ADDITIVE AND CONVOLUTIONAL NOISE CANCELING IN SPEAKER VERIFICATION USING A STOCHASTIC WEIGHTED VITERBI ALGORITHM

Néstor Becerra Yoma, Miguel Villar Fernandez

Dept. of Electrical Engineering/University of Chile

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

nbecerra@cec.uchile.cl

ABSTRACT

This paper replaces the ordinary output probability with its expected value if the addition of noise is modeled as a stochastic process, which in turn is merged with the HMM in the Viterbi algorithm. The method, which can be seen as a weighted matching algorithm, is applied in combination with spectral subtraction and RASTA to improve the robustness to additive and convolutional noise of a text-dependent speaker verification system. Reductions around 10% or 20% in the error rates and improvements as high as 30% or 50% in the stability of the decision thresholds are reported when the ordinary Viterbi algorithm is replaced with the weighted one. When compared with the baseline system, reductions of 70% or 80% are shown.

I. INTRODUCTION

Improving the robustness to noisy environments is among the most important problems that need to be solved in order to make speaker verification successful in real applications. In a previous paper (Yoma et al., 1998) a model for additive noise using DFT filters was proposed and used to estimate the uncertainty in noise canceling, which in turn was used to weight the Viterbi algorithm to take into consideration the segmental SNR in isolated word speech recognition. The model suggested that once the noise is added, an uncertainty is introduced and the original signal cannot be recovered with 100% accuracy, and the reliability (inverse of uncertainty) in noise canceling is dependent on the segmental SNR. In (Yoma and Villar, 2001) a weighted Viterbi algorithm was proposed by replacing the ordinary output probability with its expected value to incorporate the stochastic process related to the addition of noise. Reductions as high as 30% or 40% in the error rates and improvements of 50% in the stability of the decision thresholds were reported as results of experiments with additive noise. The contribution of this paper concern: a) the combination of the weighted Viterbi algorithm in combination with SS and RASTA to cancel additive and convolutional noise in speaker verification; and b) estimation of noise canceling variance in the RASTA filtering domain. The approach here proposed substantially improves the accuracy of the system and the stability of the decision thresholds, has not been found in the specialized literature, and can be considered an important step toward robust acoustic pattern recognition.

II. THE STOCHASTIC WEIGHTED VITERBI ALGORITHM (WMA)

As proposed in (Yoma et al., 1998) , the reliability of the information provided by a frame depends on the local SNR, and the weighted Viterbi algorithm uses this information in the recognition procedure. The output probability $b_s(O_t)$ is generally modeled with a mixture of Gaussians with diagonal covariance matrices (Huang et al. 1990):

$$b_s(O_t) = \sum_{g=1}^{G} p_g \cdot \prod_{n=1}^{N} (2\cdot\pi)^{-1/2} (Var_{g,s,n})^{-1/2} \cdot e^{-\frac{1}{2}\cdot\frac{(O_{t,n}-E_{g,s,n})^2}{Var_{g,s,n}}} \quad (1)$$

where g, s, n are the indices for the Gaussian components, the states and the coefficients, respectively; p_g is the weighting for each gaussian; $O_t = [O_{t,1}, O_{t,2} \ldots\ldots O_{t,N}]$ is the observation vector composed of static and delta ceptral coefficients for the period t; $E_{g,s,n}$ and $Var_{g,s,n}$ are the HMM mean and variance, respectively. According to (Yoma and Villar, 2001), as a consequence of a model for additive noise, a noisy observation vector O_t can be

considered as being a set of random variables with normal distributions defined by means ($E[O_{t,n}]$) and variances ($Var[O_{t,n}]$), in contrast to the ordinary HMM topology that assumes that O_t is composed of constants. To overcame this incompatibility the output probability $b_s(O_t)$ should be replaced with its expected value ($E[b_s(O_t)]$). Assuming that the coefficients $O_{t,n}$ are uncorrelated, the expected value of this output probability can be written as (Yoma and Villar, 2001):

$$E[b_s(O_t)] = \sum_{g=1}^{G} p_g \cdot \prod_{n=1}^{N} \frac{1}{\sqrt{2\cdot\pi\cdot Vtot_{g,s,n,t}}} \cdot e^{-\frac{1}{2}\cdot\frac{(E[O_{t,n}]-E_{g,s,n})^2}{Vtot_{g,s,n,t}}} \quad (2)$$

where $Vtot = Var_{g,s,n} + Var(O_{t,n})$; $Var(O_{t,n})$ is the uncertainty variance in the cepstral domain that is computed as a linear combination of the uncertainty variance in the log energy domain $Var\left[\log\left(\overline{s_m^2}(\phi)\right)\right]$. This variance can be considered inversely proportional to $\frac{SSE_m}{E\left[\overline{n_m^2}\right]}$, where SSE_m, $\overline{s_m^2}$, and $E\left[\overline{n_m^2}\right]$ are the spectral subtraction estimation, the energy of the clean speech and noise estimation at the output of the filter m, respectively; and ϕ is the phase difference between the noise and speech signals, which is considered as being a random variable uniformly distributed between $-\pi$ and π. Finally, $E[O_{t,n}]$ corresponds to the ordinary cepstral feature $O_{t,n}$.

Equation (2) represents an elegant and generic result, and deserves some comments. Firstly, the expression (2) means that the expected value of the output probability is also represented by a sum of Gaussian functions. Secondly, if $Var(O_{t,n}) \to 0$ (i.e. high SNR) $O_{t,n}$ can be considered as a constant and (2) is reduced to the ordinary output probability because $E[O_{t,n}] = O_{t,n}$. Finally, if $Var(O_{t,n})$ is high (i.e. low SNR) the expected value given by (2) tends to zero independently of the HMM parameters $E_{g,s,n}$ and $Var_{g,s,n}$, which means that the information provided by a noisy observation vector is not useful and has a low weight in the final decision procedure of accepting or rejecting a speaker.

III. SPECTRAL SUBTRACTION (SS) AND RASTA FILTERING

In order to cancel both additive and convolutional noise spectral subtraction (SS) and RASTA filtering were used in combination with the weighted Viterbi algorithm. The spectral subtraction estimation was defined as:

$$SSE_m = max\left\{\overline{x_m^2} - E\left[\overline{n_m^2}\right]; \beta\overline{x_m^2}\right\} \quad (3)$$

where $\overline{x_m^2}$ and $E\left[\overline{n_m^2}\right]$ are the noisy signal and noise estimation energies at the output of filter m, respectively, and β is an empiric coefficient. It is worth mentioning that (3) corresponds to a simplified version of a SS described in (Vaseghi and Milner1997). In RASTA (Hermansky et al., 1991) the temporal sequences of the static and delta cepstral coefficients are processed by the following filter,

$$H(z) = 0.1\cdot\frac{2 + z^{-1} - z^{-3} - 2\cdot z^{-4}}{z^{-4}\cdot(1 - 0.98z^{-1})} \quad (4)$$

which corresponds to the discrete equation given by,

$$y_n(t) = 0.2x_n(t+4) + 0.1x_n(t+3) - 0.1x_n(t+1) - 0.2x_n(t) + 0.98y_n(t-1) \quad (5)$$

where x_n and y_n denote the cepstral coefficient n (static or dynamic) before and after being processed by RASTA, respectively. The uncertainty on noise canceling variance in the RASTA domain, $Var(Y_n)$, was computed considering the frames as being uncorrelated, the same assumption made by HMM, and by recursively replacing $y_n(t-1)$ in (5) until $y_n(t)$ becomes a function of only x_n.

IV. THE SPEAKER VERIFICATION SYSTEM

The approach here proposed was tested on a text-dependent speaker verification system using the

Yoho database. The vocabulary is composed of 16 words, and each word is represented by a left-to-right HMM containing 8 emitting states, with a single multivariate Gaussian density per state and a diagonal covariance matrix. The false-acceptance and false-rejection curves (needed to compute the Equal Error Rate-EER) were estimated with 97 speakers and the global HMMs (Carey & Parris, 1992), used in the likelihood normalization (Furui, 1997), were trained with 41 speakers.

Each utterance (***O***) was processed with the Viterbi algorithm in order to estimate the normalized log likelihood (log *L(**O**)*):

$$\log L(\mathbf{O}) = \log P(\mathbf{O}/\lambda_i) - \log P(\mathbf{O}/\lambda_g) \qquad (6)$$

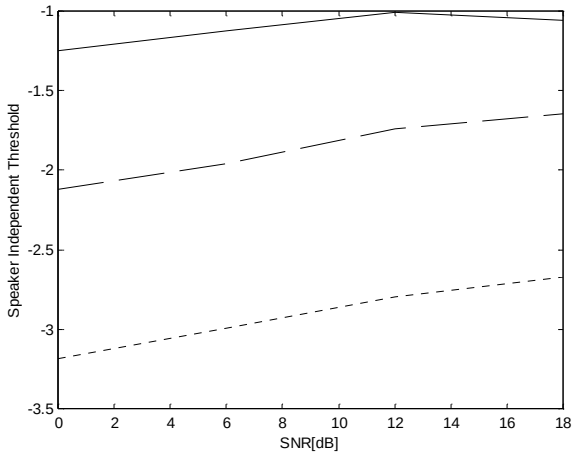
where $P(\mathbf{O}/\lambda_i)$ is the likelihood related to the speaker *i*; and $P(\mathbf{O}/\lambda_g)$ is the likelihood related to the global HMMs. Finally, the normalized log likelihood $\log L(O)$ was divided by the number of frames (*N*) in the verification utterance.

V. EXPERIMENTS

The proposed method was tested with the speaker verification system explained in section IV. The signals were divided into 25ms frames with 10ms overlap, and each frame was processed with a Hamming window before the DFT spectral estimation. The band from 300 to 3400 Hz was covered with 20 Mel DFT filters, spectral subtraction (SS) according to (3) was applied with $\beta = 0.25$, the log of the energy was estimated, and 12 static cepstral coefficients and their time derivatives were computed. The noise was estimated using only 10 non-speech frames. RASTA was applied after SS in the cepstral domain. Each word was modeled with an 8-state left-to-right topology without skip-state transition, with a single multivariate Gaussian density per state and a diagonal covariance matrix. The HMMs were trained by means of the clean signal utterances using the Baum-Welch algorithm. The verification clean utterances were used to create the noisy database by adding car noise from the Noisex database (Varga et al.1992). The convolutional noise is a 6dB/oct spectral tilt applied after the additive noise.

In the experiments whose results are here reported the techniques that were employed are indicated as follows: *Vit*, the ordinary Viterbi algorithm with output probability computed with (1); *WVit,* the weighted Viterbi algorithm with output probability estimated as $E[b_s(O_t)]$ according to (2); *SS*, spectral subtraction as in (3); and, *R* denotes RASTA filtering. To compute $E[b_s(O_t)]$ as defined in (2), $E(O_{t,n})$ and $Var(O_{t,n})$ were estimated as in (Yoma and Villar, 2001) without RASTA To include the effect due to this filtering, $E(O_{t,n})$ and $Var(O_{t,n})$ were applied to (5) according to section III. The methods here covered are compared using *a posteriori* equal error rates (EER): EER_{SS}, using speaker specific threshold ($TEER_{SS}$); and EER_{SI}, with speaker independent threshold ($TEER_{SI}$). Results with noisy speech are shown in Tables 1-2 and Fig. 1.

Figure 1: $TEER_{SI}$ vs. SNR(dB) with speech corrupted by additive (car) and convolutional noise: *WVit-R-SS* (____); *Vit-R-SS* (___ ___); *Vit* (_ _ _).



VI. DISCUSSIONS AND CONCLUSION

As can be seen in Table 1, experiments with speech signal corrupted by additve and convolutional noise show that the expected value of the output probability according to (2) combined with SS/RASTA led to reductions of 6%, 15% and 8% in the EER_{SS} at SNR=18dB, 12dB and 6dB, respectively, when compared with the ordinary Viterbi algorithm also with SS/RASTA. In the same conditions, the reductions in the EER_{SI} were 11%, 12% and 18% at, respectively, SNR=18dB, 12dB and 6dB as shown in Table 2. When

compared with the baseline system, the total improvement due to the weighted Viterbi algorithm plus SS/RASTA can be as high as 70% or 80% in EER_{SS} or EER_{SI} at SNR higher than 6dB. At SNR=0dB the reduction in the error rate is less significant.

As is shown in Fig. 1 and Table 3, the expected value of the output probability as defined in (2) substantially reduced the variability of $TEER_{SS}$ and $TEER_{SI}$. According to Table 3, the differences $TEER_{SS}$(18dB) - $TEER_{SS}$ (0dB) and $TEER_{SI}$(18dB) - $TEER_{SI}$(0dB) with SS/RASTA are, respectively, 32% and 50% lower with the weighted Viterbi algorithm than with the ordinary one. This must be due to the fact that, when the segmental SNR decreases, $Var(O_{t,n})$ increases and the output probability as defined in (2) tends to 0 for both the client and global HMM in (6).

Table 1: EER_{SS} vs. SNR (dB) with speech corrupted by additive and convolutional noise.

SNR	18dB	12dB	6dB	0dB
Vit	6.32	9.93	18.36	33.09
Vit-SS	5.52	8.48	14.04	25.58
Vit-R-SS	1.59	2.41	4.22	10.09
Wvit-R-SS	1.49	2.06	3.89	10.23

Table 2: EER_{SI} vs. SNR (dB) with speech corrupted by additive and convolutional noise.

SNR	18dB	12dB	6dB	0dB
Vit	19.59	25.15	34.15	42.80
Vit-SS	17.32	22.04	29.22	39.03
Vit-R-SS	3.21	4.20	7.44	16.73
Wvit-R-SS	2.86	3.68	6.12	15.01

The weighted matching method does not lead to an increase of the computational complexity. However, the estimation of the uncertainty variance of static and delta cepstral parameters after RASTA filtering requires a higher computational load, and further work is currently in progress to overcame this limitation. Finally, the results here presented confirm the applicability of the weighted Viterbi algorithm.

Table 3: TEER(18db) – TEER(0dB) with signal corrupted by additive and convolutional noise.

	Vit	*Vit-SS*	*Vit-R-SS*	*WVit-R-SS*
$TEER_{SS}$	0.85	1.05	0.44	0.30
$TEER_{SI}$	0.51	0.74	0.48	0.24

VII. ACKNOWLEDGEMENT

The research described in this paper was supported by a grant from Conicyt/Fondecyt-Chile.

VIII. REFERENCES

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

Furui, S. (1997). *Recent advances in speaker recognition*. Pattern Recognition Letters 18, pp. 859-872, 1997.

Hermansky, H. et al. (1991) *Compensation for the effect of the communication chennel in auditory-like analysis of speech (RASTA-PLP)*. Proc. Eurospeech 91, pp.1367-1370

Huang, X.D. et al. (1990). *Hidden Markov Models for speech recognition.* Edinburgh University Press, 1990.

Varga, A. et al. (1992*). The Noisex-92 study on the effect of additive noise in automatic speech recognition.* Technical report, DRA, UK, 1992.

Vaseghi, S.V. and Milner,B.P. (1997). *Noise compensation methods for Hidden Markov Model speech recognition in adverse environments*. IEEE Transactions on Speech and Audio Processing, 5 (1): 11-21, 1997.

Yoma, N.B. et al. (1998). *Improving performance of spectral subtraction in speech recognition using a model for additive noise.* IEEE Trans. on Speech and Audio Processing, Vol. 6, No.6, November, 1998.

Yoma, N.B. and Villar, M. (2001). *Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm*. Submitted to IEEE Transactions on SAP, February 2001.