ENHANCED ASR BY ACOUSTIC FEATURE FILTERING*

Chris J. Wellekens
Institut Eurécom - Sophia Antipolis - France[†]
(Christian.Wellekens@eurecom.fr)

ABSTRACT

Several recent results demonstrate improvement of recognition scores if some FIR filtering is applied on the trajectories of feature vectors. This paper presents a new approach where the characteristics of filters are trained together with the HMM parameters resulting in improvements of the recognition in first tests.Reestimation formulas for the cut-off frequencies of ideal LP-filters are derived as well for the impulse response coefficients of a general FIR LP-filter.

1. INTRODUCTION

Automatic speech recognition (ASR) relies on the comparison of utterances with sub-unit models. The most popular models are trained Hidden Markov Models (HMM). But, useful information is particularly concealed in acoustic waveforms and before starting any model training or recognition, an important issue is to extract pertinent features. Pitch frequency and phase are usually discarded. Some kind of harmonic analysis (LPC, cepstral, smoothed spectrum, filter banks) turns the recognition into the frequency domain. The non-stationarity of speech enforces analysis over time windows. Window length is typically 30ms shifted in time by 10ms: these values are consistent with the articulatory apparatus time constants. Very soon [1-2], the dynamics i.e. the evolution in time of the features has been recognized as crucial. It is now of common use to append each feature vector with the derivative (speed) and second derivative (acceleration) of feature coefficients. The dimension of the representation space is thus strongly increased and training of models will require more data and CPU time. Predictive HMM where the distance between a state and a feature vector is defined in terms of LPC prediction coefficients have also been proposed; thus taking account of one or several previous samples, a.o. [3-4]. However, the independancy of the added features is questionable and research has been conducted to reduce feature space dimension. Principal component analysis (PCA) has been suggested to reduce the dimension by keeping independent components only. It works on the whole training set regardless of the class appartenance of the feature vectors. Linear discriminant analysis (LDA) has thus been proposed long time ago and more recently in [5-6] making use of a priori knowledge of phonetic segmentation to increase the inter-class discrimination.

Computation of speed and acceleration vectors relies on numerical estimation based on neighboring vectors. As a consequence, information associated with one frame (10ms) is related to a wider time window. All these numerical evaluations can be viewed as a filtering process of feature vector trajectories. Recently, discriminative coefficients obtained using LDA have been considered as filtered coefficients [5]: actual filters resulting from this discriminant analysis have been analyzed and their behavior is related to derivators yielding speed and acceleration.

The aim of this paper is to derive the reestimation formulas for joint training of data filter cut-off frequencies and HMM model parameters in order to increase the likelihood of the training database.

In section 2, the expression of the filter applied to the feature vectors is discussed while section 3 is devoted to the training criterion.

Filter parameters are trained together with the HMM parameters (section 4). As a consequence, not only model parameters but also data are modified during the training. It could be alledged that preprocessing the data in order to increase the likelihood is not the best way to increase the recognition score: indeed, including the cut-off frequencies in the training is nothing else but a way to transform the data in order they fit better with Markov modeling. But we may consider the filtering as part of the model and any attempt to use the prediction error or a combination of features like speed and acceleration goes in the same direction. A common filter can be applied to all feature vector entries and to all states. More generally, different filters can be used for each entry but are similar for all

*This research is an unpublished contribution to the Large Vocabulary Speech Recognition Summer Workshop 1997, CLSP, The Johns Hopkins University, Baltimore. Special thanks are due to H.Hermansky for fruitful discussions

[†]Institut Eurécom is supported by Ascom, Cegetel, France Telecom, Hitachi, IBM, Motorola, Swisscom,Texas Instruments, Thomson CSF

states. Last, filters can be assumed different for all states and all entries.

Feature filtering contributes also to the reduction of the data flow: indeed, filtered data streams can be downsampled so that in the recognition phase, less feature vectors have to be processed per second. More specifically,

1. in the training:
   - filtering at the baseline rate (100Hz) but Viterbi alignment at a downsampled rate.
2. in recognition:
   - HMM trained at the baseline frequency (100 Hz) with a fixed cut-off frequency and recognition on downsampled vectors (according to the fixed cut-off frequency)
   - HMM trained on downsampled vectors (according to the fixed frequency)

2.FILTER DESCRIPTION

The specifications of a filter may be in time or frequency domain. The optimal frequency specifications are unkown and they may be considered as a result for this joint training. There is a possibility to describe the filter by its impulse response sample. In that case, no immediate control of frequency specification is left. In section 5, formulas are given showing how joint training of HMM and filter coefficients can be achieved. However the number of parameters in impulse response training is equal for each filter to the length of this filter. In the following, we better assume the filter is an ideal LP-filter so that respectively one cut-off frequency only is required for its complete description.

The simplest filter depending on a single parameter is the ideal low-pass filter. We use its truncated impulse response which is

$$h_p(\omega_u) = \frac{\sin(\omega_u p)}{\pi p} \qquad p \in [-P, \ldots, P]$$

where ω_u is the cut-off frequency of a LP-filter of length $2P+1$.

If cut-off frequency equals the Nyquist frequency (i.e. 50 Hz for a frame rate of 10 ms), the impulse response has only one non-zero sample at time $k=0$ and this corresponds to no filtering. Similar effect can be obtained if the filter is of length 1 ($P=0$).

In the sequel, filtered acoustic vectors will be denoted $x^t = (\xi_1, \ldots, \xi_d)$ while original vectors are denoted $z^t = (\zeta_1, \ldots, \zeta_d)$. So ζ_j is the j-th entry of the acoustic feature vector z and denoting $\zeta^{(p)}$ the entry of a vector $z^{(p)}$ located p frames after z,the filtered version of ζ is

$$\xi_j = \sum_{p=-P}^{P} h_p(\omega_u)\zeta_j^{(p)}. \tag{1}$$

This is the expression of zero-phase non-causal filtering. All entries of all vectors will be modified using this formula and HMM training will make use of the transformed vectors. An important issue is the normalization of the response. The power of the response is

$$\mathcal{P} = \sum_{p=-P}^{P} h_p^2$$

and depends of course on the cut-off frequency ω_u. Replacing h_p by $h_p/\sqrt{\mathcal{P}}$ yields the power-normalized impulse response.

In case of band-pass filtering, it is easy to notice that the truncated impulse response is the difference between the impulse responses of two LP-filters with respectively the lower (ω_l) and the upper-cut-off (ω_u) frequencies and depend of course of these two parameters only:

$$\begin{aligned} h_p &= \frac{\sin(\omega_u p) - \sin(\omega_l p)}{\pi p} \\ &= \frac{2}{\pi p}\cos(\sigma p)\sin(\delta p) \qquad p \in [-P, \ldots, P] \end{aligned}$$

with the central frequency

$$\sigma = \frac{\omega_u + \omega_l}{2}$$

and the bandwidth

$$2\delta = \omega_u - \omega_l.$$

Highpass filters are obtained as bandpass filters with a upper cutoff frequency equal to the Nyquist frequency ($\omega_u = \pi$).

3. HMM AND THEIR TRAINING ALGORITHMS

Viterbi algorithm is used in this paper for training as well as for recognition. The best path yields a partition of the data base such that each feature vector is associated with a given state. The likelihood of the training set is then

$$L = \prod_j \prod_{x \in \mathcal{Q}_j} p(x|q_j) P_t$$

where P_t is the product of all transition probabilities occurring in the best path; $\mathcal{Q}_j$ is the set of vectors associated with state q_j; the product over j runs over the set of all independent states of the models and $p(x|q_j)$ is probability density function (pdf) associated with state q_j.

In this paper, we restrict our approach to monogaussian pdf's (μ_i and Σ_i denote mean vector and covariance matrix of state q_i) since our aim is to study feasability.

In case all Σ matrices are diagonal, the log-likelihood Λ i.e $-\log(L)$ is

$$\begin{aligned}\Lambda &= 1/2\sum_{j}\sum_{x\in\mathcal{Q}_j}\sum_{k=0}^{d}\left(\frac{\xi_k-\mu_{jk}}{\sigma_{jk}}\right)^2 \\ &\quad+\sum_{j}\frac{n_j}{2}\log((2\pi)^d|\Sigma_j|)-\log(P_t) \qquad (2)\end{aligned}$$

where n_j is the number of vectors in $\mathcal{Q}_j$.

The contribution of P_t is independent of the contribution of states and can be discarded without loss of generality. The estimates of m_j and Σ_j obtained by cancelling the derivatives of Λ respectively versus m_j and Σ_j are

$$\hat{m}_j=\frac{1}{n_j}\sum_{x\in\mathcal{Q}_j}x \qquad (3)$$

and

$$\hat{\Sigma}_j=\frac{1}{n_j}\sum_{x\in\mathcal{Q}_j}(x-m_j)(x-m_j)^t. \qquad (4)$$

It is important to notice that if all feature vectors are multiplied by a common factor K, the $|\Sigma_j|$'s are multiplied by K^2 turning to additional terms in Λ: this shows that Λ is scale dependent. The best way to avoid scale dependancy is to constrain $\mathcal{P}=1$ and thus to modify the log-likelihood with a Lagrange term. Then the optimality condition discussed in section 4 depends on the Lagrange multiplier so that on ω_u too. However as it will be seen in the next section, neither ω_u nor the Lagrange multiplier can be explicitly found but result from an iterative process. To remedy this inconvenient, the cut-off frequency is computed without constraint regardless of the value of its power. However, to avoid irrelevant decay of Λ due to this gain, the filter impulse response is power renormalized at each training iteration.

4. REESTIMATION OF THE CUT-OFF FREQUENCIES

To derive a reestimation formula for the cut-off frequencies ω_u and ω_l, we cancel the derivatives of Λ versus these variables. The reestimation formulas are nonlinear. In the case of bandpass filters, a set of two nonlinear equations has to be solved. For the sake of conciseness, we consider here an LP-filter only where a single nonlinear equation gives the optimal cut-off frequency.

Clearly, all vectors depend on the cut-off frequency via equation (1). As a consequence, mean vectors and covariances depend also on them via equations (3)(4). Parameters at the k-th iteration are $m_j^{<k>}, \Sigma_j^{<k>}, \omega_u^{<k>}$.

In a conventional training where no filtering is applied, the differential of Λ

$$d\Lambda=\sum_j\left(\frac{\partial\Lambda}{\partial m_j}dm_j+\frac{\partial\Lambda}{\partial\Sigma_j}d\Sigma_j\right)$$

should be zero. This is obtained by cancelling all partial derivatives and formulas (3)(4) apply.

The total derivative of Λ versus ω_u is

$$\begin{aligned}\frac{d\Lambda}{d\omega_u} &= \sum_j\left(\frac{\partial\Lambda}{\partial m_j}\,\frac{dm_j}{d\omega_u}+\frac{\partial\Lambda}{\partial\Sigma_j}\,\frac{d\Sigma_j}{d\omega_u}\right) \\ &\quad+\sum_{all\ \ x}\frac{\partial\Lambda}{\partial x}\,\frac{dx}{d\omega_u} \qquad (5)\end{aligned}$$

Since the partial derivatives versus m's and Σ's in the bracketed terms of eq. (5) vanish due to the specific choice of the new estimators (eqs (3)-(4)), the total derivative of Λ versus ω_u will vanish if

$$\sum_{all\ \ x}\frac{\partial\Lambda}{\partial x}\,\frac{dx}{d\omega_u}=0. \qquad (6)$$

Using (1) and (2), this expression becomes under the assumption that all Σ_j's are diagonal matrices

$$\sum_j\sum_{\xi\in\mathcal{Q}_j}\sum_{k=0}^{d}\frac{\xi_k-\mu_{jk}}{\sigma_{jk}^2}\frac{d\xi_k}{d\omega_u}=0 \qquad (7)$$

or making use of (1),

$$\begin{aligned}&\sum_{p=-P}^{P}\sum_{q=-P}^{P}A_{pq}\cos(\omega_u p)\frac{\sin(\omega_u q)}{\pi q}= \\ &\qquad\sum_{p=-P}^{P}A_p\cos(\omega_u p) \qquad (8)\end{aligned}$$

where

$$A_p=\sum_j\sum_{\xi\in\mathcal{Q}_j}\sum_{k=0}^{d}\frac{1}{\sigma_{jk}^2}\mu_{jk}\zeta_k^{(p)}$$

and

$$A_{pq}=\sum_j\sum_{\xi\in\mathcal{Q}_j}\sum_{k=0}^{d}\frac{1}{\sigma_{jk}^2}\zeta_k^{(p)}\zeta_k^{(q)}.$$

It is easy to check that $A_{pq}=A_{qp}$. Coefficients A_{pq} and A_p contain the statistics collected during the backtracking of the optimal path in all training sentences.

By expliciting the special case where $q = 0$, one obtains:

$$\frac{\omega_u}{\pi}\sum_{p=-P}^{P} A_{p0}\cos(\omega_u p) = \sum_{p=-P}^{P} A_p \cos(\omega_u p) - \sum_{p=-P}^{P}\ \sum_{q=-P;q\neq 0}^{P} A_{pq}\cos(\omega_u p)\frac{\sin(\omega_u q)}{\pi q} \qquad (9)$$

which is a fixed point formulation $\omega_u = f(\omega_u)$. This fixed point equation could be iteratively solved $\omega_u^{<k+1>} = f(\omega_u^{<k>})$ at each iteration of the Viterbi process. The impulse response will simply be power normalized at each iteration to take the constrain into account. Multiple solutions exist and the sign of the second derivative of Λ should be checked to garantee a minimum.

One may argue the solution should lay in the range $[-\pi, \pi]$. However, ω_u is only used in formula (1). Clearly h_p is a periodic function of ω_u which is thus defined modulo 2π.

To increase discrimination between between states of phonemes, different features could be used to compute the emission probability associated with a state or with the states of a phone. The transformation of the features can be seen as part of the HMM description and leads to dedicated definitions of the local probabilities just as well as the dedicated gaussian pdf's do. Here we assume different filters for each state and for each feature vector entry. Cut-off frequencies are now denoted ω_{kju}. Again eq.(7) is crucial and becomes:

$$\sum_{\xi\in\mathcal{Q}_j} \frac{\xi_k - \mu_{jk}}{\sigma_{jk}^2}\frac{d\xi_k}{d\omega_{kju}} = 0. \qquad (10)$$

The definition of parameters A is now:

$$A_{pjk} = \sum_{\xi\in\mathcal{Q}_j} \frac{1}{\sigma_{jk}^2}\mu_{jk}\zeta_k^{(p)}$$

and

$$A_{pqjk} = \sum_{\xi\in\mathcal{Q}_j} \frac{1}{\sigma_{jk}^2}\zeta_k^{(p)}\zeta_k^{(q)}.$$

The number of cut-off frequencies to estimate is dS where S denotes the number of different states. If desired, it is straightforward to force all states of a same phone to have the same cut-off frequencies.

Let us assume more generally the searched filter has an impulse response $h_t \quad \forall t \in [-P, \dots, P]$ and that these coefficients should be optimized. As in (5) and (6) we come out with the condition:

$$\sum_{all\ \ x} \frac{\partial\Lambda}{\partial x}\ \frac{dx}{dh_t} = 0 \quad \forall t \in [-P, \dots, P] \qquad (11)$$

where the entries of x are still defined as in (1) but where the h-coefficients are the free parameters and no longer depend on cut-off frequencies. Using (1), (11) becomes:

$$\sum_{p=-P}^{P} h_p \sum_{\xi\in\mathcal{Q}_j} \zeta_k^{(t)}\zeta_k^{(p)} = \mu \sum_{\xi\in\mathcal{Q}_j} \zeta_k^{(t)} \quad \forall t \in [-P, \dots, P].$$

This expression should be valid for all t and the solution of this linear set yields the optimal filter. The number of free parameters is then $(2P + 1)dS$.

It is worth noticing that the power constraint discussed in section 3 can be taken explicity into account here since the coefficients result from a linear set of equations.

6. EXPERIMENTS

Preliminary experiments on a small speaker dependent data base shows that a fixed cut-off frequency (20Hz) filter of length 21 leads to 35% error rate against 37% for an unfiltered recognizer in a phoneme recognition task without grammar. This score drops to 33% if 50% dowsampling is applied and to 30% if phoneme entrance penalties are tuned. Trained filters have not yield significative results since the size of the data base was too small. Experiments are going on on TIMIT and will be reported at the conference.

7. REFERENCES

1. S.Furui, "Speaker independent isolated word recognizer using dynamic features of speech spectrum," *IEEE Trans. ASSP,*,vol.34, nr 1, pp.52-59, 1986.

2. C.J. Wellekens, "Explicit Time Correlation in Hidden Markov Models for Speech Recognition," *Proc. ICASSP-87*, vol. 1, pp. 384-386, Dallas, April 1987.

3. E.Levin, " Speech recognition using hidden control neural network architecture," *Proc. ICASSP-90*, pp. 433-436, Albuquerque (NM), 1990

4. F.Freitag, *An Application of Predictive Neural Networks to Speech recognition*, Tesi Doctoral, UPC, Barcelona, May 1998

5. S.van Vuuren, H.Hermansky,"Data-driven design of RASTA-like Filters", *Proc.Eurospeech1997*, pp.409-412, Rhodes, Greece,1997

6. N.Kumar, A.G.Andreou,"Generalization of Linear Discriminant Analysis in the Maximum Likelihood Framework", *Proc. Joint Statistical Meeting*, Chicago, August 1996.