MAXIMUM LIKELIHOOD WEIGHTING OF DYNAMIC SPEECH FEATURES FOR CDHMM SPEECH RECOGNITION

Javier Hernando

Dep. Signal Theory and Communications
Universitat Politècnica de Catalunya, 08034 Barcelona, Spain
javier@gps.tsc.upc.es

ABSTRACT

Speech dynamic features are routinely used in current speech recognition systems in combination with short-term (static) spectral features. Although many existing speech recognition systems do not weight both kinds of features, it seems convenient to use some weighting in order to increase the recognition accuracy of the system. In the cases that this weighting is performed, it is manually tuned or it consists simply in compensating the variances. The aim of this paper is to propose a method to automatically estimate an optimum state-dependent stream weighting in a CDHMM recognition system by means of a maximum-likelihood based training algorithm. Unlike other works, it is shown that simple constraints on the new weighting parameters permit to apply the maximum-likelihood criterion to this problem. Experimental results in speaker independent digit recognition show an important increase of recognition accuracy.

1. INTRODUCTION

The so-called dynamic features [1] are able to somewhat represent the time evolution of the spectrum of speech signals by providing smoothed estimates of the derivatives of the spectral parameter trajectories in the current frame, and their use reduces noticeably the recognition error rate.

Although many existing speech recognition systems do not weight dynamic features with respect to static features, it seems convenient to use some kind of weighting in order to increase the recognition accuracy of the system. When a weighting is performed, usually the same set of weights is used for every frame and so they can be estimated empirically. In most of cases, such a weighting is just manually tuned [2] or it consists simply in compensating the variances [3].

When the recognition system is based on hidden Markov modeling (HMM), there is no reason to believe that there are features which are more important for some states of the models that the others. Especially, one would expect the beginning and ending segments of a phoneme to be more context dependent than the middle part, so in that case the probability estimator of the speech recognizer should put more emphasis on dynamic features [4] [5]. Experiments have shown that static cepstra are more important than dynamic ones [6]. Thus, the recognition performance must improve using individual stream weights for every HMM state. In this case, an automatic algorithm to learn these weights is needed.

On the other hand, one main reason for the popularity and the success of HMM in its application to speech recognition has been the advent to an efficient maximum-likelihood (ML) based estimation method, the forward-backward algorithm [7]. However, in principle, ML methods applied to the estimation of these stream weights would invariably discard the stream with the lowest probability [8]. This result, although provides an obvious maximum in the objective function, does not seem reasonable for recognition purposes. Because of this fact, until now, only good results have been shown by training state-dependent stream weights in a discriminative way [5] [8] [9], including the author.

The aim of this paper is twofold: 1) to find a meaningful interpretation of these stream weights to provide more insight into their performance (section 2), and 2) to propose a solution to automatically estimate the optimum weighting of static and dynamic features based on the ML principle (section 3). The considered recognition system is based on continuous-density hidden Markov modeling (CDHMM)

2. WEIGHTING SPEECH FEATURES IN CDHMM

When dynamic features are employed in continuous-

This work has been supported by the grant TIC95-0884-C04-02

density hidden Markov modeling (CDHMM), usually the feature vector O^t is composed by concatenating static and dynamic features. In this case, good results have been obtained if dynamic features are scaled in order to equalize the variances of both kinds of features [3].

An alternative approach is to consider two separate vectors for static and dynamic features, O_1^t and O_2^t, respectively, and to assume that both streams are statistically independent. In this case, for a given state j of a model the probability that a feature vector is observed can be written as

$$b_j(O^t) = \prod_{s=1}^{2} b_{js}(O_s^t) \tag{1}$$

where s indexes both streams. If distributions are modeled by mixtures of L multivariate Gaussian functions

$$b_{js}(O_s^t) = \sum_{k=1}^{L} c_{jsk} N(O_s^t, m_{jsk}, V_{jsk}) \tag{2}$$

where N is a Gaussian pdf of mean vector m_{jsk} and covariance matrix V_{jsk}.

Many existing CDHMM-based speech recognition systems restrict the covariance matrices to be diagonal in order to increase the trainability of the models and reduce the computational complexity of the system. In this case, it is straightforward to show that both joint and separate approaches are equivalent.

In any case, the separate formulation (1) can be slightly modified to permit a very simple stream weighting to reflect the relative importance of the various streams for recognition, as

$$b_j(O^t) = \prod_{s=1}^{2} \left(b_{js}(O_s^t)\right)^{w_{js}} \tag{3}$$

where w_{js} are the weighting coefficients.

It is worth noting that, in the case of one mixture and diagonal covariance matrices, this stream weighting has a meaningful interpretation. In fact, it amounts to modifying the implicit variance-weighting of the Euclidean distance of the Gaussian exponent. It can be seen if (3) is explicitly rewritten in scalar notation as

$$b_j(O^t) = \prod_{s=1}^{2} \left[(2\pi)^{N_s} \prod_{n=1}^{N_s} \sigma_{jsn}^2 \right]^{-\frac{w_{js}}{2}} \exp\left[-\frac{1}{2} \sum_{s=1}^{2} \sum_{n=1}^{N_s} \frac{w_{js}}{\sigma_{jsn}^2} \left(O_{sn}^t - \mu_{jsn}\right)^2 \right] \tag{4}$$

where O_{sn}^t is the n-th component of the stream O_s^t, N_s is the number of components of O_s^t, and μ_{jsn} and σ_{jsn}^2 are the mean and variance of O_{sn}^t, respectively.

3. MAXIMUM LIKELIHOOD ESTIMATION OF THE WEIGHTS

An automatic algorithm for learning these state-dependent stream-weights is needed. Considering that one main reason for the popularity and the success of HMM in its application to speech recognition has been the advent to an efficient maximum likelihood based estimation method, the forward-backward algorithm [7], it would be desirable to apply ML criterion to this problem.

However, in principle, ML methods would invariably discard the stream with the lowest probability [8]. The application of the ML principle requires to impose some constraint to the stream weights. If the constraint consists in imposing that the sum of the stream weights of a state is constant, ML methods would lead to this constant for the stream which provides the highest value probability and zero for the other stream [8] [9]. This result, although provides an obvious maximum in the objective function, does not seem reasonable for recognition purposes. Because of this, only good results have been obtained by training state-dependent stream weights in a discriminative way [5] [8] [9], including the author.

In this work, the author shows that a simple and efficient forward-backward based algorithm to learn the stream weights is possible using simple constraints to the weights.

One possible constraint is

$$\sum_{s=1}^{2} (w_{js})^m = K \tag{5}$$

that is, to impose a constant L^m norm, m?0,1, of the stream weight vector (w_{j1}, w_{j2}). Notice that for m=1 this constraint would be the one mentioned above.

Let us now derive the reestimation formula of the new parameters by maximizing the partial Baum's auxiliary function [7] of the observation probabilities of the model λ $Q_b(\lambda, b_j')$ as a function of the weights constrained by (5). So, we have to solve the equations

$$\frac{\partial}{\partial w_{js}'} \left[Q_b(\lambda, b_j') - \theta \left(\sum_{s=1}^{2} (w_{js}')^m - K \right) \right] = 0 \tag{6}$$

for s=1,2, where θ is the so-called Lagrange multiplier

In terms of the well-known forward $\alpha_j(O_t)$ and backward $\beta_j(O_t)$ variables, the expression of Q_b is

$$Q_b(\lambda, b_j') = \sum_{t=1}^{T} \alpha_j(O^t)\beta_j(O^t)\log b_j'(O^t) \quad (7)$$

where T is the utterance length; and, using the separate formulation (1), it can be written explicitly in terms of the stream weights as

$$Q_b(\lambda, b_j') = \sum_{t=1}^{T} \alpha_j(O^t)\beta_j(O^t)\sum_{s=1}^{2} w_{js}' \log b_{js}'(O_s^t) \quad (8)$$

Including this expression in (6), the derivative with respect to w_{js}' leads to this set of equations

$$\sum_{t=1}^{T} \alpha_j(O^t)\beta_j(O^t)\log b_{js}'(O_s^t) - \theta m (w_{js}')^{m-1} = 0 \quad (9)$$

for s=1,2.

As these equations are linear and uncoupled, from (5) and (9) it is straightforward to obtain the following reestimation formula for the weights

$$w_{js} = \left\{ K \frac{\left(\sum_{t=1}^{T} \alpha_j(O^t)\beta_j(O^t)\log b_{js}(O_s^t) \right)^{\frac{m}{m-1}}}{\sum_{s=1}^{2} \left(\sum_{t=1}^{T} \alpha_j(O^t)\beta_j(O^t)\log b_{js}(O_s^t) \right)^{\frac{m}{m-1}}} \right\}^{\frac{1}{m}} \quad (10)$$

The numerator of this expression can be considered related to the quantity of information provided by a specific stream while the denominator normalizes this information with the contribution of both streams. The extension of this expression to multiple utterances is straightforward as in the case of the conventional HMM parameters.

Other possible constraint is

$$\sum_{s=1}^{2} m^{w_{js}} = K \quad (11)$$

that leads, in this case, to the following set of equations and reestimation formula

$$\sum_{t=1}^{T} \alpha_j(O^t)\beta_j(O^t)\log b_{js}'(O_s^t) - \theta m^{w_{js}'} \log m = 0 \quad (12)$$

$$w_{js} = \frac{1}{\log m} \log \left\{ K \frac{\sum_{t=1}^{T} \alpha_j(O^t)\beta_j(O^t)\log b_{js}(O_s^t)}{\sum_{s=1}^{2} \left[\sum_{t=1}^{T} \alpha_j(O^t)\beta_j(O^t)\log b_{js}(O_s^t) \right]} \right\} \quad (13)$$

4. RECOGNITION RESULTS

The database used in the recognition experiments was the isolated adult portion (112 speakers for training and 113 for testing) of the speaker independent digit TI [11] database. The initial sampling frequency 20 kHz was converted to 8 kHz.

The HTK recognition system, based on the Continuous-Density Hidden Markov Models (CDHMM), was appropriately modified to perform the maximum likelihood weighting of the speech features and used for the recognition experiments. In the parameterization stage, the signal was preemphasized with $1 - z^{-1}$ and was divided into frames of 30 ms at a rate of 10 ms, and each frame was characterized by its energy and 12 cepstral parameters obtained by linear prediction (LPC), with prediction order equal to 10. Regression analysis over 70 ms was applied to the static energy sequence and the static cesptrum sequence to obtain dynamic features, delta-energy and delta-cepstrum, respectively. Each digit was characterized by a first order, left-to-right, Markov model of 10 states with one mixture of diagonal covariance matrix and without skips. The same structure was used for the silence model but only with 5 states. For the conventional parameters, training was performed in two stages using Segmental k-means and Baum-Welch algorithms. Testing was performed using Viterbi algorithm.

Using the reestimation formula (10), preliminary experiments showed that the initial weighting coefficients do not need to be tuned previously to optimize recognition performance, and excellent results were obtained for m=2. In this case, the algorithm is very robust to the value of K. Figure 1 shows the number of recognition errors obtained for several values of K and m. Considering that the number of errors of the baseline system -all stream weights equal to 1- is 27 (1,09 % error rate), it can be seen that a big and consistent error reduction can be obtained with the proposed approach. For instance, for m=K=2 (Euclidean norm of the weight vector as in the baseline system), the number of errors was 14, that is, an error reduction of 48 %. And, in several tests, only there were 12 errors, that is, a 0.48 % error rate and a 56 % error reduction.

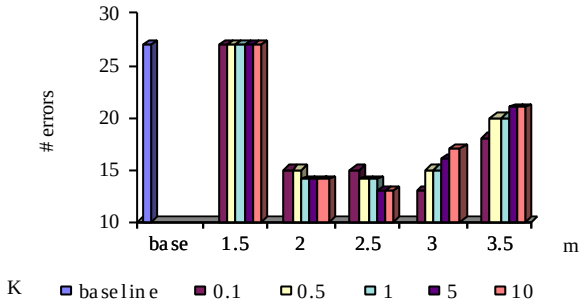

Figure 1. Recognition errors for isolated digits

In general, it was observed that the algorithm weights more the dynamic features than the static ones. As the variances of dynamic features are lower than the variances of the static features, this kind of weighting emphasizes the implicit variance-weighting in the exponent of the Gaussian distributions (4).

Furthermore, it was observed that the estimation of the weights has a big influence on the estimation of the variances of the observation distributions. In particular, the algorithm tends to increase large variances and to decrease low variances. This effect seems to be useful. Training together stream weights and observation distribution but setting at the end the weights to 1, there were 21 errors for m=K=2, that is, half of the improvement of the algorithm is due to this effect.

The reestimation formula (13) is more sensitive to the value of K. However, also good results have been obtained in the range of 12=K=20. Concretely, for m=2 and K=15 only there were 15 errors, i.e. a 0.62 % error rate and a 44 % error reduction.

5. CONCLUSIONS

In this work a ML-based algorithm to automatically estimate the optimum weighting of static and dynamic features in each state of each model of a CDHMM-based speech recognition system has been proposed. Although, in principle, ML methods applied to his problem would invariably discard the stream with the lowest probability, the authors have shown that appropriate constraints upon the stream weights lead to simple and efficient algorithms, and to an important and consistent improvement of speech recognizer performance. The recognition results obtained by means of this approach have been relevant in digit recognition: a 56 % error reduction from the baseline system. Further experiments on different recognition tasks and systems are under development to extend the results shown in this paper.

ACKNOWLEDGMENTS

The author want to acknowledge Albino Nogueiras for his suggestions, and to Javier Valverde and Juan Ayarte for their help in the software development.

REFERENCES

[1] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans. ASSP, vol. 34, pp. 52-59, 1986.

[2] K.F. Lee, *Automatic Speech Recognition. The Development of the SPHINX System*, ed. Kluwer Academic Publishers, 1989.

[3] J.G. Wilpon, C.H. Lee, L.R. Rabiner, "Improvements in Connected Digit Recognition Using Higher Order Spectral and Energy Features", Proc. ICASSP'91, pp. 349-352, Toronto, May 1991.

[4] I. Rogina, A. Waibel, "Learning State-Dependent Stream Weights for Multi-Codebook HMM Speech Recognition Systems", Proc. ICASSP´94, vol. I, pp. 217-220.

[5] Y. Normandin, R. Cardin, R. De Mori, "High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation", IEEE Trans. on SAP, vol. 2, no. 2, pp. 299-311, 1994.

[6] E.L. Bocchieri, J.G. Wilpon, "Discriminative Feature Selection for Speech Recognition", Computer Speech and Language, vol. 7, pp. 229-246, 1993.

[7] L. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Models", Inequality, vol. III, pp. 1-8, 1972.

[8] Y.L. Chow, "Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition Using the N-Best Algorithm", Proc. ICASSP'90, pp. 701-704.

[9] C. Martin, F.J. Caminero, C. de la Torre, L. Hernández, "Codebook Weights Adaptation for Discriminative Training of SCHMM-Based Speech Recognition Systems", Proc. EUROSPEECH'95, pp. 93-96.

[10] J.Hernando, J. Ayarte, E. Monte, "Optimization of Speech Parameter Weighting for CDHMM Word Recognition", Proc. EUROSPEECH'95, Madrid, September 1995, pp. 105-108.

[11] R.G. Leonard, "A Database for Speaker-Independent Digit Recognition", Proc. ICASSP'84, pp. 42.11.1-4, March 1984.