DISCRIMINATIVE WEIGHTING OF DYNAMIC FEATURES IN CONTINUOUS-DENSITY HIDDEN MARKOV MODELS FOR WORD RECOGNITION

J. Hernando and J. Ayarte

Dpto. Teoría de la Señal y Comunicaciones
Universidad Politécnica de Cataluña

ABSTRACT:

Speech dynamic features, which provide smoothed estimates of the derivatives of the spectral parameter trajectories in the current frame, are routinely used in current speech recognition systems in combination with short-term (static) spectral features. The aim of this paper is to propose a method to automatically estimate the optimum ponderation of static and dynamic features in a speech recognition system. The recognition system considered in this paper is based on Continuous-Density Hidden Markov Modelling (CDHMM), widely used in speech recognition. Our approach consists basically in 1) adding two new parameters for each state of each model that weight both kinds of speech features, and 2) estimating those parameters by means of a discriminative training algorithm that minimizes the recognition error using the recently proposed Generalized Probabilistic Descent (GPD) method. Experimental results in speaker independent digit recognition show an important increase of recognition accuracy.

1. INTRODUCTION

The so-called dynamic features [1] are routinely used in current speech recognition systems in combination with short-term (static) spectral features. As their computation encompasses several adjacent frames, they 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 ponderate 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. In the cases that a ponderation is performed, it is manually tuned [2] or it consists simply in compensating the variances of both kinds of features [3].

The aim of this paper is to propose a method to automatically estimate the optimum ponderation of static and dynamic features in a speech recognition system. The recognition system considered in this paper is based on Continuous-Density Hidden Markov Modelling

(CDHMM) [4], widely used in speech recognition. Our approach consists basically in 1) adding two new parameters for each state of each model that weight both kinds of speech features, and 2) estimating those parameters by means of a discriminative training algorithm that minimizes the recognition error using the recently proposed Generalized Probabilistic Descent (GPD) method [5].

This paper is organized in the following way. In section 2 the new parameters of the models that weight dynamic features in the CDHMM-based speech recognition system are introduced. In section 3 a brief revision of the minimum recognition error formulation and the GPD method is provided. In section 4, the discriminative learning theory is applied to estimate the new parameters introduced in section 2. Section 5 reports experimental results in speaker independent digit recognition that show an important increase of recognition accuracy. Finally, in section 6 some conclusions are summarized.

2. WEIGHTING SPEECH FEATURES IN CONTINUOUS-DENSITY HIDDEN MARKOV MODELS

In Continuous-Density Hidden Markov Modelling (CDHMM) [4], for a given state j of the model i the probability that a feature vector $\mathbf{O}_t$ is observed, $b_{ij}(\mathbf{O}_t)$, is typically modelled by a mixture of L multivariate Gaussian functions, i.e.

$$b_{ij}(\mathbf{O}_t) = \sum_{k=1}^{L} c_{ijk}\, N(\mathbf{O}_t, \mu_{ijk}, \Sigma_{ijk}) \qquad (1)$$

where N(.) is a Gaussian probability density function of mean vector μ_{ijk} and covariance matrix Σ_{ijk} and c_{ijk} is known as mixture coefficient.

When dynamic features are employed, usually, the feature vector $\mathbf{O}_t$ is composed by concatenating static and dynamic features. In some cases, 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, $\mathbf{O}^1_t$ and $\mathbf{O}^2_t$, respectively, and to assume that both vectors are statistically independent. In this case,

$$b_{ij}(\mathbf{O}_t) = \prod_{s=1}^{2} \sum_{k=1}^{L} c^s_{ijk}\, N(\mathbf{O}^s_t, \mu^s_{ijk}, \Sigma^s_{ijk}) \qquad (2)$$

where N(.) is a Gaussian probability density function of mean vector μ^s_{ijk} and covariance matrix Σ^s_{ijk} and c^s_{ijk} is the mixture coefficient.

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 that case, it is straightforward to show that expressions (1) and (2) are identical.

However, in any case, the separate formulation (2) can be slightly modified to permit a very simple ponderation of both kinds of features using two exponential weights $\{\gamma^s_{ij}\}_{s=1,2}$, i. e.

$$b_{ij}(\mathbf{O}_t) = \prod_{s=1}^{2} \left(\sum_{k=1}^{L} c^s_{ijk}\, N(\mathbf{O}^s_t, \mu^s_{ijk}, \Sigma^s_{ijk}) \right)^{\gamma^s_{ij}} \qquad (3)$$

(notice that the new $b_{ij}(\mathbf{O}_t)$ does not satisfy the stochastic restrictions).

This new separate formulation will be used in the paper and the two new parameters for each state and each model will be estimated by means of the discriminative training algorithm revised in next section.

3. DISCRIMINATIVE TRAINING ALGORITHM

Recently, a new minimum recognition error formulation and a generalized probabilistic descent (GPD) algorithm have been successfully proposed to perform discriminative training of speech recognizers [5]. The versions of both techniques used in this work will be briefly revised in this section and they will be applied in section 4 to estimate the new parameters γ^s_{ij}.

3.1. Minimum Recognition Error Rate Formulation

Given a training set $\{\mathbf{X}^i\}$, where each sample $\mathbf{X}^i$ is known to belong to one of M classes C^i, i = 1, 2, ..., M, the minimum error algorithm must find the system parameters set Λ and the corresponding decision rule that minimizes the likelihood of misclassification of any $\mathbf{X}$. In this work, each sample $\mathbf{X}$ is the sequence of feature vectors corresponding to one utterance (one digit in the recognition experiments reported in section 5). The objective criterion for optimal recognizers design can be derived as follows.

Firstly, an appropriate discriminant function $g_i(\mathbf{X},\Lambda)$ is chosen to implement the following decision rule

$$C(\mathbf{X}) = C^i, \quad \text{if} \quad g_i(\mathbf{X},\Lambda) = \max_m g_m(\mathbf{X},\Lambda) \qquad (4)$$

The discriminant function $g_i(\mathbf{X},\Lambda)$ used in this work is the logarithm of the score of the utterance $\mathbf{X}$ along its optimal path in i-th model, i.e.

$$g_i(\mathbf{X},\Lambda) = \log \pi_{iq(1)} + \sum_{t=1}^{T(\mathbf{X})} \log a_{iq(t-1)q(t)} + \sum_{t=1}^{T(\mathbf{X})} \log b_{iq(t)}(\mathbf{O}_t) \qquad (5)$$

where q(t) is the corresponding state sequence along the optimal path of the utterance **X** in the -i-th model, $\mathbf{O}_t$ is the set of features at time t of the utterance **X**, T(**X**) is the number of frames of the utterance **X**, $\pi_{iq(1)}$ is the initial probability of the state q(1) of the i-th model, and $a_{iq(t-1)q(t)}$ is the state transition probability from state q(t-1) to q(t) of the i-th model.

Then a misclassification measure is introduced to embed the decision process in a function form. In this work, if $C(\mathbf{X}) = C^i$ the following formulation is used

$$d_i(\mathbf{X},\Lambda) = -g_i(\mathbf{X},\Lambda) + \log \left\{ \frac{1}{M-1} \sum_{m,m\neq i} \exp(\eta\, g_m(\mathbf{X},\Lambda)) \right\}^{\frac{1}{\eta}} \qquad (6)$$

where $\eta > 0$. Otherwise, the opposite function is used.

Finally, a loss function is defined. In this work, the following expression is used

$$l_i(\mathbf{X},\Lambda) = \frac{1}{1+\exp(-\alpha d_i(\mathbf{X},\Lambda))} \qquad (7)$$

where α is a positive constant.

That formulation allows to directly minimize the expected recognition error by gradient descent search methods.

3.2. Generalized Probabilistic Descent Algorithm

The expected loss defined as

$$L(\Lambda) = E\{l_i(\mathbf{X},\Lambda)\} \qquad (8)$$

is usually considered the performance indicator.

Given a set of labeled training utterances, it is possible to minimize L by adaptively adjusting Λ in response to the incurred cost each time a training utterance is evaluated. The Generalized Probabilistic Descent (GPD) algorithm adjusts the parameters of the i-th model Λ_i recursively according to

$$\Lambda_i(n+1) = \Lambda_i(n) - \varepsilon(n)\, U(n)\, \nabla l_i(\mathbf{X}^n,\Lambda(n)) \qquad (9)$$

where $\Lambda_i(n)$ is the set of parameters of the i-th model in n-th iteration, $\varepsilon(n)$ is a sequence of positive step size parameters, U(n) is a sequence of positive definite matrices, ∇ is the gradient operator, $\mathbf{X}^n$ is the n-th training utterance and $\Lambda(n)$ is the set of parameters of the system in n-th iteration.

In this work an iteration is made for each training utterance (the training set is presented randomly), the matrices U(n) are taken to be equal to the identity matrix and the step size parameters satisfy the expression

$$\varepsilon(n) = \varepsilon_o \left(1 - \frac{n}{N} \right) \qquad (10)$$

where ε_o and N are positive numbers.

4. DISCRIMINATIVE TRAINING OF THE WEIGHTS

The minimum recognition error formulation and the generalized probabilistic descent (GPD) algorithm revised above can be applied to perform discriminative training of the new parameter γ^s_{ij} -the exponential weight for the s-th features of the j-th state of the i-th model- introduced in section 1 in order to ponderate static and dynamic features.

For the case of the new parameter γ^s_{ij}, expression (9) becomes

$$\gamma^s_{ij}(n+1) = \gamma^s_{ij}(n) - \varepsilon(n) \frac{\delta l_i(\mathbf{X}^n, \Lambda(n))}{\delta \gamma^s_{ij}(n)} \qquad (11)$$

where δ denotes partial differenciation and it has been supposed U(n) to be equal to the identity matrix. So it is necessary to compute the derivative of the loss function $l_i(\mathbf{X}^n, \Lambda(n))$ with respect to $\gamma^s_{ij}(n)$ (from this point, the dependence on n of the expressions involved in that computation will not be denoted explicitly).

Applying the chain rule of derivation to expression (7), the derivative of the loss function $l_i(\mathbf{X}, \Lambda))$ with respect to γ^s_{ij} can be expressed in terms of the derivative of the misclassification measure $d_i(\mathbf{X}, \Lambda)$ with respect to γ^s_{ij} as

$$\frac{\delta l_i(\mathbf{X}, \Lambda)}{\delta \gamma^s_{ij}} = \alpha \, l_i(\mathbf{X}, \Lambda) \left[1 - l_i(\mathbf{X}, \Lambda) \right] \frac{\delta d_i(\mathbf{X}, \Lambda)}{\delta \gamma^s_{ij}} \qquad (12)$$

Furthermore, from expression (6), taking into account that the discriminant function of the i-th model $g_i(\mathbf{X}, \Lambda)$ used in this work (5) does not depend on the parameters of the other models, the derivative of the misclassification measure $d_i(\mathbf{X}, \Lambda)$ with respect to γ^s_{ij} is

$$\frac{\delta d_i(\mathbf{X},\Lambda)}{\delta \gamma^s_{ij}} = - \frac{dg_i(\mathbf{X},\Lambda)}{d\gamma^s_{ij}} \qquad (13)$$

if utterance **X** corresponds to the i-th model.

Due to the fact that only the terms $b_{iq(t)}(\mathbf{O}_t)$ of the discriminant function $g_i(\mathbf{X},\Lambda)$ (5) depend on the exponential weights γ^s_{ij}, the derivative (13) is equal to

$$\frac{\delta d_i(\mathbf{X},\Lambda)}{\delta \gamma^s_{ij}} = - \sum_{t=1,q(t)=j}^{T(\mathbf{X})} \frac{\delta \log b_{ij}(\mathbf{O}_t)}{\delta \gamma^s_{ij}} \qquad (14)$$

From (3), $b_{ij}(\mathbf{O}_t)$ can be written as

$$b_{ij}(\mathbf{O}_t) = \prod_{s=1}^{2} b^s_{ij}(\mathbf{O}^s_t)^{\gamma^s_{ij}} \qquad (15)$$

where

$$b^s_{ij}(\mathbf{O}^s_t) = \sum_{k=1}^{L} c^s_{ijk}\, N(\mathbf{O}^s_t, \mu^s_{ijk}, \Sigma^s_{ijk}) \qquad (16)$$

and $\mathbf{O}^s_t$ is the vector of features -static ones for s=1 or dynamic ones for s=2- at time t of the utterance **X**. So

$$\log b_{ij}(\mathbf{O}_t) = \sum_{s=1}^{2} \gamma^s_{ij} \log b^s_{ij}(\mathbf{O}^s_t) \qquad (17)$$

and

$$\frac{\delta \log b_{ij}(\mathbf{O}_t)}{\delta \gamma^s_{ij}} = \log b^s_{ij}(\mathbf{O}^s_t) \qquad (18)$$

Finally, from expressions (11), (12), (14) and (18), the following recursion for the estimation of the parameter γ^s_{ij} results

$$\gamma^s_{ij}(n+1) = \gamma^s_{ij}(n) + \varepsilon(n)\ \alpha\ l_i(\mathbf{X}^n,\Lambda(n))\ \left[1 - l_i(\mathbf{X}^n,\Lambda(n))\right] \sum_{t=1,q(t)=j}^{T(\mathbf{X})} \log b^s_{ij}(\mathbf{O}^s_t{}^n) \qquad (19)$$

where $\mathbf{O}^s_t{}^n$ is the vector of static or dynamic features at time t of utterance $\mathbf{X}^n$. When utterance $\mathbf{X}^n$ does not correspond to the i-th model, that expression is valid just changing the sign of $\varepsilon(n)$.

5. RECOGNITION EXPERIMENTS

This section reports experimental results in speaker independent digit recognition that show the important increase in recognition accuracy obtained by means of discriminate training of the speech features weights

5.1. Database and Recognition System

The database used in the recognition experiments consists of 20 repetitions of the English digits ("zero", "one", "two", ..., "nine", "oh") corresponding to the adult speakers (112 for training and 113 for testing) of the speaker independent digit TI [6] database, that have been considered recorded in clean conditions. The initial sampling frequency 20 kHz was converted to 8 kHz.

The HTK [7] recognition system, based on the Continuous-Density Hidden Markov Models (CDHMM), was appropriately modified to perform the discriminative 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 12 cepstral parameters obtained either by linear prediction (LPC), with prediction order equal to 10, or by the mel-cepstrum technique. In some tests the energy of the frame was also used. Regression analysis over 70 ms was applied to the static cepstral sequence and the static energy sequence to obtain dynamic features, delta-cepstrum and delta-energy, 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. Training was performed in two stages using Segmental k-means, with previous manual endpointing, and Baum-Welch algorithms. Testing was performed using Viterbi algorithm.

5.2. Experimental Results

In table 1 the recognition rates obtained using LPC or mel-cepstrum analysis with and without energy E are presented. These results were obtained setting $\eta = 2$ (in expression 6), $\alpha = 0.02$ (in expression 7) and $N = 25000$ (in expression 10). All the weights γ^s_{ij} were inicialized with the value used in the baseline system, i.e. 1. The value of ε_0 was optimized for each parameterization and it can bee seen in the second column of the table. Due to the fact that all the utterances of the training set were used to train all the models, that value was multiplied by 10 when the training utterance was the digit corresponding to the model being trained.

Parameterization	ε_0	Baseline errors	DW errors	% Error reduction
LPC-cepstrum	0.0005	32	26	19
Mel-cepstrum	0.002	31	18	42
LPC-cepstrum +E	0.001	27	19	30
Mel-cepstrum +E	0.005	28	12	57

Table 1

The errors obtained using the baseline system, i.e. with $\gamma^s_{ij} = 1$, the errors obtained using the discriminative weighting (DW) method proposed in this paper and the achieved reduction error are shown in the third, forth and fifth columns of Table 1, respectively. As it can be seen the reduction error obtained whit that method is very important.

6. CONCLUSIONS

In this paper a method to automatically estimate the optimum ponderation of static and dynamic features in a speech recognition system has been proposed. The approach consists basically in 1) adding two new parameters for each state of each model that weight both kinds of speech features, and 2) estimating those parameters by means a recently proposed discriminative training algorithm, the Generalized Probabilistic Descent (GPD) method. The reduction error obtained by means of this approach has been very important.

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] L. R. Rabiner, B.H. Juang, "An Introduction to Hidden Markov Models", IEEE ASSP Magazine, vol. 3, nº 1, pp. 4-16, 1986.

[5] B.H. Juang, S. Katagiri, "Discriminative Learning for Minimum Error Classification", IEEE Trans. ASSP, vol. 40, No 12, pp. 3043-3054, 1992.

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

[7] *HTK - Hidden Markov Model Toolkit v1.5*, Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc., December 1993.