ON PARAMETER FILTERING IN CONTINUOUS SUBWORD-UNIT-BASED SPEECH RECOGNITION

Pau Pachès-Leal and Climent Nadeu

Universitat Politècnica de Catalunya
Barcelona, Spain
e-mail: paupag@gps.tsc.upc.es

ABSTRACT

Simple IIR or FIR filters have been widely used in isolated or connected word recognition tasks to filter the time sequence of speech spectral parameters, since, despite their simplicity, they significantly improve recognition performance. Those filters, when applied to continuous speech recognition, where phoneme-sized modelling units are used, induce spectral transition spreading and a cross-boundary effect. In this work, we show how the use of context-dependent units reduces the side effects of the filters and may result in improved recognition performance. When dynamic parameters are not used, filtering seems to be especially useful, even for clean speech, and when they are, filters do well under unmatched training and testing conditions.

1. INTRODUCTION

At present, IIR or FIR filtering of the time trajectories of speech spectral parameters is widely used since, albeit conceptually simple and amenable to a real time implementation, it allows a significant improvement in recognition performance.

In a logarithmic spectral domain, where a convolutive distortion becomes additive, filtering out the very low frequencies of the time trajectories helps to alleviate the linear distortion due to the slow-varying acoustic channel. However, most filters used tend to be bandpass, rather than highpass. Therefore, they can do something else than merely cancelling the DC component.

The study of the average long-term power spectrum [1] of the time sequences of spectral parameters, denoted by $T(\theta)$, where θ, often referred to as modulation frequency, is the frequency counterpart of the frame index n, shows that such filters, like those which are used to provide dynamic features or supplementary parameters, actually have two components: a differentiation, often implemented with a zero at or close to $z=1$, which attenuates the low frequency region and approximately equalizes the rest of the long-term spectrum, and a lowpass component, which discards the high frequency zone of the long-term spectrum, where the estimation error variance is greatest and which has been unduly enhanced by the first component [1].

With one feature, substitutive parameter filtering has also been shown to yield a substantial improvement for clean speech, although in this case the linear distortion due to the acoustic channel is fixed, through an enhancement of the time dynamics of the time trajectories of speech feature vectors and through a reduction of speaker variability [2].

However, those filters have most often been applied to isolated or connected word recognition tasks [2, 3], where the average word length is far higher than the effective length of the impulse response of the filters, since the latter induce a spreading of spectral transitions and make the current analysis output depend on its neighbouring context.

Such side effects are critical in continuous subword-unit-based speech recognizers, for which filters may worsen recognition performance [4]. Context-dependent subword units are frequently used in this framework since they can model speech coarticulation. In this work, we show how, as suggested by [5], the use of context-dependent units tackles the problem posed by the cross-boundary effect brought about by the filtering and may make the latter advantageous.

We apply several filters to the parameterised utterances of two continuous speech databases and for different sets of context-dependent units and number of features (one or two, that is to say, with or without addition of dynamic features or supplementary parameters). By doing so, in a number of cases, we get a twofold improvement in recognition performance: the one provided by the use of context-dependent units and the additional improvement supplied by filtering. This filtering is extraordinarily simple, and often helps to further reduce the recognition error rate significantly.

2. RESULTS WITH ONE FEATURE

Two different databases were used for phonetic classification experiments: the Spanish EUROM.1 database [6], hereafter referred to as DB1, and the Spanish SentencesUPV database (which was recorded by the Universitat Politècnica de València), we will call it DB2. Both are used for speaker-independent tests. The former comprises nearly 37000 PLU's (in 842 utterances, of which 186 are different) from 43 speakers for training, and around 12500 PLU's (in 225 utterances, 61 of them different) from 17 speakers for testing. The latter has 21667 PLU's (in 839 utterances, 120 being unique) from 7 speakers for training and 5610 PLU's (150 utterances, 50 different) from 3 speakers for testing. Both databases are parameterised with 12 Mel-frequency cepstral coefficients (MFCC) plus an energy coefficient. 25-ms analysis windows with a 10-ms analysis step were used. Each of the PLU's, as well as the silence model, is modelled by a three-state left-to-right continuous observation density HMM. The recognizer was the HMM toolkit (HTK), with three Gaussian mixtures per state. The context-dependent labelling scheme is the same as in [7]. As the number of context-dependent phones (CDP)

is too high, both in theory and in this particular database [7], and no statistically reliable model can be obtained for most of them, a threshold for the CDP's is defined, which sets the minimal number of training tokens for the corresponding CDP model to be trained. If a CDP is not frequent enough (i.e. if the number of times it appears in the training corpus text does not attain or surpass the threshold), then it is the corresponding CI (context-independent) model which is trained. The coverage rate for a set of context-dependent units is defined as the fraction of the (training or testing) corpus text that can be labelled using this set. In this work, we have only used right context dependent phones (RCP), triphones (TrP) or a combination of both. The training corpus coverage rate and the number of units for each of these sets is shown in Table I.

Set and threshold(s)	Number of models	Coverage rate
RCP 105	109	79.4%
RCP 70	139	86.3%
RCP 35	194	94.2%
TrP 70	126	43.7%
TrP 35	250	59.6%
TrP 70, RCP 70	225 (100T, 99 R)	43.7%, 84.7%
TrP 35, RCP 35	391 (224T, 141R)	59.6%, 92.9%

Table 1: Analysis of the different sets of CD units. All the sets include the 26 CI phones to attain a 100% coverage rate.

The filters we have used are (lowpass) Slepian filters [1] preceded by a (highpass) equalizing filter with transfer function $H(z)=1-rz^{-1}$, where r is close to 1. Slepian filters have two parameters, the length L and the bandwidth W, which if certain conditions hold, gives the upper cutoff frequency (bandwidth) of the filter. A Slepian filter of length L and bandwidth W is denoted as (L, W). We have also used the classical first-order derivative window, which amounts to a FIR filter of order 5, and refer to it as regression. Simple IIR filters, the so-called RASTA filters, with a single complex pole and a numerator identical to the regression filter, have also been tested. In this section, any of the three aforementioned filters substitutes for the original feature, with no addition of supplementary parameters.

2.1. Results with DB1 (clean speech)

Figure 1 shows the recognition results for DB1 and one feature. We see that filtering somewhat worsens the recognition results for CI phones. The further we move to the right, the higher the improvement of the error rate (informally defined as 100%-%accuracy) with respect to the non-filtered case. For CI phones, this improvement equals -0.97% (the best filter being the Slepian (5, 32)), whereas for the last experiment it is 11.4% (the regression filter is the best). Using CD phones allows us to have an 18.4% improvement in error rate (for the non-filtered case, and from the first to the last experiment) and filtering enables us to further reduce the error rate (by another 11.4%), although clean speech (acquired with a single microphone with little background noise) has been used. As was expected for continuous speech [1], filters of length greater than 5 were found to give worse results.

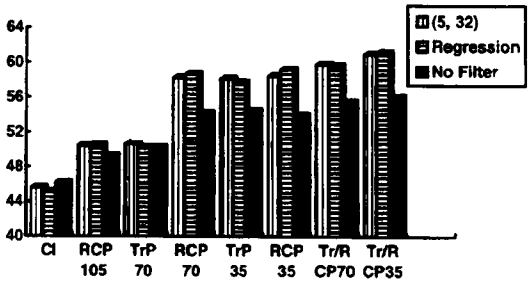

Figure 1: Recognition results (%accuracy) for DB1 and one feature.

2.2. Results with Unmatched Training and Testing Conditions

Figure 2 shows recognition results when the whole of the DB1 database is used for training and the whole of DB2 is used for testing. Those two databases were recorded with a different microphone and have some differences in dialect, speaking rate, etc.. In such conditions, CD units do not improve so much with respect to the CI case (7% for the non-filtered case), since other limiting factors than coarticulation, such as the different microphones, hold. Filtering, however, rounds off the work done by CD units by providing a further 36% improvement, which equals 24% only if no CD units are used.

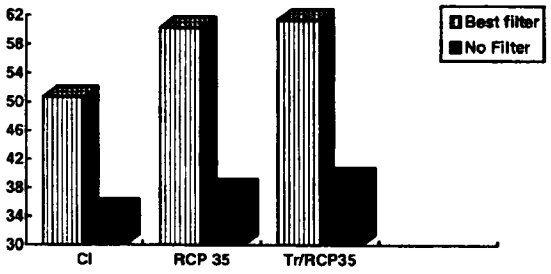

Figure 2: Recognition results (%accuracy) for the mismatched case (DB1 used for training, and DB2 for testing).

2.3. Conclusion

Context-dependent units manage to take into account the natural dependence (caused by coarticulation) between phoneme-sized units. It appears that they also succeed in incorporating the filter-induced dependence into the model, since, both for clean speech and for heavy mismatched conditions, filtering is the most advantageous for sets of context-dependent units with as many triphones as possible and triphones model a broad temporal input

context similar to the extension of the filters we have used. The reason for the success of CDPLU's seems to lie in the selection of contexts which they perform. With context-dependent units, the contexts which are pooled together to train a model are more homogeneous than with context-independent units and thus are more robust to the transition spreading produced by the filter. That increased robustness allows to take advantage of the beneficial effects of the filter such as alleviation of the effect of the acoustic channel or selection of the frequency zone where the discrimination capability of the speech units is greatest.

However, we have worked so far with just one feature. In most practical recognition systems, two features at least are used for increased performance, so the experiments carried out for one feature should be extended to a higher number of features. The next section reports experiments using two features.

3. RESULTS WITH TWO FEATURES

For the two-feature case, the regression filter of length 5 (first-order derivative window) presented above is used to supplement the first feature. In fact, two strategies have been studied: the first, hereafter referred to as series strategy, applies this filter to the first filtered feature, where the filter is a short FIR filter (a Slepian of length varying between 5 and 9) or an IIR filter (a RASTA with a pole at $z=r$), whereas the second strategy, the parallel strategy, applies the regression filter to the original non-filtered feature to give the final second feature, while the first (as is the case for the series strategy) is the original feature filtered with a FIR filter or an IIR filter.

3.1. Results with DB1 for the Series Strategy

Filter	CI	RCP 35
FIR	51.00	59.78
IIR r = 0.97	55.34	63.95
No Filter	55.15	65.15

Table 2: Recognition results (DB1, %acc.) for the series strategy

In this case, filtering seems detrimental to recognition performance and context-dependent units seem unable to cope with it. Only the IIR filter with a pole at $r=0.97$, which does little more than attenuating the low frequency region, yields good results. The long-term spectrum (averaged over all cepstral coefficients) for the non-filtered case (the first feature for the last row of table 2) can be seen in figure 3, and that for the IIR filter with $r=0.97$, by itself and convolved with the impulse response of the regression filter, is shown in figure 4. Only the low frequency region of the spectra is shown. As can be seen from figure 4, the two filtered cases, which correspond to the first and second feature of the last but one row of table 2, only differ in the low frequency region: both annul the DC component but afterwards the first (the IIR by itself) presents a sharp peak very near 0Hz, while the second (the IIR followed by the regression filter) goes up smoothly from the zero at 0Hz to a rounded peak near 6 Hz. The IIR with $r=0.9$ shows the same pattern, except that the two long-term spectra are less different (the initial peak being less sharp). Similar plots to those of figure 3 for Slepian filters show that the long-term spectra of the filter by itself and that of the filter convolved with the impulse response of the regression filter are too much alike. The redundancy in the frequency domain explains the mediocre results, similar to those with one feature. The filter which works best and outdoes even the non-filtered case (for the CI experiment) is the only one where there is some complementariness in the long-term spectrum even if it is so in a small yet perceptually important low frequency region [2]. The IIR filter with $r=0.9$ gives results between those of the IIR filter with $r=0.97$ and those of the FIR filters; in fact, the simple IIR filters outperform any FIR filter from $r=0.8$. In order to try to obtain more different spectra, the second strategy was adopted.

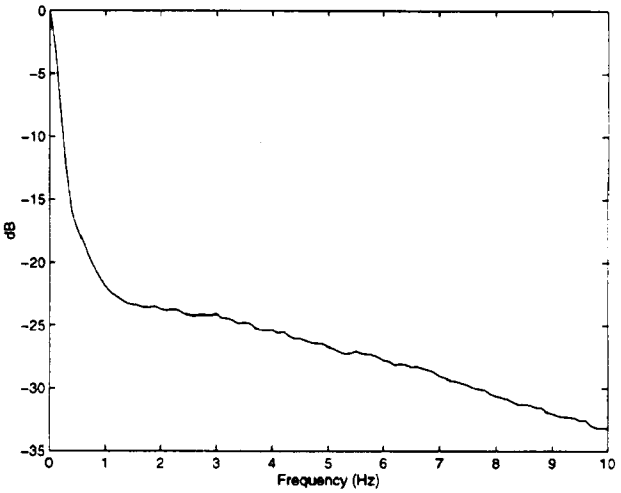

Figure 3: average long-term spectrum of DB1 for the non-filtered case.

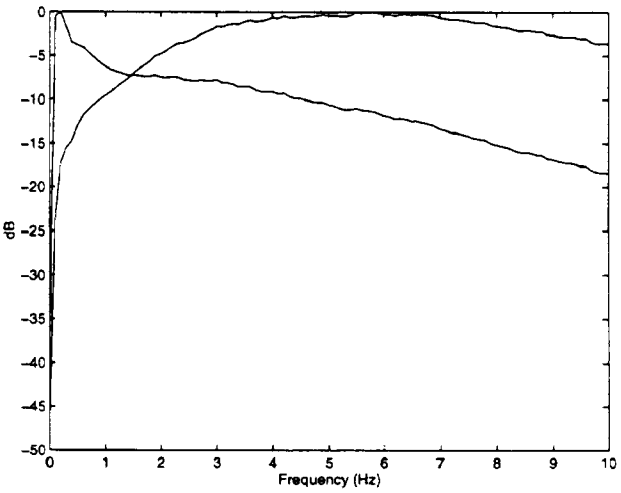

Figure 4: average long-term spectrum of DB1 for IIR $r=0.97$ by itself and followed by a regression filter.

3.2. Results with DB1 for the Parallel Strategy

As can be seen from table 3, the parallel strategy is not effective either and it seems pointless to filter (with or without context-dependent units) when dynamic parameters are included in the feature vector, since the filtered feature (or dynamic feature) already improves the discrimination by filtering the first feature, and since both are handed over to the recognizer. Only removing the DC component (IIR *r=0.97*) seems worthwhile.

Plots similar to figure 4 show that the long-term spectrum of the second feature for the parallel strategy resembles the first feature filtered with a FIR more closely than does the second feature for the series strategy, which agrees with the fact that FIR filters give worse results in table 3 than in table 2. For the IIR filters, the two long-term spectra for the second feature in both cases look very much alike, which explains the similarity of the results obtained.

Filter	CI	RCP 35
FIR	43.57	55.55
IIR r = 0.97	56.26	65.06
No Filter	55.15	65.15

Table 3: Recognition results (DB1, %acc.) for the parallel strategy

In order to determine whether filtering was useful when using two features, we carried out the experiments of section 2.2 (mismatch) using two features.

3.3. Results with Unmatched Training and Testing Conditions, Parallel Strategy

Those results are given in table 4. In this case it is worth filtering: filters improve recognition in all cases and the improvement is higher when context dependent units are used. Even FIR filters, ineffective in the parallel strategy due to the redundance in the bands of the long-term spectrum for DB1, work well. The result for RCP35 is better than that for Tr/RCP35 in accuracy but not in correctness. The result has been kept as it is for the sake of consistency with the other results reported in this work.

Filter	CI	RCP 35	Tr/RCP 35
FIR	50.20	58.44	58.52
IIR r = 0.97	61.26	65.88	62.44
No Filter	48.16	50.90	51.91

Table 4: Results (%accuracy) for the parallel strategy in the mismatch case.

3.4. Conclusion

For clean speech without mismatch and using two features, at least when the second is obtained by filtering with a regression filter, filtering anything else than the zero frequency component seems to have no advantages. Therefore the use of context-dependent units only provides a better modelling of coarticulation, rather than allowing to take advantage of the benefits supplied by filtering (since there are no such advantages if the two features are considered). In mismatched training and testing, filtering does seem advantageous and context-dependent modelling does increase its efficiency, much like what happened with just one feature.

4. CONCLUSION

Simple IIR and FIR filters have been applied to continuous speech recognition. If context-dependent units are used, the side effects of the filters may be remedied and their benefits taken advantage of, which results in a significant improvement in recognition performance that adds up to the one provided by the use of context dependent units. When no dynamic parameters are used, filtering and CD units yield a substantial improvement even for clean speech. When conventional dynamic parameters are incorporated into the feature vector, filtering and CD units seem useful especially for unmatched training and testing conditions. CD modelling must be very effective for improving performance of continuous speech recognition over telephone lines, where filters are usually used. In order to make filtering and CD units useful under all conditions, alternative dynamic features should be found.

5. REFERENCES

1. Nadeu, C., and Juang, B.H., "Filtering of Spectral Parameters for Speech Recognition", *Proc. ICSLP'94* (Yokohama), pp. 1927-30, 1994.
2. Nadeu, C., Pachès-Leal, P., and Juang, B.H., "Filtering the Time Sequence of Spectral Parameters for Speaker-Independent CDHMM Word Recognition", *Proc. Eurospeech'95* (Madrid), pp. 923-26, 1995.
3. Hanson, B.A., and Applebaum, T.H., "Subband or Cepstral Domain Filtering for Recognition of Lombard and Channel-Distorted Speech", *Proc. ICASSP'93* (Minneapolis), pp. II-70, II-73.
4. Chang, J., and Zue, V., "A Study of Speech Recognition System Robustness to Microphone Variations: Experiments in Phonetic Classification", *Proc. ICLSP'94* (Yokohama), pp. 995-98, 1994.
5. Hermansky, H., and Morgan, N., "RASTA Processing of Speech", *IEEE Trans. on Speech and Audio Processing*, Vol.2, No. 4, October 1994.
6. ESPRIT Project: Speech Technology Assessment in Multilingual Applications (SAM-A). Document SAM-A/6002, 1993
7. Bonafonte, A., Estany, R., and Vives, E., "Study of Subword Units for Spanish Speech Recognition", *Proc. Eurospeech'95* (Madrid), pp. 1607-10, 1995.