ROBUST CONNECTED WORD SPEECH RECOGNITION USING WEIGHTED VITERBI ALGORITHM AND CONTEXT-DEPENDENT TEMPORAL CONSTRAINTS

Nestor Becerra Yoma[1,2] *Lee Luan Ling*[1] *Sandra Dotto Stump*[2]

[1]*DECOM/FEEC/UNICAMP*
CP 6101, 13083-970 Campinas-SP, Brazil
nestor@decom.fee.unicamp.br
[2]*MACKENZIE UNIVERSITY*
Rua da Consolação 896, 01302-000 São Paulo-SP, Brazil

10.21437/Eurospeech.1999-635

ABSTRACT

This paper addresses the problem of connected word speech recognition with signals corrupted by additive and convolutional noise. Context-dependent temporal constraints are proposed and compared with the ordinary temporal restrictions, and used in combination with the weighted Viterbi algorithm which had been tested with isolated word recognition experiments in previous papers. Connected-word recognition tests show that the weighted Viterbi algorithm depends on the accuracy of the state duration modelling and the approach here covered can lead to reductions as high as 90 or 95% in the error rate at moderate SNR using spectral subtraction, an easily implemented technique, even with a poor estimation for noise and without using any information about the speaker. It is also shown that the weighting procedure can reduce the error rate when cepstral mean normalization is also used to cancel both additive and convolutional noise.

1. INTRODUCTION

In [1], a model for additive noise using DFT filters was proposed and used to compute the uncertainty or variance related to the spectral subtraction (SS) process to weight the Viterbi (HMM) algorithm. In [2] the weighted Viterbi algorithm was compared and used in combination with state duration modelling, and a weighting function without a free variable was proposed. The experiments in [1] [2] concerned isolated words and showed that weighting the information along the signal can substantially improve the recognition accuracy when the speech signal is corrupted by additive and convolutional noise, using spectral subtraction (SS) and cepstral mean normalization (CMN), two easily implemented techniques. Considering the simplicity of the techniques involved and testing conditions, the results presented in [1] [2] were very encouraging. However, preliminary experiments with connected digits showed that the weighted Viterbi algorithm depended on the accuracy of the state duration modelling and further work was needed to improve the applicability of the approach

The contributions of this paper concern: a)context-dependent (CD) state duration modelling; b) combination of the weighted Viterbi algorithm with temporal constraints for robust connected word recognition; c) comparison of contex-independent (CI) with context-dependent temporal constraints; d) improvement of SS combined with CMN to cancel additive and convolutional noise as far as connected word recognition is concerned. The approach covered by this paper has not been found in the literature, seems to be generic and interesting from the practical applications point of view, and is an important step toward robust feature extraction.

2. WEIGHTED VITERBI ALGORITHM

The weighting coefficient (based on the noise cancelling uncertainty variance [1] [2]) can be included in the Viterbi algorithm by raising the output probability of observing the frame T_t to the power of $w(t)$, where t is the time index [2]. In the experiments here reported, each word was modelled using a left-to-right topology (Fig. 1) with a single multivariate Gaussian density per state and a diagonal covariance matrix. The weighting function is defined as [2]

$$w = \frac{1}{N}\sum_{n=1}^{N}\frac{\sigma^2_{\lambda,i,n}}{\sigma^2_{\lambda,i,n}+Var[c_n|X]} \qquad (1)$$

where $\sigma^2_{\lambda,i,n}$ is the variance of coefficient n, state i and model λ. The function shown in (1) compares the uncertainty variance of coefficient n ($Var[c_n|X]$) [2] with the variance of the coefficient n in a phonetic class or state of a HMM.

The weight $w(t)$, which is between 0 and 1, tends to compress the range of variation of the output probability $b_i(T_t)$: if $w(t)$ is close to 0 (low reliability), $[b_i(T_t)]^{w(t)}$ will be close to 1 regardless of $b_i(T_t)$ and this probability loses discriminability. Consequently,

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

the importance of the transition probabilities a_{ij} to discriminate between two models increases in those segments where the local SNR is lower. However, the transition probabilities offer a low discriminative ability and are related to the modelling of state durations which is not well achieved using the geometric distribution of the ordinary HMM topology. In other words, the weighting procedure enhances the need of a more realistic state duration distribution. For the case of isolated digit [1] [2], the weighted Viterbi algorithm leads to a substantial improvement in the recognition accuracy with and without temporal constraints although the best results were achieved with state duration modelling [2]. This must result from the facts that the term concerning the product of transition probabilities does not present big changes for different state alignment and that the recognition tends always to rely on those frames with higher segmental SNR. However, for connected word recognition, the state and word alignment is used to decide the sequence of words and to achieve this it is necessary to segment properly the speech signal in terms of where one word should start and finish. Consequently, for connected word recognition the role of the temporal constraints seems to be dramatically more important than for the isolated case to make the weighted algorithm successful.

3. STATE DURATION MODELLING

A high improvement in the recognition accuracy was reported in [4], with signal corrupted by additive noise, and in [5], with clean speech, when temporal constraints were introduced in the training and in the testing procedure respectively. Including temporal constraints in the recognition process is conceptually simpler and generally requires a lower computational load. In [5] every state was associated to a gamma distribution whose parameters (mean and variance) were estimated by means of Viterbi alignment using the training database after the HMMs had been trained, and the state duration contribution was taken into consideration in the Viterbi algorithm using a proposed metric. In [2], every state was also associated to a gamma distribution whose parameters were estimated as in [5] although the state duration contribution was included with conditional transition probabilities. This approach was tested with isolated digits and proved effective in reducing the error rate when the signal was corrupted by additive noise. The experiments reported in this paper concern connected digits and a simple state duration model based on the max and min possible state durations was chosen to be used in combination with the weighted Viterbi algorithm. The parameters max and min were estimated as in [2] after the HMMs have been trained by means of the Viterbi alignment using the training database. To include the restrictions imposed by max_i and min_i, where i denotes the state (see topology shown in Fig. 1), the transition probabilities were modified:

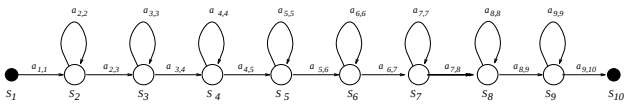

Figure 1. Eight-state left-to-right HMM without skip-state transition.

$$a^{\tau}_{i,i} = \begin{cases} 1 & \text{if } \tau < min_i \cdot K_{min} \\ 0 & \text{if } \tau \geq max_i \cdot K_{max} \\ a_{i,i} & \text{otherwise} \end{cases} \quad (2)$$

$$a^{\tau}_{i,i+1} = \begin{cases} 0 & \text{if } \tau < min_i \cdot K_{min} \\ 1 & \text{if } \tau \geq max_i \cdot K_{max} \\ a_{i,i+1} & \text{otherwise} \end{cases} \quad (3)$$

where $a_{i,i}$ and $a_{i,i+1}$ are transition probabilities estimated during the HMM training algorithm and τ denotes the state duration. The constants K_{min} and K_{max} introduce a tolerance to the min and max duration for every state.

4. CONTEXT-DEPENDENT STATE DURATION MODELLING

In HMM based speech recognition, word modelling (i.e. a HMM per each vocabulary word) is generally used for small vocabulary recognition systems (e.g. connected digits). Consequently, in a small vocabulary connected word recognition task each HMM attempts to capture the coarticulation effect between contiguous words indepently of the context (i.e. a word is represented by only one HMM). In [5] and [2], where word modelling was used, every state duration distribution was modelled with only one probability density which is reasonable for the isolated word task but is not very accurate for connected word recognition. In the latter case, the state duration (at least for the first and last states) depends on the context and this paper proposes that the word's position in the string carries information which should be used to improve the accuracy of temporal restrictions. Table 1 shows the mean duration of the first and last states, respectively, of the word "zero" HMM in triplets (i.e. three digit strings) when the word is string-initial, when it is between two other words and when it is string-final. As can be seen in Table 1, the context-independent state duration distribution

is not a precise model due to the fact that the mean state duration strongly depends on whether there is or not a "silence" interval before and/or after the word. The state duration model used in this paper in combination with the weighted Viterbi algorithm, as explained in section 3., concerns just the possible *min* and *max* state durations used by (2) and (3). For the task here considered, two set of state duration models were estimated: context-independent (CI) and context-dependent (CD). The CI model is composed by the overall min_i^O and max_i^O state durations independently of the word's position in the string. The CD model is composed by three set of *min* and *max* state durations: min_i^I and max_i^I, when the word is string-initial; min_i^M and max_i^M, when the word is preceded and followed by other words; and min_i^F and max_i^F, when the word is string-final.

Table 1. Mean duration of the first and last states of the word "zero" HMM according to the word's position in the string.

State	*string-initial*	*middle*	*string-final*
First state	15.0	5.3	6.2
Last state	2.0	2.4	10.1

5. EXPERIMENTS

The proposed methods were tested with speaker-dependent connected digits (English) recognition experiments. The tests were carried out employing the two speakers (one female and one male), and the car noise from the Noisex database [3]. The Noisex database has widely been used by several authors [6] [7] and allowes the comparison with results presented elsewhere. The signals were downsampled to 8000 samples/sec and were divided in 25ms frames with 12.5ms overlapping. The band from 300 to 3400 Hz was covered with 14 Mel DFT filters. At the output of each channel the energy was computed, SS [2] was applied and the log of the energy was estimated. In every frame 10 cepstral coefficients were computed. The noise estimation was made using just 200ms of non-speech signal. Where convolutional noise experiments were performed a +6dB/oct spectral tilt was applied to the signals and CMN was employed after SS. Each word was modelled using an 8-state left-to-right topology without skip-state transition (Fig. 1), with a single multivariate Gaussian density per state and a diagonal covariance matrix. The HMMs and the *max* and *min* state durations (sections 3. 4.) were estimated by means of the clean signal training utterances. The constants K_{min} and K_{max} were made equal to 0.8 and 1.5, respectively. The recognition error rate was computed as $\frac{S+D+I}{N} \cdot 100$ where S, D and I are the number of subtitutions, deletions and insertions errors, respectively, and N is the total number of words in the testing utterances.

Table 2. Recognition error rate (%) for speech signal corrupted by additive noise (car). The recognition experiments were done with SS: Vit, the ordinary Viterbi algorithm; $WVit$, the weighted Viterbi algorithm; $Vit-CI$ and $Vit-CD$, the ordinary Viterbi algorithm with context-independent and context-dependent temporal constraints, respectively; and $WVit-CI$ and $WVit-CD$, the weighted Viterbi algorithm with context-independent and context-dependent temporal constraints, respectively.

SNR	*18dB*	*12dB*	*6dB*	*0dB*
Vit	21.0	44.7	84.0	95.0
WVit	27.3	38.0	53.7	76.7
Vit-CI	11.0	30.7	55.4	79.3
Vit-CD	2.0	7.7	26.7	62.7
WVit-CI	18.4	27.4	40.4	60.7
WVit-CD	1.0	2.0	4.7	19.0

Table 3. Recognition error rate (%) for speech signal corrupted by convolutional noise (6dB/oct spectral tilt). The recognition experiments were done with CMN: Vit, the ordinary Viterbi algorithm; and $Vit-CI$ and $Vit-CD$, the ordinary Viterbi algorithm with context-independent and context-dependent temporal constraints, respectively.

SNR	*6dB/oct spectral tilt*
Vit	4.4
Vit-CI	2.7
Vit-CD	0.4

6. DISCUSSION AND CONCLUSION

Experiments with only *additive noise* (Table 2) showed that the weighted version of the Viterbi algorithm without temporal constraints did not lead to any improvement in the recognition accuracy which did not confirm the results presented in [2] where experiments with isolated words showed that the weighted Viterbi algorithm led to a high reduction in the error rate with and without temporal constraints. This must be due to the fact that, as explained in section 2., the weighting procedure enhances the need of a more realistic state duration distribution which is

Table 4. Recognition error rate (%) for speech signal corrupted by additive and convolutional noise. The recognition experiments were done with SS and CMN: Vit, the ordinary Viterbi algorithm; and $Vit-CD$ and $WVit-CD$, the ordinary and weighted Viterbi algorithms, respectively, with context-dependent temporal constraints.

SNR	*18dB*	*12dB*	*6dB*	*0dB*
Vit	8.0	22.0	60.4	90.7
Vit-CD	5.0	6.3	20.0	49.3
WVit-CD	3.7	4.0	5.7	23

much more critical for connected than isolated word recognition. Actually, as can be seen in Table 2, when CI temporal constraints are used in combination with the weighted Viterbi algorithm, the error rate presents reductions of 12, 39, 52 and 36% at SNR=18, 12, 6 and 0dB, respectively. However, the best results were achieved with the weighted Vitertbi algorithm in combination with CD temporal constraints which led to reductions as high as 95, 96, 94 and 80% in the error rate at SNR=18, 12, 6 and 0dB, respectively.

Experiments with only *convolutional noise* (Table 3) show that CD temporal restrictions gave a higher recognition accuracy than CI temporal constraints and almost a 100% reduction in the error rate when compared with the ordinary Viterbi algorithm. Tests with *additive and convolutional noise* (Table 4) indicate that the weighting procedure in combination with CD temporal restrictions can also be effective if CMN is applied after SS.

When compared with other techniques [6] [7], poorer error rates were observed with the weighted HMM for the car noise at SNR=0dB, although the experiments of this paper concern static parameters only and a poor estimation of the additive corrupting signal made in just 200ms and without using any information about the convolutional noise. Connected word recognition experiments showed that CD temporal constraints are much more effective than the ordinary CI temporal restrictions with the same computational load and, in combination with the weighted Viterbi algorithm, leads to a substantial reduction in the arror rate with easily implemented techniques (SS and CMN). The approach here covered should also be applicable to more complex tasks (e.g. medium and high vocabulary continuous speech recognition) that uses context-dependent phone HMMs which in turn results in context-dependent state duration modelling. The simplicity of the techniques involved and testing conditions make the method here presented interesting from the practical application point of view and further work is currently in progress to improve the applicability of the approach to cancel both additive and convolutional noises in speech recognition and speaker verification. Finally, it is worth mentioning that weighted matching algorithms could also be used with other noise cancelling methods.

7. ACKNOWLEDGMENT

Nestor Becerra Yoma was supported by a grant from FAPESP, Sao Paulo, Brasil.

REFERENCES

[1] N.B.Yoma, F.R.McInnes, M.A.Jack. *Improving performance of spectral subtraction in speech recognition using a model for additive noise.* IEEE Tans. Speech and Audio Processing, Vol.6, No6, 1998.

[2] N.B.Yoma, F.R.McInnes, M.A.Jack. *Weighted Viterbi algorithm and state duration modelling for speech recognition in noise.* Proc. ICASSP'98, pp.709-712.

[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 Speech Research Unit, U.K., 1992.

[4] K. Laurila. *Noise robust speech recognition with state duration constraints.* Proc. ICASSP 97, Vol.2, pp.871-874

[5] D.Burshtein. *Robust Parametric Modeling of Durations in Hidden Markov Models.* IEEE Transactions on Speech and Audio Process. V.4, No3, May, 1996.

[6] M.F Gales. *Model-Based Techniques for Noise Robust Speech Recognition.* PhD Thesis, Engineering Department, Cambridge University, Sept. 95.

[7] T.Claes,D.VanCompernolle. *SNR-Normalization for robust speech recognition.* Proc. ICASSP, pp 331-334, 1996.