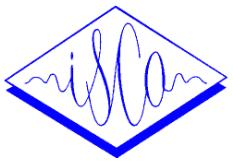

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPEAKER DEPENDENT TEMPORAL CONSTRAINTS COMBINED WITH SPEAKER INDEPENDENT HMM FOR SPEECH RECOGNITION IN NOISE

Néstor Becerra Yoma

Dept. of Electrical Engineering/University of Chile

Av. Tupper 2007, P.O.Box 412-3, Santiago, CHILE

nbecerra@cec.uchile.cl

ABSTRACT

This paper addresses the problem of speech recognition in noise using speaker-dependent temporal constraints in the Viterbi algorithm in combination with speaker-independent HMM. It is shown that the speaker-dependent re-estimation of state duration parameters requires a low computational load and a small training database, and can lead to reductions in the error rate as high as 30% or 40% with clean signals and with signals corrupted by additive noise, without noise canceling methods. Moreover, the approach here covered could also be seen as a speaker adaptation method in which only temporal restrictions parameters are adapted.

I. INTRODUCTION

In a previous paper [1], where state duration modeling was tested in speech recognition in noise, speaker-dependent (isolated or connected word) experiments suggested that: temporal constraints can lead to high reductions in the error rate with signals corrupted by additive or convolutional noise; and the accurate statistical modeling of state duration (e.g. with gamma probability distribution) does not seem to be very relevant if *max* and *min* state duration restrictions are imposed. However, state duration modeling did not lead to a high improvement in a speaker-independent (SI) connected word task which suggests that the introduction of temporal constraints in the Viterbi algorithm should be more useful when the state duration parameters are trained and employed on a speaker-dependent (SD) basis (although the HMMs could still be SI). In this paper, SD temporal constraints are trained by means of SI HMM using a small adaptation database. It is shown that the SD re-estimation of state duration parameter requires a low computational load and a small trainingdatabase, and can lead to reductions in the error rate as high as 30% or 40% with clean signals and with signals corrupted by additive noise. The approach here covered could also be seen as a speaker adaptation method in which only temporal restrictions parameters are adapted.

The contributions of this paper concern: a) the combination of SD temporal restrictions with SI HMM for speech recognition with clean and noisy signals; b) a re-estimation method for state duration parameters; and c) the comparison of the truncated gamma and geometric probability distributions for SD temporal restrictions with SI HMM. The approach covered by this paper has not been found in the literature and seems to be generic, and interesting from the practical application point of view because state duration modeling does not need any information about the testing environment.

II. TEMPORAL RESTRICTIONS IN THE VITERBI ALIGNMENT

Given the topology shown in Fig.1, in [1] the temporal restrictions were included in the Viterbi algorithm by means of replacing the ordinary transition prababilities with:

$$a_{i,i}^{\boldsymbol{t}} = \begin{cases} 1 & \text{if} \quad \boldsymbol{t} < t_{min_i} \\ 0 & \text{if} \quad \boldsymbol{t} > t_{max_i} \\ \dfrac{D_i(\boldsymbol{t}) - d_i(\boldsymbol{t})}{D_i(\boldsymbol{t})} & otherwise \end{cases} \quad (1)$$

$$a_{i,i+1}^{\boldsymbol{t}} = \begin{cases} 0 & \text{if} \quad \boldsymbol{t} < t_{min_i} \\ 1 & \text{if} \quad \boldsymbol{t} > t_{max_i} \\ \dfrac{d_i(\boldsymbol{t})}{D_i(\boldsymbol{t})} & otherwise \end{cases} \quad (2)$$

where: τ is the number of frames in state i up to time t; $t_{min_i} = tol\_min \cdot min_i(t)$ and $t_{max_i} = tol\_max \cdot max_i(t)$; $min_i(t)$ and $max_i(t)$ are the possible *min* and *max* durations, respectively; the constants *tol_min* and *tol_max* introduce a tolerance to the min and max duration for every state; $d_i(\tau)$ is the probability of state duration equal to τ; and $D_i(\boldsymbol{t}) = \sum_{t=\boldsymbol{t}}^{\infty} d_i(t)$. Equations (1) and (2) correspond to the truncated conditional probability

$$a_{i,j}^{\boldsymbol{t}} = \text{Prob}(s_{t+1} = j \mid s_t = s_{t-1} = \ldots = s_{t-\boldsymbol{t}+1} = i)$$

where $j=i$ or $j=i+1$ according to the topology shown in Fig.1.

2.1. Geometric and Gamma probability functions

If the geometric distribution is used, $\dfrac{D_i(\boldsymbol{t}) - d_i(\boldsymbol{t})}{D_i(\boldsymbol{t})}$ in (1) coincides with $a_{i,i}$, and $\dfrac{d_i(\boldsymbol{t})}{D_i(\boldsymbol{t})}$ in (2) with $a_{i,i+1}$, where $a_{i,i}$ and $a_{i,i+1}$ are the ordinary transition probabilities estimated during the HMM training algorithm. However, the gamma function better fits the empirical state duration distributions and $d_i(\tau)$ can also be modeled as a gamma probability function whose parameters are estimated with the mean $(E_i(t))$ and the variance $(Var_i(\boldsymbol{t}))$ of state durations. The state duration parameters $(E_i(\boldsymbol{t}), Var_i(\boldsymbol{t}), max_i(\boldsymbol{t})$ and $min_i(\boldsymbol{t}))$ are computed for every state in each model by means of estimating the optimal state sequence for training utterances using the Viterbi algorithm after the HMMs have been trained.

2.2. Speaker-independent and speaker-dependent temporal restrictions

In [1], speaker-independent (SI) experiments used temporal parameters that were estimated with all the training speakers, and speaker-dependent (SD) tests were achieved with HMMs and temporal restrictions trained and tested with only one speaker. In this paper, the HMM are trained on a SI basis although the temporal parameters are SD.

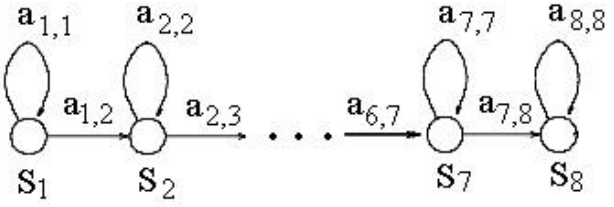

Figure 1: Eight-state left-to-right HMM without skip-state transition.

III. SD TEMPORAL RESTRICTIONS AND SI HMM

The results here presented were achieved using the Tidigits database from LDC [3]. Tidigits contains speech which was originally designed and collected for the purpose of designing and evaluating algorithms for SI recognition of connected digit sequences. There are 326 speakers (111 men, 114 women, 50 boys and 51 girls) each pronouncing 77 digit sequences. Each speaker group is partitioned into test and training subsets. Eleven digits were used: "zero", "one", "two", ... , "nine", and "oh". Seventy-seven sequences of these digits were collected from each speaker, and consisted of the following types: 22 isolated digits (two tokens of each of the eleven digits) 11 two-digit sequences; 11 three-digit sequences; 11 four-digit sequences 11 five-digit sequences; and 11 seven-digit sequences. As a consequence, the vocabulary is composed of 11 words, and a left-to-right HMM

containing 8 emitting state (Fig. 1) represent each word. The SI HMMs were estimated with 1100 three-digit sequences provided by 100 training speakers (50 males and 50 females). SI temporal parameters ($E_i(t)$, $Var_i(t)$, $max_i(\boldsymbol{t})$ and $min_i(\boldsymbol{t})$) were estimated using the training 4-digit utterances after the HMMs had been trained by means of Viterbi alignment.

The testing database was composed of 440 three-digit sequences provided by 40 testing speakers (20 males and 20 females). In order to compute the SD state duration parameters, 440 four-digit sequences from the testing database were employed. These utterances were used to compute the state duration parameters ($E_i(\boldsymbol{t})$, $Var_i(\boldsymbol{t})$, $max_i(\boldsymbol{t})$ and $min_i(\boldsymbol{t})$) by means of Viterbi alignment on a SD basis: the HMMs were SI but the testing utterances and the ones used to estimate the temporal parameters belonged to the same speaker.

3.1 Temporal restrictions for the first and last states

It was observed that the highest coarticulation effect in state duration takes place in the first and last states, and that the coarticulation effect was not proportionally represented by WPI (Word Position Independent) [1] temporal restrictions because these parameters were estimated with four-digit sequences and tested with three-digit utterances. Moreover, the error introduced by the end-point detection algorithm should also be mainly restricted to the first and last states. As a consequence, the state duration parameters should not be re-estimated for those states, where the best approximation would be the speaker independent temporal restrictions.

3.2 Estimation of SD temporal parameters

The state duration parameters were computed according to:

Estimation Algorithm

If number of adaptation samples per word >= N and if state is neither the first nor the last one,

$E_i(t)$, $Var_i(t)$, $max_i(\boldsymbol{t})$ *and* $min_i(\boldsymbol{t})$ *are speaker dependent and estimated with the adaptation utterances.*

Else,

$E_i(t)$, $Var_i(t)$, $max_i(\boldsymbol{t})$ *and* $min_i(\boldsymbol{t})$ *are speaker independent (estimated with the training database).*

where *N* denotes the minimum number of samples per word to be used in the state duration parameter estimation.

IV. EXPERIMENTS

The proposed method was tested with SI connected digit recognition experiments. The signals were downsampled to 8000 samples/sec and divided in 25ms frames with 12.5ms overlapping. Each frame was processed with a Hamming window before the spectral estimation and the band from 300 to 3400 Hz was covered with 14 Mel DFT filters. At the output of each channel the energy was computed and the log of the energy was estimated. Ten static cepstral coefficients and their time derivatives were estimated. Each word was modeled with an 8-state left-to-right topology (see Fig.1) without skip-state transition, with a single multivariate Gaussian density per state and a diagonal covariance matrix. The HMMs were estimated by means of the clean signal utterances. $E_i(\boldsymbol{t})$, $Var_i(\boldsymbol{t})$, $max_i(\boldsymbol{t})$ and $min_i(\boldsymbol{t})$ were estimated after the HMMs had been trained by means of Viterbi alignment. The constants *tol_min* and *tol_max* were made equal to 0.8 and 1.5, respectively. In some cases it was observed that the variation in state duration was equal to zero and a threshold was introduced to set a floor for $Var_i(\boldsymbol{t})$. The 440 testing clean three-digit utterances were used to create the noisy database by adding car and speech noise from the Noisex database [3] at 4 global-SNR levels: +18dB, +12dB, +6dB, and 0dB. The global-SNR was defined as in [4].

Three experiments were done: the ordinary Viterbi algorithm, *Vit*; the Viterbi algorithm with *max* and *min* state duration plus state duration distribution with gamma function, *Gamma*; and finally the

Viterbi algorithm with *max* and *min* state duration plus the ordinary geometric distribution, *Geom.* The state duration parameters were WPI [1]. Tables 1, 2 and 3 show the results with clean speech, and speech corrupted by car and speech noise, where *N* denotes the minimum number of samples per word used in the state duration parameter estimation (section 3.2). The recognition error rate was computed as $\frac{S+D+I}{W}\cdot 100$ where *S, D,* and *I* are the number of substitution, deletion and insertion errors, respectively, and *W* is the total number of words in the testing utterances.

Table 1: Recognition error rate (%) in experiments with clean speech. The ordinary Viterbi algorithm gives a recognition error equal to 6.21.

N	SI	5	4	3	2
Gamma	5 .8	5.1	4.9	4.1	4.4
Geom	6.2	6.0	5.5	4.9	5.0

Table 2: Recognition error rate (%) in experiments with speech corrupted by car noise.

	Gamma				*Geom*			
	SNR (dB)				*SNR(dB)*			
N	18	12	6	0	18	12	6	0
∞	8.0	10.2	14.8	24.9	9.2	12.1	17.5	27.2
5	7.2	9.1	13.4	24.0	8.1	10.8	16.7	26.4
4	6.4	8.2	12.7	23.2	7.4	10.1	15.3	24.9
3	5.9	7.3	12.1	22.6	6.4	8.9	13.5	23.6
2	5.2	6.4	10.8	21.4	5.6	7.6	12.0	23.2
Vit	9.2	12.1	17.5	27.3	9.2	12.1	17.5	27.3

IV. DISCUSSION AND CONCLUSION

The results presented in this paper suggest that the SD re-estimation of state duration parameters requires a low computational load and a small training database, and can lead to reductions in the error rate as high as 30% or 40% with clean signals and with signals corrupted by additive noise. According to Tables 1,2 and 3, the highest improvements in the error rate were achieved with clean speech and at SNR=18 and 12dB: 34%, 40% and 38%, respectively, using the truncated gamma distribution. At lower SNRs the state duration modeling did not lead to important improvements. It is worth mentioning that no noise canceling method (e.g. spectral subtraction) was used. When compared with the geometric truncated distribution, the gamma distribution gave better results and it seems that as few as two or three samples per word would be needed to estimate the state duration parameters. Finally, the approach here covered seems to be generic and applicable to more complex tasks.

Table 3: Recognition error rate (%) in experiments for speech corrupted by speech noise.

	Gamma				*Geom*			
	SNR (dB)				*SNR(dB)*			
N	18	12	6	0	18	12	6	0
∞	10.1	17.2	41.4	76.3	12.4	19.5	42.1	75.5
5	9.5	16.2	40.2	75.5	11.7	17.9	40.3	74.7
4	8.8	15.5	38.9	73.9	10.5	16.9	39.7	73.3
3	8.0	14.7	37.1	73.3	9.0	16.1	37.7	73.2
2	7.7	13.7	37.6	72.9	8.5	15.1	37.4	73.0
Vit	12.4	19.5	42.1	76.2	12.4	19.5	42.1	76.2

VI. REFERENCES

[1] N.B. Yoma et al. "On including temporal constraints in viterbi alignment for speech recognition in noise". Accepted for publication in IEEE Trans. on Speech and Audio Processing.

[2] Linguistic Data Consortium, University of Pennsylvania, http://www.ldc.upenn.edu.

[3] A.Varga, H.J.M.Steeneken, M.Tomlinson, and D.Jones. *The Noisex-92 study on the effect of additive noise in automatic speech recognition.* Technical report, DRA, UK, 1992.

[4] O.Ghitza. *Robustness against noise: the role of timing-synchrony measurement.* Proc. ICASSP´87, pages 2372-2375, 1987.