SYLLABLE LATTICE BASED RE-SCORING FOR SPEAKER VERIFICATION

Minho Jin*+, Frank K. Soong+, Chang D. Yoo*,

*Div. of EE, Dept. of EECS, Korea Advanced Institute of Science and Technology
373-1 Guseong Dong, Yuseong Gu, Daejeon 305-701, Korea
jinmho@kaist.ac.kr, cdyoo@ee.kaist.ac.kr
+Microsoft Research Asia
5F, Sigma Center, No. 49, Zhichun Road, Beijing 100080, P.R.C
frankkps@microsoft.com

ABSTRACT

The Gaussian mixture based GMM-UBM approaches have shown good performance in speaker verification without using contextual and longer-term temporal information in speech signal. This paper proposes to use the information provided in the arcs of a decoded syllable lattice to verify a claimed speaker's identity. The forward algorithm is used to summarize this information in the whole syllable lattice instead of the best decoded string. The performance is evaluated on the mandarin Chinese database. With two minute of target speaker's enrollment data, the proposed algorithm shows 1.03% of equal-error rate for short input utterances. By combining with the GMM-UBM, the system shows a 0.74% of equal-error rate.

1. INTRODUCTION

The speaker verification is to verify the claimed speaker's identity via speech. It can be classified into text-dependent (TD) and text-independent (TI). The TD speaker verification takes the input speech and its true transcription while the TI speaker verification takes only the input speech. This paper focuses on the TI speaker verification.

The GMM-UBM has shown good performance in TI speaker verification[1][2], where each frame is assumed to be independent. On the other hand, Doddington[3] showed that the contextual information have potential to verify speaker identity. Many approaches have proposed to utilize this information. Among them, the phonetic speaker verification[4] based on the relative phone n-gram achieved good performance.

In TD speaker verification where the exact contextual information is given, the hidden Markov model (HMM) based approaches are very successful[5]. In [6][7], the large vocabulary continuous speech recognition (LVCSR) was used to reduce the performance gap between TD and TI speaker verification. These approaches assume the best transcription to be correct and perform the speaker verification. Thus, an erroneous decoding can degrade the verification performance.

The oracle word-error rate (WER) is usually much better than the 1-best WER. Therefore using whole lattice instead of the best transcription gives much more chance for the correct path to appear in the test statistics. Recently, the phone lattice decoding achieved good improvements in both language identification[8] and speaker verification[9]. In this paper, the syllable lattice is used because our experiments were performed on mandarin Chinese, a syllabically paced language. The performance improvement was experimentally shown over the previous best transcription-based approaches.

This paper is organized as follows: Section 2 illustrate the speaker verification with HMM and calculation of lattice-based score. Section 3 shows the experimental results, and Section 4 summarizes and discuss further works.

2. LATTICE RE-SCORING FOR SPEAKER VERIFICATION

2.1. Speaker Modeling with hidden Markov Models

The speaker verification is usually performed with a background model and the claimed speaker's model. The background model is trained as general speaker-independent (SI) model, and the target speaker model is estimated on target speaker's enrollment data. Fig. 1 illustrates the system design.

The target speaker's model is usually adapted from the SI background model using the target speaker's enrollment data. If the transcription is not provided, unsupervised adaptation is performed. In our experiments, the transcriptions of enrollment data are obtained by SI-HMMs with free syllable decoding loop, and the maximum likelihood linear regression (MLLR) [10] and the maximum a posterior (MAP) [11] adaptations are performed successively.

*This work was carried out during the first author's internship at Microsoft Research Asia.

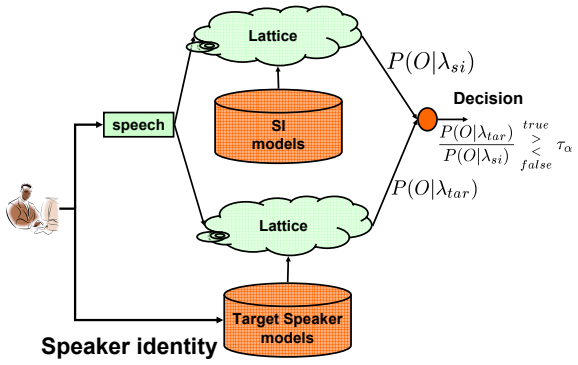

Fig. 1. *System Design*

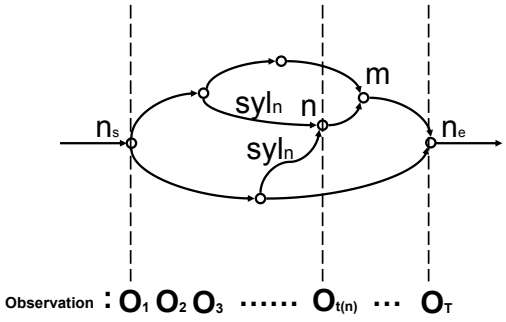

Fig. 2. *Forward Algorithm for Lattice*

2.2. Lattice for Speaker Verification

The Neyman-Pearson criterion leads the test statistics for speaker verification as follows:

$$\frac{P(O|\lambda_{tar})}{P(O|\lambda_{si})} \underset{false}{\overset{true}{\gtrless}} \tau_\alpha \quad (1)$$

where O is the input utterance, $P(O|\lambda_{tar})$ and $P(O|\lambda_{si})$ are the likelihoods from target speaker's model and background model, and τ_α is threshold chosen beforehand. In previous LVCSR based speaker verification[6][7] the 1-best hypotheses are used compute

$$P(O|\lambda_{si})_{1-best} = P(O, Q^*|\lambda_{SI}) \quad (2)$$
$$P(O|\lambda_{tar})_{1-best} = P(O, Q^*|\lambda_{tar}) \quad (3)$$

where Q^* is the state alignment of 1-best hypothesis obtained by SI models λ_{SI}. Since the test statistics only depend on the best transcription, an erroneous decoding can degrade the speaker verification performance. In this paper, $P(O|\lambda)$ is approximated by the lattice re-scoring.

2.2.1. Grammar Network

The phone lattice based decoding showed good performance in both language identification[8] and speaker verification[9]. In this paper, we used the syllable lattice instead of a phone lattice in mandarin Chinese, a syllabically paced language. The number of syllables is limited (i.e., slightly over 400 without tones), and the accuracy of free syllable decoding is fairly decent in mandarin Chinese. We used the phone set design based on the segmental tonal modeling[12], and the free syllable network used for lattice construction.

2.2.2. Lattice Forward Algorithm

In our experiments, syllable lattice as shown in Fig. 2 was used. Each arc is an instance of syllable whose label is assigned by its entering node. For example, all arcs entering the same node n are instances of same syllable syl_n starting from different nodes. The forward probability α_n of node n located on time $t(n)$ is defined as follows:

$$\alpha_n = P(O_1 O_2 \ldots\ldots\ldots O_{t(n)}, n_{t(n)} = n|\lambda) \quad (4)$$

where $n_{t(n)}$ is the hypothesized node at time $t(n)$, λ is the model used to construct the lattice. n_s and n_e are the starting node and the ending node, respectively. For node m, the forward probability of the node m with all the preceding nodes, ns, is as follows::

$$\alpha_m = \sum_{\substack{all\ n\ s.\ t. \\ n->m}} \alpha_n P(m|n) \quad (5)$$

With (5), the likelihood of whole input observations O can be calculated as follows:

$$P(O|\lambda) \approx P(O_1 O_2 \ldots\ldots\ldots O_T, n_T = n_e|\lambda) = \alpha_{n_e} \quad (6)$$

where n_e is the ending node of the network. Using (6), $P(O|\lambda_{si})$ and $P(O|\lambda_{tar})$ can be similarly approximated, and verification decision is made, based upon (1).

3. EXPERIMENTS

3.1. Experimental Setup

The speaker verification is evaluated using the mandarin Chinese database. This database contains clean speech collected by microphone with 16kHz sampling rate. 39^{th} dimension Mel-frequency cepstral coefficients (MFCC), consisting of 12 cepstral coefficients plus energy and their 1^{st} and 2^{nd} derivatives were used.

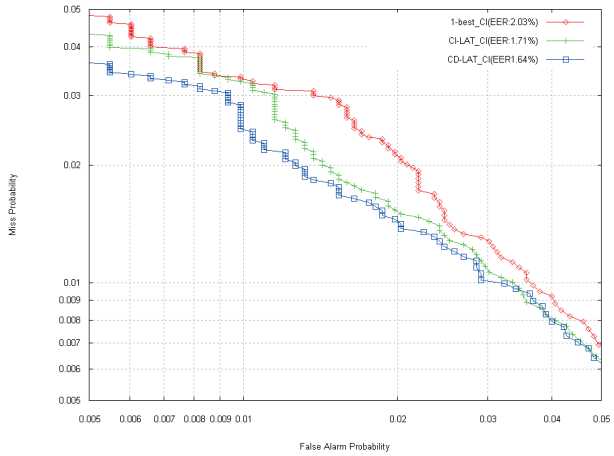

Fig. 3. *1-best Scoring(1-best_CI) and Lattice Re-scorings with CI lattice(CI-LAT_CI) and CD lattice(CD-LAT_CI).*

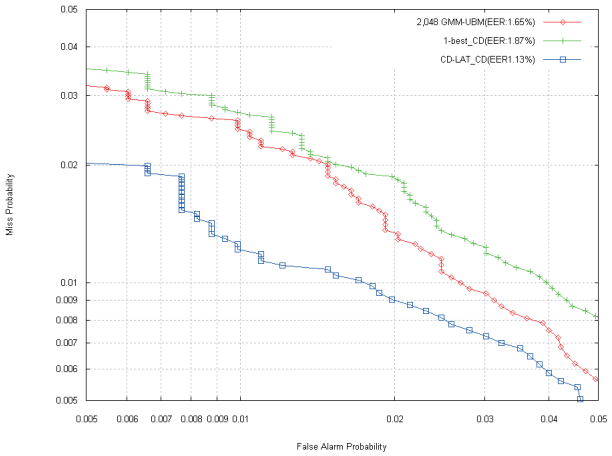

Fig. 4. *2,048 GMM-UBM, 1-best Scoring(1-best_CD), and Lattice Re-scoring(CD-LAT_CD)*

The phoneme models based on the 97 phone set [12] is used to represent the whole syllable set in mandarin Chinese. Each phone is modeled as a three-state HMMs with 16 Gaussians per state. The SI models were trained on roughly 90 hours of data collected from 300 speakers (156 female and 144 male speakers). This SI-HMMs showed 50.8% of tonal syllable recognition accuracy with a free syllable decoding. 19 (10 male and 9 female) speakers are selected as true speakers, and 180 speakers are selected as impostor speakers for evaluation. These 300, 19 and 180 speakers form mutually exclusive sets. Test utterances are segmented into 1,832 true trials and 26,250 impostor trials with an average duration of 2.0 sec. The target speaker model is adapted from the SI model by unsupervised MLLR-MAP with 2 min. of enrollment data as described in Sec.2.1. The proposed algorithm is evaluated with context-dependent (CD) and context-independent (CI) models. The syllable lattices were generated by using HTK[13].

3.2. Performance with Context Independent Model

Fig. 3 illustrates three detection error trade-off (DET) curves. 1-best_CI calculate the score with CI models based on 1-best transcription. CI-LAT_CI calculate the score with CI models based on the syllable lattice generated by CI models, while CD-LAT_CI calculate the score with CI models based on the lattice generated by CD models. The use of syllable lattice from both CI and CD models improve the performance over the 1-best transcriptions. The CD models generated more accurate lattice which improves the speaker recognition performance.

3.3. Performance with Context Dependent Model

DET curves of GMM-UBM, 1-best scoring with CD model and our proposed approach with CD model are illustrated in Fig. 4. In GMM-UBM, the background model of 2,048 Gaussian kernels is trained on the same data used for SI-HMMs, and target speaker's model is adapted by MAP adaptation. The performance of CD model is better than that of CI model for both 1-best scoring and lattice re-scoring. The performance of 1-best scoring is slightly worse than that of GMM-UBM, which was similarly observed in [6] for short utterances. The performance is significantly improved with syllable lattice re-scoring compared with 1-best scoring. The proposed algorithm showed error-rate reduction of 31.5% and 39.6% compared to GMM-UBM and 1-best scoring.

3.4. MLLT for Accurate Modeling and Combination

The maximum likelihood linear transformation (MLLT)[14] is performed to increase the model accuracy. The MLLT was used for GMM-UBM and showed significant improvement[15]. In our proposed algorithm, the MLLT improves the tonal syllable recognition accuracy to 54.5%, which amounts to 7.3% of error-rate reduction compared to MFCC. Additionally, linear combination of the two scores is calculated as follows:

$$\begin{aligned} TS_{comb}(O) &= a * TS_{GMM-UBM}(O) \\ &+ (1-a) * TS_{LAT-SCR}(O) \end{aligned} \quad (7)$$

where $TS_{GMM-UBM}(O)$ is the test statistics of observation O with GMM-UBM algorithm and $TS_{LAT-SCR}(O)$ is the test statistics with the proposed algorithm.

Compared with Fig. 4, the performance was improved for both the proposed algorithm and the GMM-UBM. In Fig. 5,

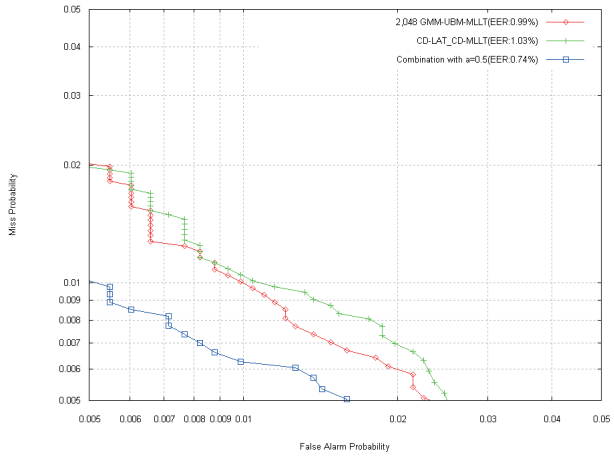

Fig. 5. *2,048 GMM-UBM with MLLT, Lattice Re-scoring with MLLT(CD-LAT_CD-MLLT), and Combination with* $a = 0.5$

the GMM-UBM with MLLT showed slightly better performance than the lattice re-scoring with MLLT. However, the difference is not statistically significant, and the linear combination with $a = 0.5$ showed the best EER of 0.74%.

4. CONCLUSION

The syllable lattice re-scoring was investigated for speaker verification and compared to the 1-best scoring and the GMM-UBM. The proposed algorithm showed that the whole decoded syllable lattice shows distinctive potential for verifying a speaker's identity. And the use of the temporal information by ASR can result in comparable performance to GMM-UBM in speaker verification with short utterances. The performance were improved by MLLT in both the proposed algorithm and the GMM-UBM, and their linear combination gives the best performance. For phoneme model, we performed MLLR-MAP adaptation in an unsupervised manner. In our future work, lattice-based adaptation will be used to exploit more information from the decoded lattice information.information

5. ACKNOWLEDGMENTS

This work was supported by grant No. R01-2003-000-10829-0 from the Basic Research Program of the Korea Science and Engineering Foundation and by University IT Research Center Project.

6. REFERENCES

[1] T. Quatieri D. Reynolds and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[2] Ganesh Ramaswamy and et al, "The ibm system for the nist-2002 cellular speaker verification evaluation," in *Proc. ICASSP*, 2003, vol. 2, pp. 61–64.

[3] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. Eurospeech*, 2001, vol. 4, pp. 2517–2520.

[4] M. A. Kohler W. D. Andres and J. P. Campbell, "Phonetic speaker recognition," in *Proc. Eurospeech*, 2001, pp. 149–153.

[5] J.P. Campbell Jr., "Speaker recognition: a tutorial," *Proc. of the IEEE*, vol. 85, pp. 1437–1462, Sept. 1997.

[6] F. Weber, B. Peskin, M. Newman, A. Corrada Emmanuel, and L. Gillick, "Speaker recognition on single- and multispeaker data," *Digital Signal Processing*, vol. 10, pp. 75–92, 2000.

[7] Tilmothy Hazen and et. al., "Integration of speaker recognition into conversational spoken dialogue systems," in *Proc. Eurospeech*, 2003, pp. 1961–1964.

[8] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language Recognition Using Phone Lattices," in *Proc. ICSLP*, 2004.

[9] Andrew Hatch, Barbara Peskin, and Andreas Stolcke, "Improved Phonetic Speaker Recognition Using Lattice Decoding," in *Proc. ICASSP*, 2005, vol. 1, pp. 169–172.

[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[11] Chin-Hui Lee, Chih-Heng Lin, and Biing-Hwang Juang, "A study on speaker adaptation of the parameters of continuous density hidden markov models," *IEEE Trans. Signal Proc.*, vol. 39, no. 4, pp. 806–814, 1991.

[12] Chao Huang and et al, "Segmental tonal modeling for phone set design in mandarin lvcsr," in *Proc. ICASSP*, 2004, pp. 901–904.

[13] Steve Young and et. al., "The htk book," 2002.

[14] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. ICASSP*, 1998, pp. 661–664.

[15] Jiri Navratil, Upendra V. Chaudhari, and Ganesh N. Ramaswamy, "Speaker verification using target and background dependent linear transforms and multi-system fusion," in *Proc. Eurospeech*, 2001.