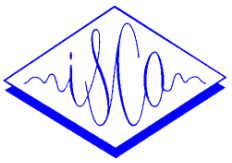


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

SPEAKER VERIFICATION IN NOISE USING TEMPORAL CONSTRAINTS

Néstor Becerra Yoma(*), Tarciano Facco Pegoraro(**)

(*) Dept. of Electrical Eng./University of Chile, Av. Tupper 2007, P.O.Box 412-3, Santiago, CHILE

(**)Ericsson do Brasil, Maria Prestes Maia 300, Sao Paulo-SP, BRAZIL

nbecerra@cec.uchile.cl

ABSTRACT

This paper addresses the problem of state duration modeling in combination with spectral subtraction and Rasta filtering to cancel both additive and convolutional noise in a text-dependent speaker verification task. The results presented in this paper suggest that temporal constraints can lead to reductions of 30 and 14% in the error rates at SNR equal to 0 and 6dB, respectively, without noise canceling techniques. However, with noise canceling methods, temporal restrictions give a lower improvement. The results here shown propose that state duration modeling can be useful in those cases when the noise reduction is low.

I. INTRODUCTION

In [1] state duration modeling was tested in speech recognition in noise and it was concluded that in speaker-dependent (SD) tasks 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 led to a lower improvement in a speaker-independent (SI) connected word task which suggests that the introduction of temporal constraints in the Viterbi algorithm could be more useful when the state duration parameters are trained and employed on a SD basis (although the HMMs could still be SI), or in speaker verification systems. In [2] temporal constraints were included in the Viterbi alignments with a proposed penalization procedure in a text-dependent speaker verification task. It was shown that, in experiments with clean signals, temporal restrictions did not lead to any improvement.

However, with speech signals corrupted by additive noise, state duration modeling with spectral subtraction could led to reductions of 20 and 10% in the error rate at SNR equal to 0 and 6dB, respectively. The results presented also suggested that the lower the SNR, the higher the improvement. In this paper state duration modeling in combination with noise canceling methods (spectral subtraction, and Rasta) is tested in the context of text-dependent speaker verification with speech signal corrupted by convolutional and additive noise. The results here shown suggest that: temporal constraints according to [2] lead to significant reductions in the error rate depending on the noise cancelation techniques being employed; and the lower the noise canceling effectiveness, the higher the improvement due to state duration modeling. Finally the accurate statistical modeling of state duration does not seem to be very relevant if *max* and *min* state duration restrictions are imposed.

II. THE SPEAKER VERIFICATION SYSTEM

2.1. The database

The Viterbi algorithm with temporal constraints was tested on a text-dependent speaker verification system using the Yoho database [3]. The YOHO Speaker Verification Corpus (Yoho) supports development, training and testing of speaker verification systems that use limited vocabulary, free-text input. The vocabulary is composed of two-digit numbers ("thirty-four", "sixty-one", etc),

spoken continuously in sets of three (e.g. "62-31-53" or "sixty-two thirty-one fifty-three"). The database is divided into "enrollment" and "verification" segments; each segment contains data from all 138 speakers (108 males and 30 females). There are four enrollment sessions per speaker and each session contains 24 utterances. Each verification segment contains 10 sessions and each session contains four utterances per speaker.

2.2. The HMM representation

Each two-digit number can be decomposed in two words: a decade ("30", "40", "50", "60", "70", "80" or "90") and a digit ("1", "2", "3", "4", "5", "6", "7", "8", or "9"). As a consequence, the vocabulary is composed of 16 words, and each word is represented by a left-to-right HMM containing 8 emitting state (Fig. 1), with a single multivariate Gaussian density per state and a diagonal covariance matrix. The false-acceptation and false-rejection curves (needed to compute the Equal Error Rate-EER) were estimated with 97 speakers (77 males and 20 females) and the global HMMs [4], used in the likelihood normalization [5], was trained with 41 speakers (31 males and 10 females). All the HMMs were trained with the Baum-Welch algorithm. The speaker dependent HMMs (16 per speaker) were estimated using 96 (4 enrollment sessions and 24 utterances per session), and the 16 global HMMs (one per vocabulary word) were trained with 3936 utterances (41 speakers and 96 training utterances per speaker).

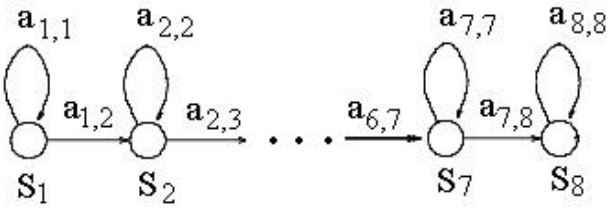
The false-rejection errors were estimated using the 40 verification utterances (10 sessions and four utterances per session) per speaker. The false-acceptation curve for a given speaker was computed with only one utterance per impostor. Each utterance (***O***) was processed with the Viterbi algorithm in order to estimate the normalized log likelihood (log *L(**O**)*):

$$\log L(\mathbf{O}) = \log P(\mathbf{O} / l_i) - \log P(\mathbf{O} / l_g) \qquad (1)$$

where $P(\mathbf{O} / l_i)$ is the likelihood related to the speaker *i*; and $P(\mathbf{O} / l_g)$ is the likelihood related to the global HMMs. Both models, l_i and lg, correspond to the sequence of word HMMs that compose the testing sequence **O** (text-dependent speaker verification). In order to estimate the false-rejection and false-acceptance error curves, the normalized log likelihood $\log L(O)$ was divided by the number of frames (*N*) in the verification utterance (after end-point detection):

$$\log L(\mathbf{O})' = \frac{\log L(\mathbf{O})}{N} \qquad (2)$$

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



III. TEMPORAL RESTRICTIONS

The temporal restrictions were included in the Viterbi algorithm according to the conditional transition probabilities proposed in [2]:

$$a_{i,i}^{\boldsymbol{t}} = \begin{cases} 1 & \text{if} \quad \boldsymbol{t} < t_{min_i} \\ min(a_{i,j}) & \text{if} \quad \boldsymbol{t} > t_{max_i} \\ \dfrac{D_i(\boldsymbol{t}) - d_i(\boldsymbol{t})}{D_i(\boldsymbol{t})} & otherwise \end{cases} \qquad (3)$$

$$a_{i,i+1}^{\boldsymbol{t}} = \begin{cases} min(a_{i,j}) & \text{if} \quad \boldsymbol{t} < t_{min_i} \\ 1 & \text{if} \quad \boldsymbol{t} > t_{max_i} \\ \dfrac{d_i(\boldsymbol{t})}{D_i(\boldsymbol{t})} & otherwise \end{cases} \qquad (4)$$

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 $D_i(t) = \sum_{t=\tau}^{\infty} d_i(t)$, and $min(a_{i,j})$ is a threshold that was empirically estimated and depends on the percentage of frames that are allowed not to comply with the minimum and maximum state durations. In this paper the probability function $d_i(\tau)$ is approximated with a gamma or with a geometric distribution. The state duration parameters ($E_i(t)$, $Var_i(t)$, $max_i(t)$ and $min_i(t)$) are computed for every state in each model by means of estimating the optimal state sequence for every training utterance using the Viterbi algorithm after the HMMs have been trained.

The estimation of $\log P(\mathbf{O}/\lambda_i)$ and $\log P(\mathbf{O}/\lambda_g)$ in (1) were made with the same SD temporal parameters (computed with SD HMMs) according to [2]. The temporal parameters ($E_i(t)$, $Var_i(t)$, $max_i(t)$ and $min_i(t)$) were WPI (Word Position Independent) [1]. As suggested in [2], there is not significant difference between WPI and WPD (Word Position Dependent) parameters for the task here considered.

IV. EXPERIMENTS

The proposed methods were tested with the text-dependent speaker verification system explained in section II. The signals were divided in 25ms frames with 10ms overlapping, each frame was processed with a Hamming window before the DFT spectral estimation (SS) [6] was applied to reduce the additive noise. The band from 300 to 3400 Hz was covered with 20 Mel DFT filters, the log of the energy was estimated, and 12 static cepstral coefficients and their time derivatives were computed. Besides the cepstral and Δ-cepstral parameters, the frame log energy (logE) and its time derivative (Δ-logE) were also estimated. Rasta [7] filtering were applied on the static cepstral parameters. 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 using the Baum-Welch algorithm. The state duration parameters ($E_i(\tau)$, $Var_i(\tau)$, $max_i(\tau)$ and $min_i(\tau)$) were estimated using the enrolment utterances after the HMMs had been trained by means of Viterbi alignment. 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 $Var_i(\tau)$. For each client (96 speakers), SD temporal parameters were computed. In contrast, SI state duration parameters were estimated with all the 41 impostors employed to train the global HMMs (speaker independent models). In experiments with additive noise, the testing clean utterances were used to create the noisy database by adding car noise from the Noisex database [8] at 4 global-SNR levels: 18, 12, 6, and 0dB. The convolutional noise was introduced with a FIR whose frequency response was approximately flat up to a break point frequency of 250Hz followed by a +6dB/oct tilt above 250Hz.

Table 1: Results with only convolutional noise.

Techniques	EER_{SS} (%)	EER_{SI} (%)
Vit	1,54	8,83
Vit / Rasta	0,50	1,12
Vit-P-Gamm / Rasta	0,52	1,19
Vit-P-Geom / Rasta	0,52	1,19

In order to test the validity of the state duration modeling from the text-dependent speaker verification point of view, the following configurations were tested with the noise canceling techniques: the ordinary Viterbi algorithm, *Vit*; and the penalization procedure according to (3)(4), with

tol_max=1.5, *tol_min*=0.8 and $\log\{min(a_{i,j})\}$=-10 [2], using the ordinary geometric distribution (*Vit-P-Geom*) and with the gamma function (*Vit-P-Gamm*). The methods here covered are compared employing *a posteriori* equal error rates (EER): EER_{SS}, using speaker specific thresholds; and EER_{SI}, with a speaker independent threshold. Results are shown in Tables 1, 2 and 3. Without noise, the baseline system gave EER_{SS} and EER_{SI} equal to 0.36 and 0.96%, respectively.

Table 2: Results with convolutional and additive (car) noise.

	EER_{SS} (%)			
Techniques	18dB	12dB	6dB	0dB
Vit	3.42	5.94	17.8	30.0
Vit / SS / Rasta	0.72	1.26	2.78	8.53
Vit-P-Gamm / SS / Rasta	0.78	1.35	2.65	7.57
Vit-P-Geom / SS / Rasta	0.77	1.36	2.67	7.66

Table 3: Results with only additive (car) noise.

	EER_{SS} (%)			
Techniques	18dB	12dB	6dB	0dB
Vit	0.68	1.80	6.34	22.9
Vit-P-Gamm	0.68	1.86	5.44	16.2
Vit-P-Geom	0.66	1.87	5.45	16.2

V. DISCUSSION AND CONCLUSION

As can be seen in Table 1, temporal constraints did not lead to reductions in the error rate when applied with Rasta to cancel convolutional noise. This must be due to the fact that Rasta considerably reduces the convolutional noise and this result suggests that if the testing data matches the training data, state duration modeling does not lead to significant improvements, which in turn is consistent with [2]. Comparing *Vit-P-Gamm* / SS / Rasta and *Vit-P-Geom* / SS / Rasta in Table 2, temporal restrictions did not give any significant improvement except at SNR=0dB where state duration modeling led to a reduction of 10% in the error rate. This should result of the fact that SS and Rasta together can cancel reasonably well both convolutional and additive noise at SNR equal to 18, 12 and 6%. However, Table 3 shows that state duration modeling can lead to reductions in the error rate of 30 and 14% at SNR equal to 0, 6dB, respectively, without any noise canceling method. Tables 1-3 also suggest that 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, which in turn confirms [2]. To conclude, the results here presented suggest that, in a text-dependent speaker verification task, the lower the noise canceling effectiveness (e.g. non-stationary environments), the higher the improvement due to state duration modeling.

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] N.B.Yoma, T.F:Pegoraro. "Robust speaker verification with state duration modeling". Submitted to Speech Communication.

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

[4] M. Carey and E. Parris. "Speaker verification using connected words". Proceedings on Institute of Acoustics, 14 (6), pp. 95-100, 1992.

[5] S. Furui. "Recent advances in speaker recognition". Pattern Recognition Letters 18, 859-872, 1997

[6] S.V. Vasegui and B.P. Milner. "Noise compensation methods for Hidden Markov Model

speech recognition in adverse environments". IEEE Transactions on SAP, 5 (1): 11-21, 1997.

[7] Hermansky, H.; Morgan, N.; 1994. RASTA Processing of Speech. IEEE Transactions on SAP, 2 (4), pp. 578-589.

[8] 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.