Filtering the Time Sequence of Spectral Parameters for Speaker-Independent CDHMM Word Recognition

Climent NADEU, Pau PACHÈS-LEAL* and Biing-Hwang JUANG***

* Universitat Politècnica de Catalunya, Barcelona, Spain
** AT&T Bell Laboratories, Murray Hill, USA
E-mail: nadeu@tsc.upc.es

ABSTRACT

In this work, we show how speaker-independent CDHMM word recognition performance can be significantly improved for clean speech by filtering the time sequence of spectral parameters to enhance its time dynamics. Experimental results with the standard TI connected digits database show the filter can achieve more than 30% reduction of string recognition error. As shown in this paper, that improvement is partially due to the speaker variability reduction obtained by attenuating the very low modulation frequencies. The widely used cepstral mean subtraction technique also improves the recognition rate, but it can not achieve such a noticeable improvement as the parameter filter. In fact, the best results are obtained when the peak of the long-term spectrum of the filter output is at around 3 Hz, a frequency which corresponds to the average syllable rate of the employed database.

1. Introduction

The first step in the pattern matching approach to the problem of speech recognition is to convert frame-by-frame a speech waveform into a set of time sequences, one for each spectral parameter, where the time index indicates the frame number. Every time sequence of spectral parameters (TSSP) carries not only the phonetic content of the utterance but also some speaker characteristics, acoustic distortion and noise included in the speech signal, as well as the error of the spectral estimation process [1].

The TSSPs are filtered to remove or attenuate their very low frequency components that are contaminated by the quasi-invariant linear distortion due to microphone, telephone channel, etc. involved in the speech signal, provided that the TSSPs lie in the logarithmic domain [2-3]. However, a significant improve-ment has rarely been reported in recognition of clean speech by filtering the time sequences of parameters.

Recently, by using the long-term power spectrum of the TSSP, some insight into the filters applied to those parameter sequences was gained [1]. We will denote that spectrum by $T_m(q)$, where m is the parameter index and q, which has been referred to as the modulation frequency [4], is the frequency counterpart of the frame index n. As explained in [1], the parameter filters show two basic components. The first one is a differentiation, usually implemented with a zero at z=1, which approximately equalizes the spectrum of the TSSP $T_m(q)$, except the very low frequency region, where the spectral components are largely attenuated and the zero frequency component is removed. As the TSSP is viewed in this approach as a stationary random process, the equalization performed by that zero produces a certain decorrelation of the TSSP. Equivalently, that equalization can be interpreted as an augmentation of the time resolution (in the sense of a dynamics amplification) of the TSSP that enhances the discrimination capability of the speech classifier.

The second component of the parameter filters is a kind of smoothing that attenuates the likely unreliable higher frequency components of the TSSP spectrum that have been enhanced by the equalizing zero. For instance, the IIR (RASTA) filter used in [2] essentially consists of that equalizing zero at z=1 plus a pole whose magnitude is close to one, which controls the pass-band cutoff frequencies. Interestingly enough, those two components are also present in every filter that computes supplementary differential parameters (dynamic features) [1].

A large reduction (61%) of speaker-independent single digit recognition error was obtained with telephone speech only by properly filtering the time sequences of cepstral coefficients [1]. In the present work, we have applied the same kind of filtering to the TI digits database (Section 2). In spite of the fact that the data does not include variable linear distortion (since a unique microphone was used to collect it) an even larger error reduction than for the telephone database is achieved by filtering. In fact, as will be shown in Section 3, the modulation frequency region around zero is more affected by the speaker characteristics than by the phonetic content, and thus the attenuation of the TSSP spectrum in that region produces a higher recognition rate improvement in the TI digits speaker-independent task.

On the other hand, in Section 4, the widely used cepstral mean subtraction (CMS) technique is compared with a fixed-length CMS filter and the other filters used in Section 2. Although CMS improves the recognition results, it cannot achieve such a noticeable improvement as the parameter filters, because it depends on the utterance length and lacks a tuning parameter to shape the TSSP's spectrum.

This work has been partly funded by the Spanish Government project TIC92-1026-C02-02

2. Recognition Results Obtained by Filtering

We have carried out speech recognition experiments by filtering the TSSPs in several ways, and using the filtered TSSPs as the speech representation, with no addition of supplementary differential parameters. Training and testing were carried out with the single and connected digit utterances of the adult portion of the TI database [5]. After decimating the signals from 20 KHz to 8 KHz sampling rate and pre-emphasizing them with a zero at z=0.95, Hamming windowed frames of 30 ms were taken every 10 ms. Every frame was represented by 12 LPC cepstral parameters, obtained from a 10-order model, plus the energy normalized on the whole utterance. A speech recognition system based on continuous observation density hidden Markov models (CDHMM) was used (HTK software). Each of the digit models consisted of 8 (emitting) states, and the silence model had 3 states. Only one diagonal covariance Gaussian mixture was employed per state. As the testing conditions were equivalent to those in [1], we will be able to compare the recognition results obtained with the TI database and the telephone database used in that work.

The TSSPs, i.e. the time sequence of each cepstral coefficient $c_m(n)$ and that of the energy, were filtered for every utterance before performing training and testing. Therefore, the word models were re-trained after each new filtering of the TSSPs. As in [1], the filter consisted of the cascade combination of two filters: an equalizer $1-rz^{-1}$, and a Slepian filter, i.e. a filter whose impulse response is a discrete prolate spheroidal wave sequence. The former approximately equalizes $T(q)$, the long-term power spectrum of the TSSP averaged over all the cepstral coefficients. Its zero r was identified with the coefficient of the first-order linear predictor of the set of TSSPs obtained by means of mean-square estimation. Thus, $r=R(1)/R(0)$, where $R(n)$ is the inverse Fourier transform of $T(q)$. We used r=0.95, a value which results from a spectrum estimate obtained by averaging over the whole adult training set of TI digit utterances, and over the first 12 LPC-cepstral coefficients. This value produces a rather flat filtered TSSP's spectrum, like the value r=0.97 in [1].

The Slepian filter shows some optimal properties and has two control parameters: the impulse response length L and the bandwidth W. Following the design criteria described in [1], L=7 and W=16 Hz was found to be the optimum choice. The effective bandwidth is about 16 Hz, larger than for isolated digits (it was equal to 10 Hz in [1]), due to the higher speaking rate of connected digits. Also note that the filter length is smaller, a fact which helps to reduce the cross-boundary effect of filtering in connected word speech recognition.

	String	Word	Del	Subs	Ins
Unfiltered	20.34	7.68	2.02	4.86	0.41
Filtered	13.74	4.44	0.97	2.97	0.62

Table1 Percentage of recognition error rates for the unfiltered and the Slepian filtered parameters.

Table 1 shows the percentage of recognition error of the filtered TSSPs with respect to the unfiltered TSSPs by using the Slepian filter. A 32% reduction of string error and a 42% reduction of word error are achieved by filtering, and every performance measure is improved, except the insertion rate. The number of insertions is larger than for the unfiltered case due probably to the above mentioned cross-boundary effect.

An IIR filter (the typical RASTA filter) has been tested as well. Its numerator is $-2-z^{-1}+z^{-3}+2z^{-4}$ (zeros at q=0, 0.58π, and π), and the only (real) pole r has been empirically optimized. The best performance with one mixture corresponds to a pole at r=0.75, which yields 14.55% string error rate and 4.07% word error rate, results similar to those of the Slepian filter in Table 1.

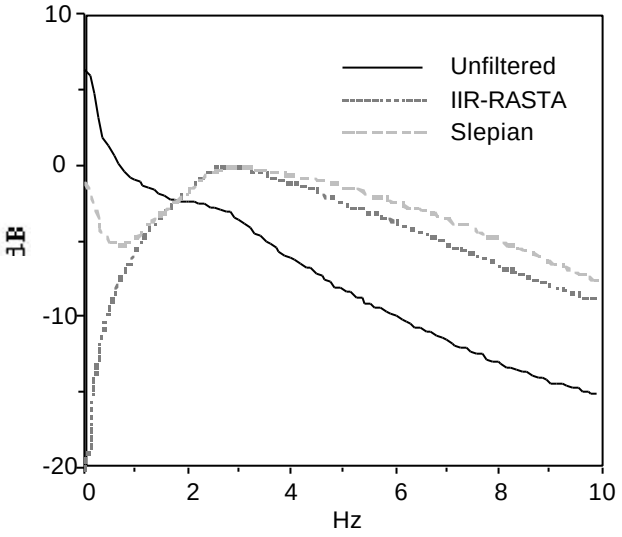

Fig.1. TSSP's spectra of unfiltered and filtered parameters estimated over all the training utterances of the adult TI digits database.

As can be seen in Fig. 1, there is a strong similarity between the high power bands of the spectra at the output of both optimized filters. This observation agrees with the role given in [1] to the TSSP's spectrum and its band equalization to interpret the parameter filtering process. Moreover, the spectral peak is located around 3 Hz, a result that agrees with previous experimental works from Houtgast and Steeneken [4] about the most important modulation frequencies for speech intelligibility in all the speech spectral bands. Actually, the 3 Hz modulation frequency reflects the syllable rate in speech; in the TI digits database, the syllable rate is 2.8 Hz.

When performing training and testing with only single digits, the improvement is much larger: 71% improvement with respect to the unfiltered case with an

optimum Slepian filter. Notice that this improvement obtained by filtering with single digits is even larger for the TI database than for the database employed in [1] (61%). This may be surprising since the latter was collected with two types of microphones and through the telephone network, so it has a different linear distortion for every speaker. Consequently, if the filter aimed only at cancelling linear distortion, it should give a higher improvement than for the TI database, where the linear distortion is constant for every speaker and utterance. However, the adult portion of the TI database contains 112 speakers for training and 113 for testing, a larger number than the database used in [1], which has 50 speakers for training and 22 for testing. In the next section we will show how the filter performs a reduction of speaker variability that accounts for this higher improvement.

3. Reducing the Speaker Variability

It is a well known fact that the long-term spectrum of speech signals is influenced by the speaker identity. As these long-term spectral characteristics are time-independent or slowly variant, they are included in the very low frequency interval of the TSSP's spectrum *T(q)*. To verify it in the TI digit database used in this work, we have carried out some variance measurements. For this purpose, we have used all the single digit utterances of the adult portion of the database. It consists of the string 'silence-digit-silence', where 'digit' refers to one element of the set {oh, zero, one,...,nine}. Every utterance is considered as a different class. Since the whole database was collected through the same microphone and acquisition system, the only sources of variation within a given class are the speaker voice characteristics and the dialectal diversity.

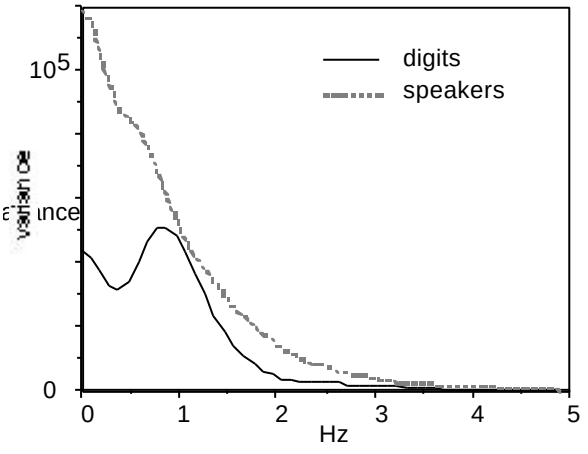

Fig.2. Inter-class (digit) variance and averaged intra-class (speaker) variance of the TSSP's spectra estimated over all the single digit utterances of the training adult TI database.

The intra-class variance, averaged over all the 11 digit classes, and the inter-class variance of the TSSP's spectrum are plotted in Fig. 2 between 0 and 5 Hz. Note that the most significant differences between them are located at the very low frequency interval. As can be observed in Fig. 2, the variance of the frequency components up to 1 Hz appear more influenced by the speaker voice than by the digit class, since the intra-class variance is clearly larger than the inter-class variance at those frequencies.

We have made yet another observation to gain more insight into the variance of *T(q)* between 0 and 1 Hz, and its relationship with the discrimination capability of the speech recognition system. By estimating separately the intra-class variances of the various digit utterances, we can correlate their relative values at the low frequency band with the confusion matrices of speech recognition tests. For this purpose, we defined a variance ratio for each digit as the quotient between the value of the variance of the digit at frequency zero and its average value over *q*, and we related it with the speech recognition errors that are eliminated by filtering the TSSPs with the Slepian filter presented in Section 2. Even though the estimated variances correspond to single digit utterances, the six digits with the largest variance ratios (six, eight, nine, one, oh, five) account for 93% of the eliminated errors (insertion, substitution and deletion) in the connected digit experiment reported in Table 1..

In the following, we will try to gain some insight into the influence of the TSSP's spectrum on the probability computation. Given a sequence of *N* iid observations O and a fixed-state sequence **q** in a continous observation Gaussian density HMM λ with one mixture per state, the conditional log probability of the observation sequence is

$$\log p(\mathbf{O} / \lambda, \mathbf{q}) = -\frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\log 2\pi s_m^2(n) - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\frac{|c_m(n)-\mu_m(n)|^2}{s_m^2(n)} \quad (1)$$

where $c_m(n), 1 \le n \le N$, is the time sequence of the *m*-th spectral parameter (which in our tests is the *m*-th cepstral coefficient), $1 \le m \le M$, and $\mu_m(n)$ and $s^2_m(n)$ are, respectively, the mean and variance of the *m*-th spectral parameter in the Gaussian density of the state corresponding to the *n*-th observation.

Applying the Parseval's theorem [6] to the time integration, and assuming that the variance of every parameter is independent of the state (grand variance), it follows that only the last term in (1) depends on the observation sequence and it is proportional to

$$-\sum_{m=1}^{M}\frac{1}{s_m^2}\int_{-p}^{p}|C_m(q)-M_m(q)|^2 dq \quad (2)$$

where $M_m(q)$ is the Fourier transform of $m_m(n)$, $1 \le n \le N$.

According to the last expression, the log probability can be computed in the modulation frequency domain. Given a sequence of states and, hence, given the function $M_m(q)$, the variable part of the log probability only depends on the Fourier transforms $C_m(q)$ of the M different TSSPs of the utterance. Since the average magnitude of those transforms has its largest values at low frequencies, these frequencies will generally dominate the probability computation. Consequently, the speech recognition system will get closer to speaker-independence by removing or highly attenuating with the filter the very low frequency components of the TSSP, which are more affected by the speaker characteristics than by the phonetic content.

4. Cepstral Mean Subtraction and Speaking Rate

Cepstral mean subtraction (CMS) [3] is a widely used technique to cancel linear distortion in speech recognition. It eliminates the zero frequency component of every TSSP by subtracting from each of its (frame) samples the average value in the utterance. So the whole utterance has to be available before performing CMS. CMS has rarely been used to increase the recognition rate in the case of clean speech. Nevertheless, a clear performance improvement (20%) was already reported in [7] with the TI digits database by using log spectral energies as parameters.

In our speech recognition tests with the whole adult portion of the TI digits database, CMS showed 17.66% string error rate and 6.49% word error rate, scoring between those of the unfiltered and Slepian filtered TSSP (see Table 1). Actually, CMS also performs a kind of filtering. To gain more insight into its behaviour, we note that it can be regarded as a circular convolution between the TSSP of length N and the sequence (assuming N odd)

$$h(n) = d(n) - \frac{1}{N}, \quad -\frac{N-1}{2} \le n \le \frac{N-1}{2} \qquad (3)$$

Since each utterance has a different length N, $h(n)$ is a variable length sequence, and thus the effect of the CMS on the continuous frequency components of the TSSP depends on N. This variable effect is stronger in the low frequency range between q_1=0 and q_2=2p/N, where the samples of the DFT of $h(n)$ go from value 0 to value 1, and it may become important for short utterances. (In the single digit case, the average length of the utterances is 96 frames, so q_2 is, approximately, 1 Hz). The variable effect implies a lower consistency of the trained word models and a larger variability of the CMS-processed TSSP of the test utterances.

A way of avoiding that dependency of the filtering process on the utterance length N consists of defining a filter whose impulse response has the same shape as $h(n)$ but with a fixed length M, independent of N and smaller than it, such as the one used in [8], but using both past and future samples. Herewith, we will refer to it as the fixed-length cepstral mean subtraction filter. Figure 3 plots the fixed-length CMS string error rate for the whole adult TI database. It appears that there exists an optimum value (33) of the filter length M. Note that the recognition score is almost like that of the Slepian filter and clearly better than that of the conventional CMS.

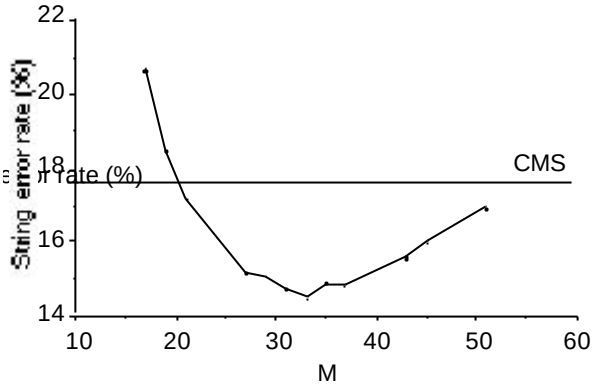

Fig. 3. String recognition error rate of fixed-length CMS for several values of the filter length M. The straight line indicates the CMS string error rate.

Also, the optimum M value is close to the average syllable length in the database, which is about 35 frames. Actually, the fixed-length CMS filter is a high-pass filter whose cut-off frequency is, approximately, 2π/M. Hence, due to the monotonically decreasing $T(q)$ curve, the optimal value M=33 yields a peak of the filtered TSSP spectrum at 3 Hz, which corresponds to the average syllable rate. According to that correspondence between the value of M and the average syllable length, the best M value for single digit utterances should be higher than for connected digits, since the speaking rate is lower. In fact, the largest improvement in single digit recognition by fixed-length CMS filtering is obtained for M=39. Note that, as regards the cross-boundary effect, the fact of using digit models allows a longer M value than if shorter speech units were modelled.

Indeed, the average value of N is much higher than the optimal fixed-length CMS value M=33, and thus the average spectrum of the CMS-processed TSSP has its peak at a modulation frequency lower than the optimum 3 Hz position. Furthermore, if the digit string is longer, N increases, while the optimal M value decreases, since the speaking rate is larger.

5. Conclusions

Speaker-independent CDHMM word recognition results can be noticeably improved for clean speech by filtering the TSSP. Experimental results with the TI connected digit database show the filter can achieve 71% reduction of string recognition error in the single digit case and more than 30% in the connected (single

and multiple) digit case. That improvement is at least partially due to the speaker variability reduction obtained by attenuating the very low modulation frequencies. Best results correspond to enhancing the modulation frequency band located around the average syllable rate 3 Hz, an observation that coincides with previous perception studies. Although CMS also improves the recognition rate, it can not achieve so noticeable an improvement as the parameter filter, because it depends on the utterance length and it lacks a tuning parameter to shape the TSSP's spectrum.

Acknowledgments

The authors wish to thank José B. Mariño for helpful discussions. They also like to express their gratitude to Manuel Toril for his assistance in the experimental work.

References

[1] C. Nadeu, B.H. Juang, "Filtering spectral parameters for speech recognition", *Proc. ICSLP'94*, pp. 1927-30.

[2] H. Hermansky, N. Morgan, “RASTA processing of speech”, *IEEE Trans. on SAP*, Vol. 2, No. 4, Oct. 1994.

[3] B.A. Hanson, T.H. Applebaum, J.C. Junqua, "Spectral dynamics for speech recognition under adverse conditions", in *Advanced Topics in Automatic Speech and Speaker Recognition*, C.-H. Lee, K.K. Paliwal and F.K. Soong, Eds., Kluwer Acad. Publ., 1995.

[4] T. Houtgast, H.J.M. Steeneken, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria”, *J.ASA*, Vol. 77 (3), March 1985, pp.1069-77.

[5] R.G. Leonard, "A database for speaker-independent digit recognition", *Proc. ICASSP'84*, pp. 42.11.1-4.

[6] A.V. Oppenheim, R.W. Schafer, *Discrete-Time Signal Processing*, Prentice-Hall, 1989.

[7] R. Haeb-Umbach, D. Geller, H. Ney, "Improvements in connected digit recognition using linear discriminant analysis and mixture densities", *Proc. ICASSP'93*, pp. 239-42.

[8] A.E. Rosenberg, C.-H. Lee, F.K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification", *Proc. ICSLP'94*, pp. 1835-38.