FRAME LEVEL LIKELIHOOD TRANSFORMATIONS FOR ASR AND UTTERANCE VERIFICATION

Konstantin P. Markov, Satoshi Nakamura

ATR Spoken Language Translation Research Labs.
2–2–2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan

ABSTRACT

In most of the current speech recognition systems based on HMM, existing decoding and utterance verification methods make use of state output likelihood as a measure of the acoustic match between the input data and the acoustic models. In this paper, we present a new and more generalized approach to the formation of the acoustic match score. The essence of this approach is to transform the likelihood of each acoustic vector with respect to any particular HMM state according to some non-linear function. We have investigated two types of such transformation functions. The first one, performs likelihood normalization, and the second one transforms likelihoods into exponentially ordered weights. The transformed likelihoods, as new acoustic scores, are used further for decoding, recognition and verification instead of the conventional likelihoods. In our evaluation experiments we used TIMIT database for phoneme recognition and verification and a database of 710 speakers and a total of 4252 distinct words, for isolated word recognition and verification. The results we achieved show that the transformed likelihood scores, in average, increase slightly the recognition accuracy and reduce the verification error rates up to 30%.

1. INTRODUCTION

Acoustic modeling based on hidden Markov model has proven to be very successful method and most of the current state-of-the-art speech recognition systems utilize the HMM for modeling temporal and spectral characteristics of speech signals. Speech recognition is a task where the aim is to find that sequence of speech units (phonemes, words, etc.) which most probably would have generated given sequence of observation vectors. Essential for this process is to have confident and reliable information about the match between the observed data and the models. Currently, the HMM output likelihood is the most widely used numerical representation of this match. Generally speaking, the likelihood is a matching score, as is the Euclidean distance in vector quantization or in former template based approach to speech recognition. Since all further processing is based on these scores, although a higher level knowledge can be applied, from pattern recognition point of view it is essential that the probability density functions (pdf) of the acoustic scores for correct and incorrect classes do not overlap. In practice, however, this cannot be achieved, mainly because we don't have accurate knowledge of the true data distribution and due to errors in estimation of the parameters of this distribution. Extensive research has been conducted in order to find reliable methods for estimation of data distribution. However, what really matters is the pdfs of the acoustic match scores for correct and incorrect classes and the distance between them as was discussed in [1].

Our approach to the above problem is to transform likelihood scores according to some function. In this transformation we can include some additional knowledge about non-target (incorrect) classes, which could make our scores more confident measure of the acoustic match and increase the separation between these scores pdfs for correct and incorrect classes.

We have investigated two kinds of such transformation functions. The fist one performs a likelihood normalization, technique widely used in speaker verification [2, 3], but applied here at frame level. It is based on likelihood ratio which has been used directly for utterance verification in several studies [4, 5, 8], but in our study, we use it only to obtain new acoustic scores. The second type of likelihood transformation, called Weighting Models Rank (WMR), transforms the likelihood into exponentially ordered weights which are used further as acoustic scores. The weights correspond to the model's rank from a list of all models sorted according to their likelihood score with respect to a particular input frame. This approach is similar to the rank-ordering method proposed in [6]. The difference is that in out case weights are ordered exponentially and that they are not used directly for utterance verification.

2. FRAME LEVEL LIKELIHOOD TRANSFORMATIONS

Transformed likelihood scores are obtained using the following general formula:

$$Sc(o|\lambda) = f(p(o|\lambda)) \tag{1}$$

where $p(o|\lambda)$ denotes the likelihood of observation vector o with respect to model λ representing some particular data class. In order to be meaningful, the transformation function $f(x)$ should be non-linear and satisfy the following condition [11]:

$$if \quad x_1 > x_2, \quad then \quad f(x_1) > f(x_2) \tag{2}$$

As a special case, $Sc(o|\lambda)$ is equal to the standard likelihood score when the transformation function is of the type $f(x) = x$.

2.1. Likelihood normalization

Given a single frame likelihood $p(o_t|\lambda_i)$ from the i^{th} class model with respect to frame o_t, the likelihood is transformed using the following function:

$$Sc(o_t|\lambda) = \frac{p(o_t|\lambda_i)}{\frac{1}{B}\sum_{b=1}^{B} p(o_t|\lambda_b)} \tag{3}$$

where $p(o_t|\lambda_b)$ are the frame likelihood scores from background class models given the same frame o_t. Different choices of the background class set give different transformation functions. Note that the above likelihood transformation approximates the likelihood ratio, but for a single frame.

Given the class model i, we have experimented with the following background class sets:

- **All others** - the background class set consists of all classes, except the class i.
- **Cohort** - the background class set consists of K acoustically closest classes to the class i. The cohort classes are determined on the training data in advance.

2.2. Weighting models rank (WMR)

The main idea is to transform the frame likelihood $p(o|\lambda)$ into a weight w which does not depend on the absolute value of this likelihood, but depends on its relative position with respect to the likelihoods from all other classes.

The WMR transformation consists of the following two steps.

- **Step 1**. For each test vector $o_t, t = 1, 2, \ldots, T$, calculate all likelihoods $p(o_t|\lambda_i), i = 1, \ldots, N$ and sort them in a decreasing order. Each model is assigned a *rank* - r_λ, which corresponds to the position of the model in the sorted list and is an integer ranging from 1 to N. The weight w is defined as a function of r_λ:

$$w(r_\lambda) = g(r_\lambda) \tag{4}$$

- **Step 2**. For each model λ_i, find its rank r_{λ_i}, i.e. its place in the N-best list, and instead of the likelihood $p(o_t|\lambda_i)$ use the corresponding weight $w_t(r_{\lambda_i})$ as a model's frame score.

Now we can define the WMR type likelihood transformation function as:

$$Sc(x_t|\lambda) = \exp(w_t(r_\lambda)) \tag{5}$$

Obviously, in this technique, the most important issue is what types of function $g()$ to use. Previous study [11] has shown that the following exponential function is appropriate:

$$g_{exp}(r_\lambda) = \exp(A - Br_\lambda), r_\lambda = 1, \ldots, N \tag{6}$$

How to choose the parameters A and B is also explained in [11].

3. RECOGNITION SYSTEM

3.1. Overview

Our speech recognition system is based on continuous density hidden Markov models. State output pdf consists of mixture of Gaussian components with diagonal covariance matrices. Each sub-word unit (phoneme in our experiments) is modeled by a 3 state left-to-right HMM without state skips.

HMMs are trained using up to 10 iterations of the standard Baum-Welch algorithm. Initial models are made by random data sampling.

3.2. Decoding algorithm

Our decoder is a simple one-pass Viterbi decoder. It can operate in two modes: with and without language model. The former mode gives pure acoustic level utterance likelihood and the output is acoustically most probable phoneme sequence. In the later mode, phoneme bigram or finite state grammar (FSG) can be used as language model. When the decoder is given an input frame o_t, likelihoods from all states S_j from all

models $p(o_t|S_j)$ are calculated and then Viterbi search is performed as:

$$\delta_t(j) = \max_i[\delta_{t-1}(i) + \log a_{ij}] + \log p(o_t|S_j) \qquad (7)$$

where $\delta_t(j)$ is the best score along a single path at time t, which accounts for the first t frames and ends in state j.

A slight modification is needed in this algorithm in order to incorporate the transformed likelihood scores. After likelihoods are calculated, the only additional step is the transformation step. Then the Viterbi search is done as usual. Therefore, the modified algorithm will be:

$$\delta_t(j) = \max_i[\delta_{t-1}(i) + \log a_{ij}] + \log Sc(o_t|S_j) \qquad (8)$$

3.3. Utterance verification

The utterance verification algorithm we adopted is based on word scores and uses likelihood ratio for making decision. An utterance (an isolated word in our experiments) is accepted if:

$$L = \frac{L(O|FSG)}{L(O|NoGrammar)} \qquad (9)$$

is above some threshold. In this equation, $O = o_1, \ldots, o_T$ is the observation sequence, $L(O|FSM)$ and $L(O|NoGrammar)$ are accumulated likelihood (or transformed likelihood) scores with and without FSG. Similar likelihood ratio has been used previously for key phrase spotting [7].

This utterance verification algorithm requires two decoding passes: one with grammar and one without grammar.

4. EXPERIMENTAL RESULTS

4.1. Speech material

The TIMIT corpus was used for phoneme recognition and verification experiments. Speech data were converted into 39 dimensional MFCC feature vectors (power, 12 cepstral coefficients and their delta and delta delta) with window of 25 ms. and shift of 10 ms. 48 monophone left-to-right HMMs with 3 states / 8 mixtures were trained from the suggested training data.

The second database which was used for word recognition and verification consists of isolated word recordings from 710 speakers uttering several repetitions of groups of 4255 distinct words. 540 speakers and 6474 utterances were used for training of 26 monophone 3 state / 8 mixture left-to-right HMMs. The test set consisted of 170 speakers and 2030 utterances. Front-end speech processing was the same as for TIMIT database.

4.2. Phoneme recognition and verification

When implementing the likelihood transformation technique we have to define the distinct classes. In our experiments, each HMM state represents a separate class.

Since all TIMIT utterances are phonetically labeled, in addition to connected phoneme recognition experiments we were able to run phoneme classification experiment. The evaluation results are summarized in Table 1 where the third row shows the recognition rates using phoneme bigram language model. Each column shows the results for different acoustic score. "Lik." which stands for likelihood, is our baseline. Columns "ALL" and "Coh." show the performance of the transformed likelihood scores using Eq.(3). Column "WMR" is for likelihood scores transformed using WMR technique. As can be seen, the new scores perform not worse than the standard likelihood with exception of WMR scores. However, the drop in the recognition rates is negligible. These results are in contrast to those reported in [8], where frame likelihood normalized scores were used for decoding as well, and significant drop in the performance was observed.

Table 1: 48 phoneme set recognition results.

	Scores			
	Lik.	All	Coh.	WMR
Classification (%)	66.76	66.82	66.84	65.94
Recognition (%)	50.89	51.48	51.57	50.60
+ bigram (%)	61.96	62.27	62.57	61.82

In order to test whether the new scores have led to increased separability between different phonemes, we performed phoneme verification experiment, where decision was taken by comparing phoneme scores with a threshold. Table 2 shows the verification equal error rates (EER). Using new scores the EER dropped in average by 30%.

4.3. Isolated words recognition and verification

The same set of scores was used in these experiments. Recognition rates were obtained after the first decoding pass using FSG. The second pass was used in order to calculate the likelihood ratio of Eq.(9).

Table 2: Phoneme verification equal error rates for different scores.

	Scores			
	Lik.	All	Coh.	WMR
EER (%)	41.3	28.9	28.7	28.5

Table 3 shows both the word error rates and the utterance verification EER. In the case of isolated words, the WMR scores performed best reducing the WER by 4.3% and the word verification EER by 20%. The different performance of the WMR scores with respect to phoneme recognition case can be explained with the different segment duration over which the WMR scores are accumulated. Average phoneme duration is several times shorter than the words duration and the effect of WMR scores is more significant for longer utterances.

Table 3: Isolated word error and utterance verification equal error rates.

	Scores			
	Lik.	All	Coh.	WMR
WER (%)	6.90	6.80	6.72	6.60
EER (%)	30.33	27.23	27.10	24.24

5. CONCLUSION

We have proposed and evaluated new types of acoustic match score which are generalization of the standard likelihood score. The new scores accommodate some additional information about the non-target classes which allows for bigger separability between scores distributions for target and non-target classes.

Evaluation experiments on two databases show that, in contrast to some other studies, there is no drop in the recognition performance when using the normalized likelihood for decoding ("All" and "Coh." cases) and that a significant reduction of the verification error rates can be achieved using the proposed likelihood scores.

6. REFERENCES

[1] B. S. Atal, "Automatic speech recognition: A communication perspective," in *Proc. ICASSP*, pp. 457–460, 1999.

[2] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech Communication*, vol. 17, pp. 109–116, Aug. 1995.

[3] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, Aug. 1995.

[4] E. Lleida and R. Rose, "Efficient decoding and training procedures for utterance verification in continuous speech recognition," in *Proc. ICASSP*, pp. 507–510, 1996.

[5] N. Moreau and D. Jouvet, "Use of a confidencs measure based on frame level likelihood ratios for the rejection of incorrect data," in *Proc. Eurospeech*, vol. I, pp. 291–294, 1999.

[6] Q. Lin, S. Das, D. Lubenski, and M. Picheny, "A new confidence measure based on rank-ordering subphone scores," in *Proc. ICSLP*, pp. 3249–3252, 1998.

[7] Q. Lin, D. Lubenski, M. Picheny, and P. S. Rao, "Key-phrase spotting using an intergated language model of N-grams and finite-state grammar," in *Proc. ICSLP*, pp. 255–258, 1997.

[8] L. K. Leung and P. Fung, "A more efficient and optimal LLR for decoding and verification," in *Proc. ICASSP*, pp. 689–692, 1999.

[9] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. on SAP*, vol. 4, pp. 420–429, Nov. 1996.

[10] M. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative utterance verification for connected digits recognition," *IEEE Trans. on SAP*, vol. 5, pp. 266–277, May 1997.

[11] K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using non-linear frame likelihood transformation," *Speech Communication*, vol. 24, pp. 193–209, June 1998.