C	D
Fixed-line(N)	2.7	2.5	7.0	9.0
Manual	2.7	2.2	4.4	5.3
Compose	3.0	3.4	6.2	5.3
Likelihood	2.7	2.2	4.4	4.5

shows that the accuracy obtained by COD-HMM selection is equivalent to the manually selected model, and that the improvements are always superior to the CODEC-independent HMM.

5. Conclusions

To improve the accuracy of speech recognition over cellular-phone networks, we investigate the use of CODEC-dependent acoustic models and rapid CODEC-adaptation using model selection based on maximum likelihood criterion. These methods succeeded in reducing recognition errors in cellular phone speech by 33 %. An issue for future study is new CODECs such as G.729(CS-ACELP), which are recently installed on cellular phone systems to improve the quality of cellular phone speech[8]. We therefore need to also evaluate our method with these new CODECs. Cellular phone subscribers also have a tendency to call from noisy locations such as in crowds, vehicles and train platforms so enhancing the robustness of speech recognition versus environmental influences is also important.

6. References

[1] A.Acero, “Acoustical and environmental robustness in automatic speech recognition,” Kluwer Academic Publishers, Boston, 1993.

[2] R. Gruhn, H. Singer, H. Tsukada, A. Nakamura, M. Naito, A. Nishino, Y. Sagisaka, S. Nakamura., “Cellular Phone Based Speech-To-Speech Translation System ATR-MATRIX,” Proc. of ICSLP 2000, Vol. IV, pp. 448–451, 2000.

[3] ITU-T, Recommendation, G.711: Pulse code modulation (PCM) of voice frequencies.”

[4] ARIB STD-T53, “CDMA Cellular System ARIB STANDARD,” Association of Radio Industries and Buisinesses.

[5] RCR STD-27, “PERSONAL DIGITAL CELLULAR TELECOMMUNICATION SYSTEM RCR STANDARD,” Association of Radio Industries and Buisinesses.

[6] RCR STD-28, “PERSONAL HANDY PHONE SYSTEM RCR STANDARD,” Association of Radio Industries and Buisinesses.

[7] K. Yamaguchi, H. Singer, S Matsunaga, S. Sagayama., “Speaker-Consistent Parsing for Speaker-Independent Continuous Speech Recognition,” IEICE trans. INF.&SYST, Vol. **E78-D**, No. 6, pp. 470–475, 1995.

[8] ITU-T Recommendations, “G729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP).”