T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	N-Best-Based Instantaneous Speaker Adaptation Method for Speech Recognition
Author	Tomoko Matsui, Sadaoki Furui
Journal/Book name	ICSLP96, Vol. 2, No. , pp. 973-976
発行日 / Issue date	1996, 10

Powered by T2R2 (Science Tokyo Research Repository)

Proc. Int. Conf. Spoken Language Processing, pp. 973-976 (1996)

N-BEST-BASED INSTANTANEOUS SPEAKER ADAPTATION METHOD FOR SPEECH RECOGNITION

Tomoko Matsui *Sadaoki Furui*

NTT Human Interface Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo, Japan

ABSTRACT

An instantaneous speaker adaptation method is proposed that uses N-best decoding for continuous mixture-density hidden-Markov-model based speech recognition systems. An N-best paradigm of multiple-pass search strategies is used that makes this method effective even for speakers whose decodings using speaker-independent models are error-prone. To cope with an insufficient amount of data, our method uses constrained maximum a posteriori estimation, in which the parameter vector space is clustered, and a mixture-mean bias is estimated for each cluster. Moreover, to maintain continuity between clusters, a bias for each mixture-mean is calculated as the weighted sum of the estimated biases. Performance evaluation using connected-digit (four-digit strings) recognition experiments performed over actual telephone lines showed more than a 20% reduction in the error rates, even for speakers whose decodings using speaker-independent models were error-prone.

1. INTRODUCTION

In continuous mixture-density hidden Markov model (HMM)-based speech recognition systems, the performance of speaker-independent phoneme HMMs for some speakers is often low. Techniques that adapt the parameters of speaker-independent phoneme HMMs to each speaker and thus improve the performance are therefore important. These techniques are generally classified as either supervised, in which the training sentences are known, or unsupervised, in which arbitrary utterances are used. They can also be classified as off-line, in which the system collects a limited amount of data and uses it for adaptation, incremental, in which the adaptation transformation is estimated every utterance without supervision and the adapted models are used for the next utterance, and instantaneous, in which recognition utterances are used to estimate the adaptation transformation [1]. Instantaneous adaptation is especially useful in applications where there is only a very brief interaction between the speaker and the system. This technique must work without supervision, using only a small amount of data, such as a few words or a single sentence.

In general, unsupervised adaptation techniques use a decoding algorithm, such as the Viterbi algorithm, to align the input speech frames to the phonemes by using speaker-independent phoneme HMMs [1][2]. The mixture-density distributions are then adapted using the alignment. However, the decoding is error-prone for some speakers, so the adaptation effect does not usually work.

In this paper, we propose an instantaneous speaker-adaptation method that is effective even for error-prone speakers. It uses an alignment obtained based on the maximum likelihood value using speaker-adapted phoneme models instead of speaker-independent phoneme ones. We assume that the correct decoding shows a high likelihood value through adaptation even when it shows a low value for speaker-independent phoneme HMMs. In this method, speaker adaptation using all possible alignments must be attempted, and the likelihood values for the phoneme models adapted using these alignments must be compared. Therefore, how to reduce the search space without losing the correct decoding is a serious problem.

In large-vocabulary speech recognition, multiple-pass search strategies have been explored as a way to substantially reduce the search space, without increasing the error rate [3]. One of these strategies, the N-best paradigm, computes alternative hypotheses for a sentence that can later be rescored using more detailed and more comprehensive knowledge sources. We use the N-best paradigm in our speaker adaptation method to calculate likely alignments. In our method, phoneme models that become more precisely adapted to the speaker as the number of estimation iterations increases are used.

Estimating all the parameters of phoneme HMMs robustly is difficult when only a small amount of data is available. Maximum a posteriori (MAP) estimation, which combines estimates obtained from the adaptation data with a priori information of the speaker-independent system, is particularly useful in dealing with problems posed by sparse training data for which the maximum-likelihood (ML) approach gives inaccurate estimates [4]. This estimation, however, updates only the parameters of phoneme models for which observations occur in the adaptation data. Several techniques for updating the parameters of phoneme models not observed in the adaptation data have been developed. In [5] and [6], the estimated spectral bias is used to decrease the uniform mis-

$$b_n^i = \frac{\sum_{\{j,k|\phi(j,k)=i\}} [\tau_{jk}(\mu_{jk} - m_{jk}) + \sum_{t=1}^{T} c_{jkt}(x_t - m_{jk})]r_{jk}}{\sum_{\{j,k|\phi(j,k)=i\}} (\tau_{jk} + \sum_{t=1}^{T} c_{jkt})r_{jk}} \quad (1)$$

match between input speech and all phoneme models, and this estimated bias is subtracted from each speech frame or added to each mixture mean. In [2] and [7], the adaptation transformation is constrained to be an Affine transformation, and the parameters of the transformation are estimated separately for different clusters of mixture-density distributions. To cope with an insufficient amount of data, our method calculates each mixture-mean bias as the weighted sum of the biases which are estimated for each mixture-mean cluster based on MAP estimation.

2. N-BEST-BASED ADAPTATION

Our method consists of three steps (Figure 1):

1. Multiple alignments of input speech $\{\Lambda_1, \Lambda_2, \ldots, \Lambda_N\}$ are obtained using speaker-independent phoneme HMMs and the N-best decoding technique.
2. The parameters of the phoneme models are adapted for each decoding.
3. The decoding providing the maximum-likelihood value is selected for the speech, and the speaker-adapted phoneme models for that decoding are used.

According to the N-best paradigm of multiple-pass search strategies, Steps 2 and 3 are iterated as the adaptation in Step 2 gradually becomes more precise. Then, the hierarchical codebook adaptation algorithm [8][9], which has been proposed for speaker adaptation in vector-quantization based systems, is applied: the reference codebook elements are clustered hierarchically in an increasing number of clusters, and adaptation is performed hierarchically from global to local individuality of the speaker. In practice, the mixture-mean biases shared by the distributions in the same cluster are estimated, and the number of biases (i.e., the number of clusters) increases as the number of estimation iterations increases. The following sections explain how we estimate the mixture-mean biases and cluster the mixture means.

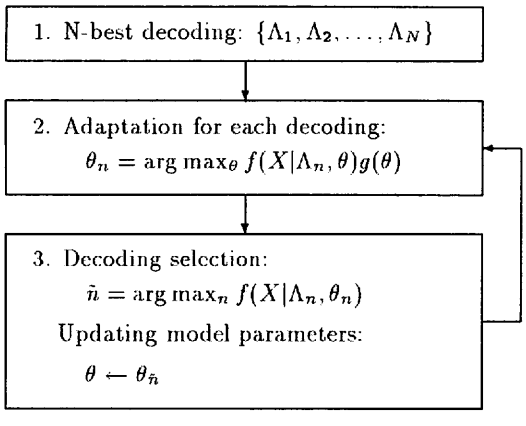

Figure 1: Block diagram of our method

2.1. Bias estimation

For each alignment Λ_n, the mixture-mean bias b_n^i of cluster i is approximated while $f(X|\Lambda_n, \theta_b)g(\theta_b)$ is locally maximized using MAP estimation, where X is the observation sequence, θ_b is the HMM parameter set prescribed by bias b in the relation $m_{jk} \rightarrow m_{jk} + b$ (m_{jk}: mean vector of mixture component k in state j), $f(\cdot)$ is the likelihood function, and $g(\cdot)$ is the a priori density function. The expectation-maximization reestimation formula is derived in Eq. (1), where μ_{jk} and τ_{jk} are the a priori density parameters, r_{jk} is the precision vector, and c_{jkt} is the probability of being in state j with mixture component k at time t given that the HMM generates observation vector x_t; $\phi(\cdot)$ is a membership function indicating the cluster to which the mixture component in the state belongs. Then, to maintain continuity between the clusters, for each mixture mean of all phoneme HMMs, the bias was calculated as the weighted sum of biases $\{b_n^1, b_n^2, \ldots, b_n^I\}$ based on the distances between the centroids and the mixture-mean vector [8][9].

2.2. Hierarchical clustering

In our method, a binary-tree-structure is made from input speech. The number of leaves depends on the speech length and the depth of the tree depends on the number of estimation iterations. The mixture-density distributions are clustered into 2^{M-1} classes based on the distances between the centroids and the mean vectors of the distributions, where M is the number of estimation iterations.

3. EXPERIMENTAL CONDITIONS

A speaker-independent HMM for each digit was created by the Baum-Welch algorithm using digit string utterances (Japanese one, two, or four-digit strings with a short silence at the beginning and end) spoken by 177 male speakers (24,194 total strings). The database was collected by NTT over actual telephone lines in a metropolitan area and recently used for the extended work of connected-digit recognition in our laboratory [10]. In our experiments, four-mixture Gaussian HMMs were used as context-independent digit-HMMs. The number of states in each digit-HMM varied depending on the number of phonemes in the digit; each phoneme is represented by three states and the average number of phonemes per digit is three, so there are nine states per digit on average. We used 13 digit-HMMs to represent the digits 0 to 9, since in Japanese some digits have two pronunciations; a one-state, four-mixture Gaussian HMM was

N-best	number of strings	
1-best	507	[84.5]
2-best	41	[6.8]
3-best	10	[1.7]
4-best	2	[0.3]
5-best	4	[0.7]
6-best	3	[0.5]
7-best	6	[1.0]
8-best	3	[0.5]
9-best	1	[0.2]
10-best	1	[0.2]
Other	22	[3.7]

Table 1: Number of correct strings included within ten best decodings by using speaker-independent digit models ([ ]: percentage of strings).

used as the silence HMM.

The database used for adaptation and recognition tests consisted of four-digit strings spoken by 400 male speakers [11]. The data was collected by NTT Data over actual telephone lines in seven different areas. The string recognition rates were used to evaluate our method. In the experiments, the ten best alignments were decoded by default. The percentage of correct strings included within the ten best decodings by using speaker-independent digit models was 96.3% (Table1). One hundred of the 400 speakers were selected so that the histogram of the recognition rates for the speaker-independent digit-HMMs was the same as that for the 400 speakers. Six different strings were used per speaker. The 100 speakers were classified into two sets: one set consisted of 25 speakers whose recognition rates using speaker-independent digit-HMMs were under 80% (average: 54.7%); the other set consisted of the remaining 75 speakers whose rates were over 80% (average: 94.4%).

The cepstral and delta-cepstral coefficients were calculated with an order of 12 for each. LPC analysis was used with a frame period of 8 ms, and a frame length of 32 ms. Cepstral mean subtraction was performed for each string.

4. RESULTS

Table2 lists the four-digit string recognition rates [and the error reduction rates compared with the baseline performance]. Speaker-independent digit models were used when the one best and when the ten best alignments were decoded. "16-HBE" is our method for estimating 16 mixture-mean biases for each string by using hierarchical clustering with five estimation iterations (one bias was estimated in the first iteration, two in the second, four in the third, eight in the fourth, and 16 in the fifth). "1-BE" is our method for estimating one mixture-mean bias by using three estimation iterations without hierarchical clustering. In each iteration, the parameters of all digit-HMMs were updated using a bias given by the decoding with the maximum-likelihood value.

Method	N-best	under 80%		over 80%		Ave.
baseline	-	54.7		94.4		84.5
16-HBE	1	62.0	[16.1]	95.1	[12.5]	86.8
	10	64.7	[22.1]	95.8	[25.0]	88.0
1-BE	1	56.7	[4.4]	94.4	[0.0]	85.0
	10	59.3	[10.2]	94.9	[8.9]	86.0

Table 2: Effect of N-best number on string recognition rate (%) ([ ]: error reduction rate (%)).

Method	under 80%		over 80%		Ave.
1-BE	59.3	[10.2]	94.9	[8.9]	86.0
2-BE	64.0	[20.5]	94.7	[5.4]	87.0
4-BE	61.3	[14.6]	95.1	[12.5]	86.7
8-BE	62.7	[17.7]	95.3	[16.1]	87.2

Table 3: Effect of clustering on string recognition rate (%) for BE method ([ ]: error reduction rate (%)).

Method	Iteration	under 80%		over 80%		Ave.
8-HBE	1 (1)	60.7	[13.2]	94.9	[8.9]	86.3
	2 (2)	62.7	[17.7]	94.7	[5.4]	86.7
	3 (4)	63.3	[19.0]	94.7	[5.4]	86.8
	4 (8)	64.7	[22.1]	95.6	[21.4]	87.8
8-BE	4 (8)	63.3	[19.0]	95.3	[16.1]	87.3

Table 4: Effect of hierarchical clustering on string recognition rate (%) ([ ]: error reduction rate (%), (): number of biases).

The bias and the order of the likelihood values for the decodings are updated for each iteration. For both the 16-HBE and 1-BE methods, the recognition rates when using the ten best alignments were higher than when using only the one best alignment. These results indicate that N-best decoding works effectively in our method.

Table3 lists the four-digit string recognition rates [and the error reduction rates] when several numbers of biases were estimated using K-BE methods in which K mixture-mean biases were estimated using three estimation iterations without hierarchical clustering. In each iteration, the parameters of all digit-HMMs were updated using K biases given by the decoding with the maximum-likelihood value. The more-than-two BE methods performed better than the 1-BE method. However, the optimal number of biases seems to be difficult to automatically determine with this method.

Table4 shows the difference in performance between hierarchical and direct clustering for the 8-HBE method. In this method, eight mixture-mean biases are estimated using four estimation iterations. The recognition rates for the fourth iteration were higher than those for 8-BE. These results indicate that hierarchical clustering is effective.

Method	correct ⇓ correct	incorrect ⇓ correct	correct ⇓ incorrect	incorrect ⇓ incorrect
1-HBE	502	16 [81.3]	5	77
2-HBE	500	20 [70.0]	7	73
4-HBE	498	23 [69.6]	9	70
8-HBE	500	27 [63.0]	7	66
16-HBE	498	30 [70.0]	9	63

Table 5: Number of strings recognized correctly and incorrectly before and after use ([ ]: percentage of strings recognized as second best using speaker-independent digit models).

5. DISCUSSION

We analyzed the recognition ability of our method. Table 5 lists the number of strings recognized correctly and incorrectly before and after use of our K-HBE method. As the number of biases increased, the number of strings recognized incorrectly before use but correctly after use increased. Conversely, the number of strings recognized incorrectly both before and after use decreased as the number of biases increased. The number of strings recognized correctly before and after use and the number of strings recognized correctly before use but incorrectly after use did not vary widely.

In Table 5, the numbers in square brackets show the percentage of strings recognized as second best using speaker-independent digit models in strings recognized incorrectly before use but correctly after use. The percentage for the 1-HBE method was higher than that for the more-than-two HBE methods. Therefore, by estimating more than two biases, strings seriously misrecognized (e.g., recognized as third or fourth best) using speaker-independent digit models can be recognized correctly.

In a few cases, strings were recognized correctly before use but incorrectly after use. Further investigation is needed of this side effect.

6. CONCLUSION

We have presented an instantaneous speaker adaptation method using N-best decoding for continuous mixture-density HMM-based speech recognition systems. Connected-digit (four-digit strings) recognition experiments performed over actual telephone lines showed that this method, which can work with only a small amount of data, is effective even for speakers whose decodings using speaker-independent models are error-prone. In experiments using 100 speakers, our 16-HBE method had a 22.1% error-reduction rate for speakers whose recognition rates using speaker-independent digit HMMs were under 80%, and 25.0% for speakers whose rates were over 80%.

Further study includes investigation of more general adaptation transformations including using an Affine transformation in our method.

ACKNOWLEDGMENTS

We are grateful to the NTT Data Communications Systems Corporation for allowing us to use their database for our experiments. We are thankful to Ken Hatano of Tokyo Institute of Technology for his valuable support in our experiments. We thank the members of the Furui Research Laboratory of the NTT Human Interface Laboratories for their valuable and stimulating discussions.

REFERENCES

1. G. Zavaliagkos, R. Schwartz and J. Makhoul, *Batch, incremental and instantaneous adaptation techniques for speech recognition*, Proc. ICASSP, pp. I-676-679, 1995.
2. 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.
3. R. Schwartz, L. Nguyen and J. Makhoul *Automatic speech and speaker recognition: Chapter 18 Multiple-pass search strategies*, edited by C.-H. Lee et al., Kluwer Academic Publishers, pp. 429-456, 1995.
4. J.L. Gauvain and C.-H. Lee, *Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains*, IEEE Trans. Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, 1994.
5. A. Sankar and C.-H. Lee, *Robust speech recognition based on stochastic matching*, Proc. ICASSP, pp. I-121-124, 1995.
6. Y. Zhao, *Iterative self-learning speaker and channel adaptation under various initial conditions*, Proc. ICASSP, pp. I-712-715, 1995.
7. V.V. Digalakis, D. Rtischev and L.G. Neumeyer, *Speaker adaptation using constrained estimation of Gaussian mixtures*, IEEE Trans. Speech and Audio Processing, Vol. 3, No. 5, pp. 357-366, 1995.
8. Y. Shiraki and M. Honda, *Speaker adaptation algorithm based on piecewise moving adaptive segment quantization method*, Proc. ICASSP, pp. II-657-660, 1990.
9. S. Furui, *Unsupervised speaker adaptation based on hierarchical spectral clustering*, IEEE Trans. on ASSP, Vol. 37, No. 12, pp. 1923-1930, 1989.
10. T. Matsuoka, N. Uemoto, T. Matsui and S. Furui, *Elaborate acoustic modeling for Japanese connected digit recognition*, Proc. IEEE Automatic Speech Recognition Workshop, Snowbird, pp. 169-170, 1995.
11. M. Morishima, T. Isobe and K. Murakami, *Telephone speech database and the experiment using CDHMM*, Proc. Acoustical Society of Japan, Fall Meeting, 2-8-8, 1994.