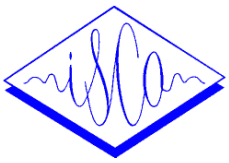

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Phoneme HMM Evaluation Algorithm without Phoneme Labeling

Yasuhiro Minami, Tatsuo Matsuoka and Kiyohiro Shikano

NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180, Japan

ABSTRACT

This paper proposes a phoneme HMM evaluation algorithm that does not require phoneme labeling and that gives a result which is independent of the recognition task. This algorithm is applied to an evaluation of speaker independent HMMs trained by using a speech database uttered by 64 speakers. The algorithm is compared with the conventional evaluation algorithm based on phoneme labels. The results show that the proposed algorithm is highly useful for HMM phoneme model evaluation.

10.21437/ICSLP.1992-198

1. Introduction

Speech recognition systems based on phoneme hidden Markov models (HMMs) have become popular for recognizing continuous speech. The performance of the phoneme HMMs greatly affects the overall performance of these kinds of systems, so better speech recognition systems will require better phoneme HMMs. We therefore need algorithms for evaluating phoneme HMMs to decide whether or not they are better.

One of the conventional algorithms for this purpose uses phoneme-labeled speech data to evaluate phoneme HMMs in terms of their phoneme recognition accuracy[1][2]. This algorithm uses speech phoneme data extracted from continuous speech based on phoneme labeling. Phoneme-labeled speech data is indispensable to this algorithm, but making such data precisely requires much time and effort because experts have to label the phonemes by looking at spectrograms of the speech data. And the results of this algorithm are sometimes affected by labeling errors in the data. Another conventional algorithm evaluates phoneme HMMs in terms of the sentence or word recognition accuracy of a recognition system. With a speech recognition system, it is easy to evaluate phoneme models with this algorithm and it does not require a labeled database. It does, however require a large amount of calculation and its result strongly depends on the task which is being treated by the system.

The algorithm proposed here does not require either phoneme labeling or a large amount of calculation, and its result is independent of the recognition task. This paper shows experimentally that the proposed algorithm can evaluate phoneme HMMs as effectively as the conventional algorithm that uses phoneme-labeled data. The proposed algorithm is therefore never affected by labeling errors.

2. Structure of Phoneme HMM

The structure of phoneme HMMs used in this paper (Fig. 1) has four states and three loops. The HMMs are continuous Gaussian-mixture-density models with diagonal Gaussian output probabilities. Their feature vectors are 16 cepstral coefficients, 16 delta cepstral coefficients, and delta power (i.e., 33 dimensions in all). The feature vectors were computed by a 16th-order Linear-Predictive-Coefficients (LPC) analysis, using a 32-ms

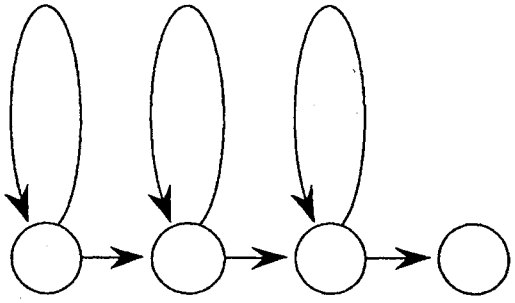

Fig. 1. Structure of phoneme HMM.

Hamming window with an 8-ms shift.

To cover all Japanese phonemes, we used 43 phoneme models: /s, sh, h, z, ch, ts, p, t, k, b, d, g, m, n, N, r, w, y, a, i, u, e, o, aa, ii, uu, ee, oo, ei, ou, sy, hy, zy, cy, py, ky, by, gy, my, ny, ry/, silence, and long stop models.

A concatenated training algorithm[3] uses unlabeled database to train the HMMs. As shown in Figure 2, the concatenated training algorithm uses the text descriptions (phoneme sequences) of utterances to construct a sentence HMM from phoneme HMMs. In this figure, the front HMM's final state is connected with the next HMM's initial state. Silence HMMs are inserted at the beginning, at the end, and at the punctuation marks in the sentence, since speech data usually have some silences in these parts. The forward-backward algorithm[4] was used to train sentence HMM's parameters, sentence by sentence. After all the sentence HMM's parameters had been calculated, the phoneme HMM's parameters were updated by decomposing sentence HMMs into phoneme HMMs.

We did not use random values as initial values of sentence HMM's parameters because the database included some long sentences. If a long-sentence HMM's parameters were trained from random values, some of the output probabilities would be small at beginning stages and multiplying these values with the forward-backward algorithm might cause an underflow problem. To avoid these problems, the speaker-independent HMMs trained using a labeled database from 10 speakers with 115 sentences each were used as initial models for training.

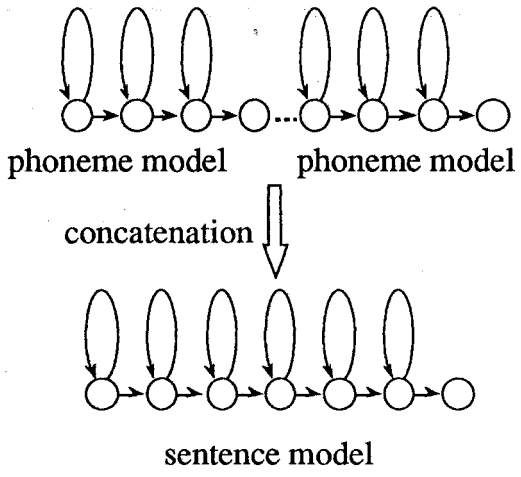

Fig. 2. Concatenated HMM model.

We used the continuous speech database that the Acoustical Society of Japan (ASJ) made from 64 speakers by using 150 utterances for each speaker. The sentences of these utterances were chosen from a set of 503 phoneme-balanced sentences made by ATR (Advanced Telecommunications Research Laboratories). Database utterances that were not mispronounced were used for training. To evaluate the proposed evaluation algorithm, 6 sets of HMMs were made from the database by varying the number of speakers from two to 64. Table 1 lists the number of speakers and the number of utterances used for training. The five training iterations were carried out with the concatenated HMM training algorithm.

3. Evaluation Algorithm without Phoneme Labeling

Although the utterances in the speech database are unlabeled, text descriptions (phoneme sequences) are supposed to be given. The proposed algorithm uses the text descriptions (phoneme sequences) of the utterance to form a sentence HMM from the phoneme HMMs.

Figure 3 shows the concept of the proposed algorithm. As shown in the top sentence HMM in the figure, first a correct sentence HMM is constructed by using an utterance text description. The phoneme HMM to be evaluated in the sentence HMM (the phoneme surrounded by the square in the top sentence HMM) is swapped one by one with other phoneme HMMs. Thus all possible phoneme

Table 1. Numbers of training speakers and utterances.

speakers (male/female)	utterances
2 (1/1)	300
4 (2/2)	600
8 (4/4)	1200
16 (8/8)	2400
32 (14/18)	4801
64 (30/34)	9603

HMMs are substituted in turn for this phoneme HMM. The sentence HMMs from second to bottom in the figure are phoneme-swapped sentence HMMs. The likelihood value for the sentence HMM is calculated every time a phoneme HMM is swapped, and quality of the phoneme HMM is evaluated by the likelihood value of the sentence HMMs. When the likelihood value of the correct sentence HMM is greater than any other values of the phoneme-swapped sentence HMMs, it means that the phoneme is recognized correctly, and that the phoneme HMM is well trained. To evaluate every phoneme HMM in the sentence HMM, the phoneme swapping is carried out one by one from the first phoneme in the sentence HMM to the last phoneme. To calculate the phoneme recognition rate, the number of phonemes which is recognized correctly in a database is counted and is divided by the number of phonemes in the database.

4. Calculation Reduction of Phoneme HMM Evaluation Algorithm

Figure 4 illustrates a trellis whose vertical and horizontal axes show the states of the correct sentence HMM and the input speech frames. The likelihood of speech data is usually calculated by sweeping through this trellis, but a large amount of calculation would be required, if the whole trellis is calculated every time a phoneme HMM is swapped. We propose a new algorithm which uses the forward path trellis and backward path trellis to calculate this likelihood value efficiently. First the forward and backward trellises for the correct sentence HMM are calculated and stored in the matrix (the states by frames). These trellises are used to reduce the amount of calculation. Let us suppose that the second phoneme HMM, the /r/ HMM, is swapped. Then the forward path trellis before this swapped phoneme and the backward path trellis after it have already been calculated. Only the swapped part of the trellis (unshaded portion in the figure) needs to be calculated in order to obtain the likelihood value of the phoneme-swapped sentence HMM. This likelihood value is efficiently calculated by concatenating the forward path trellis, swapped phoneme trellis, and backward path trellis. Consequently, when phoneme-swapped sentence HMMs are constructed every time a phoneme HMM is swapped, only the part of the trellis for the swapped phoneme HMM needs to be calculated.

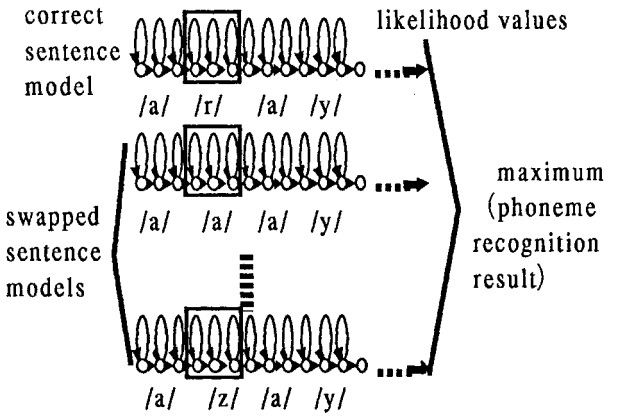

Fig. 3. Concept of the proposed algorithm.

5. Evaluation of Phoneme HMM Evaluation Algorithm

The proposed algorithm was used to evaluate phoneme HMMs trained in the concatenated fashion. The proposed algorithm was compared with the conventional phoneme recognition algorithm that uses phoneme labeling. We used speech data obtained from six speakers (four

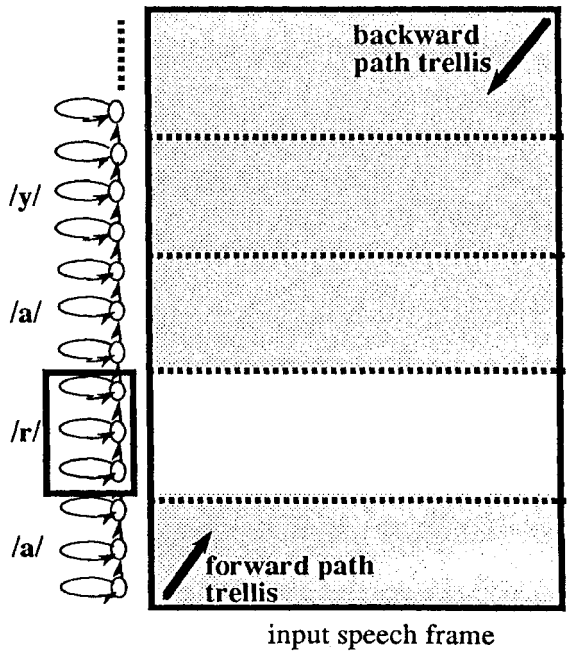

Fig. 4. Trellises of forward and backward paths.

female and two male) uttering the first 50 sentences in the phonetically balanced continuous speech database made by ATR. We evaluated 23 phoneme HMMs: /s, sh, h, z, ch, ts, p, t, k, b, d, g, m, n, N, r, w, y, a, i, u, e, o/. Both algorithms showed increases in phoneme recognition rates with increases in the number of training speakers up to 64 (Fig. 5). The results of the proposed algorithm are proportional to those of the conventional algorithm. Although the phoneme recognition rates of the conventional algorithm are always higher than those of the proposed algorithm, this is not important because it is only necessary to know which model is better. For example, this figure shows that both algorithms tell us that the 32-speaker training is almost enough for building speaker-independent phoneme HMMs. The proposed algorithm can evaluate phoneme HMMs as well as the conventional algorithm can, even though it does not use phoneme labeling. Furthermore, the proposed algorithm is never affected by segmentation errors.

6. Conclusions

We have described a phoneme evaluation algorithm that does not require phoneme labeling, and we have evaluated this algorithm by comparing it with the conventional algorithm. Using concatenated HMMs made from a database by varying the number of speakers from two speakers to 64 speakers has shown that the proposed algorithm gives results that are proportional to those given by conventional algorithm and that it can evaluate phoneme HMMs as well as the conventional algorithm does. We will apply the proposed algorithm to evaluation of various HMMs, such as the discrete HMM, the tied-mixture HMM, and the continuous HMM.

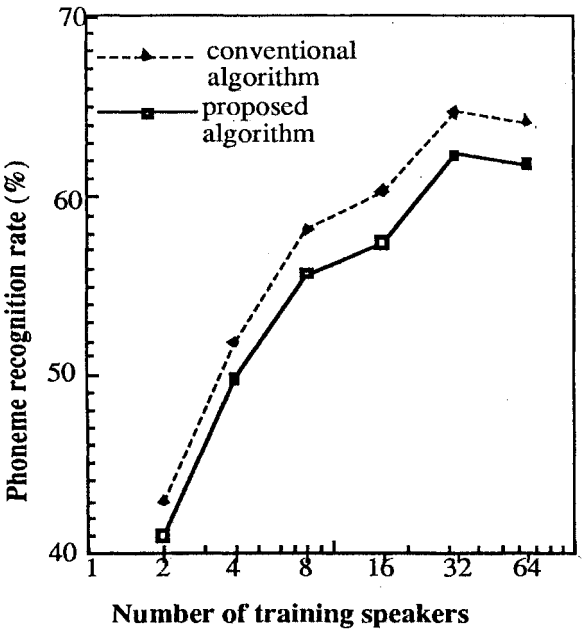

Fig. 5. Relationship between phoneme recognition rate and the number of training speakers.

7. Acknowledgments

We thank Dr. Sadaoki Furui, the director of the Furui Research Laboratory of the NTT Human Interface Laboratories, and members of the laboratory for useful discussions. We used the ASJ speaker independent continuous speech database and ATR speech database.

References

[1] Y. Minami, T. Hanazawa, H. Iwamida, E. McDermott, K. Shikano, S. Katagiri, M. Nakagawa, "On the Robustness of HMM and ANN Speech Recognition", ICSLP'90, pp. 1345-1348 (November 1990).

[2] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang, "Phoneme recognition using time-delay neural networks", IEEE Trans. on ASSP, pp. 328-339, vol. 37, No. 3 (March 1987).

[3] K. F. Lee and H. W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM", ICASSP'88, pp. 123-126 (April 1988).

[4] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition", The Bell System Technical Journal, vol. 62, No. 4 (April 1983).