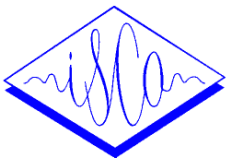

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

Time Derivatives, Cepstral Normalization, and Spectral Parameter Filtering for Continuously Spelled Names over the Telephone

J-C. Junqua[1], D. Fohr[2], J-F. Mari[2], T.H. Applebaum[1], and B.A. Hanson[1]

[1]*Speech Technology Laboratory, Panasonic Technologies Inc., U.S.A.*

[2]*CRIN-CNRS & INRIA Lorraine, France*

ABSTRACT

In this paper, we focus on spectral parameter filtering for reducing the mismatch between training and testing and report experimental results on a continuously spelled name recognition task over the telephone. We studied various time derivative feature combinations, the influence of RASTA processing, short-term and long-term cepstral mean normalization, and the influence of the amount of training data on recognition performance. Based on the results of these experiments, we derived a new front-end for our task, leading to an error rate improvement of almost 30% in name retrieval as compared to previous published results. We also discuss the interaction between the different techniques studied when used in combination.

10.21437/Eurospeech.1995-221

I. INTRODUCTION

Time derivatives, spectral mean normalization, and spectral parameter filtering constitute several of the major techniques contributing to the significant improvements reported recently in the development of front-ends for automatic speech recognition in adverse conditions. More generally, these techniques are useful when there is a mismatch between training and testing conditions. These different approaches can all be interpreted from the point of view of *filtering the sequence of spectral parameters* in a suitable domain.

Time derivatives can be viewed as linear filters of time spectral parameters where two basic effects, a differentiation and a smoothing, are combined [13]. It has been shown that there is a trade-off between the estimation error variance, which increases when the time derivative window length used in the calculation decreases, and the temporal resolution, which decreases as the window length increases [13]. When successive time derivatives are combined in a speech representation, there is a complex interaction between the effects of the various parameter filters. This is also the case when time derivatives are computed from a set of normalized (e.g. with cepstral mean normalization) or already filtered parameters (such as RASTA-based coefficients). Consequently, when different features or various spectral parameter filtering techniques are combined in the extraction of the speech parametric representation, it is not clear how the different filtering effects interact.

Similarly to time derivatives, cepstral mean normalization (CMN) can also be viewed as a spectral parameter filtering which suppresses the slowing varying distortions [7]. In [11] is was shown that the average of the cepstral vectors (long-term cepstrum) represents the telephone channel. Cepstral normalization has been shown to be effective using long-term averages (e.g. [11]) or short-term averages (e.g. [14]).

Spectral normalization can also be achieved by means of a band-pass or high-pass filtering which suppresses the lower modulation frequencies of a subband analysis (e.g. [8; 9]). The time constant of the band-pass filter determines the amount of suppressed information.

The main differences between all these equalization techniques lie in the filter shape and the filter time constant. In an IIR filter-based technique such as RASTA, the filtered output depends also on the speech signal history. If we compare the spectral shape of these filters, as done in [7], the first time derivative filter has a more selective frequency response than the RASTA filter, and the cepstral mean filter cut-off is typically lower than that of the RASTA and first derivative filter. In all these spectral parameter filtering techniques the length of the filter time response is critical and is directly related to the frame rate and the degree of temporal smoothing introduced by the number of frames used in the filtering calculation.

In this paper, we experimentally assess the contribution of the various spectral parameter filtering techniques just mentioned and discuss how these different techniques interact. The spectral parameters studied are derived from two different analysis methods: the Perceptually-based Linear Prediction Analysis (PLP) [5] and Linear Prediction (LP) mel-cepstral analysis. The evaluation is done in the context of a continuously spelled name recognition task over the telephone.

II. TASK, RECOGNITION STRATEGY AND DATABASE

Automatic speech recognition of spelled names is a difficult task because of the confusable letters contained in the alphabet, the distortions introduced by the telephone channel and the variability due to an undefined telephone handset. In [10] we proposed a solution to this problem by means of a multi-pass recognizer, propagating N-best hypotheses through different processing modules. This recognizer is based on a frame synchronous HMM recognizer with adaptive beam search. N-best candidates are propagated through the different passes and dynamic grammars are used to provide the final constraints.

The database used in our experiments is a subset of the speech telephone corpus collected at Oregon Graduate Institute (OGI) [3]. Over four thousand people called in response to public requests. They were prompted by a recorded voice to say their first and last names, with and without pauses, together with other information. 60 repetitions of the alphabet and more than 1,200 different calls were selected for the training, 558 calls for the validation and 491 calls for the test. As every speaker belongs only to one set (training, validation or test) the experiments conducted are speaker-independent. For some of the tests, to introduce a mismatch between training and testing conditions, a fixed, second-order pole band-pass filter [12] was applied to the test data, leading

to "distorted channel" data. The confidence interval of the results presented is generally less than +/- 1.6%.

III. EXPERIMENTAL RESULTS WITH SEVERAL SPECTRAL PARAMETER FILTERING TECHNIQUES USED ALONE OR IN COMBINATION

As is already well known, the first derivative has been shown to provide impressive improvements when combined with static features in the speech parametric representation (e.g. [4]). The second derivative, somewhat more noisy than the first derivative, was also reported to provide recognition improvements (e.g. [1]).

In the context of our task, we first evaluated the contribution of time derivatives on the recognition performance with both an 8th-order PLP and Mel Frequency Cepstral Coefficients (MFCC) derived from a 14th-order LP analysis. The energy and, for the experiments using the regression features, the regression values of the energy were included in the parametric speech representation. 9 PLP coefficients (including the energy) and 11 MFCC (including the energy) were used in the static feature which was used to compute time derivatives. For the two analyses, the window length and the frame shift were, respectively, 20 and 10 msec for the PLP analysis and 32 and 16 msec for the LP-based MFCC parametrization. Figures 1 and 2 summarize these results (some of these experiments have also been reported in [10]).

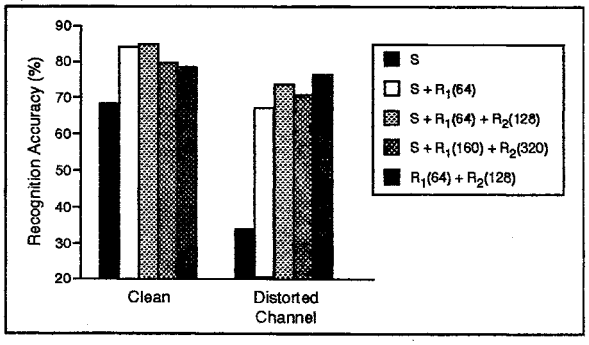

FIGURE 1. Letter recognition accuracy obtained with various MFCC-based feature sets (S stands for static feature, and R_1 and R_2 for first and second-order regression features). The number in the parentheses indicates the window length used in the calculation of the time derivatives (in msec).

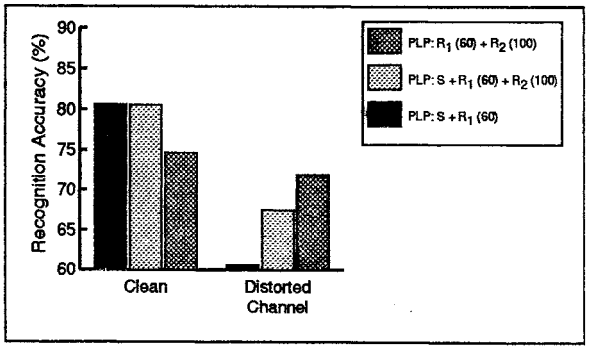

FIGURE 2. Letter recognition accuracy obtained with various PLP-based feature sets.

In these figures and all the other ones in this paper, recognition accuracy was computed by taking the percentage of the number of correctly recognized letters minus the number of insertions over the total number of letters. These two figures show that

- including a second derivative slightly improves recognition accuracy for clean speech;
- long regression windows for the first and second derivatives decrease recognition accuracy;
- the combination of R_1+R_2 compensates well for the mismatch between training and testing conditions; however, R_1+R_2 alone decreases the recognition accuracy for clean test data;
- a 14th-order LP-based mel frequency analysis performs better than an 8th-order PLP on this task (especially when there is a mismatch between training and testing).

In [2] long time derivative windows have been found to provide the best performance for isolated words. However, our results together with some other recent studies (e.g. [13]) suggest that this may not be true for continuous speech. Some insights, which can explain these results, can be found in [13]. In this paper, the authors showed that the long-term power spectrum of the time sequence of spectral parameters has a larger bandwidth for continuous speech than for isolated words. A short regression window provides less frequency resolution (greater bandwidth) than a long regression window, and thus is more suitable to capture the information in continuous speech.

When there is a mismatch between training and testing, the normalization effect produced by time derivatives, not present in the static features, helps compensate for the mismatch between training and testing conditions. As shown in Figures 1 and 2, in case of large mismatch, the static feature may be dropped.

Figure 3 shows experimental results when time derivatives are combined with another spectral parameter filtering technique, namely RASTA processing. For these results the filter time constant of RASTA processing was optimized on this task and set to 0.90.

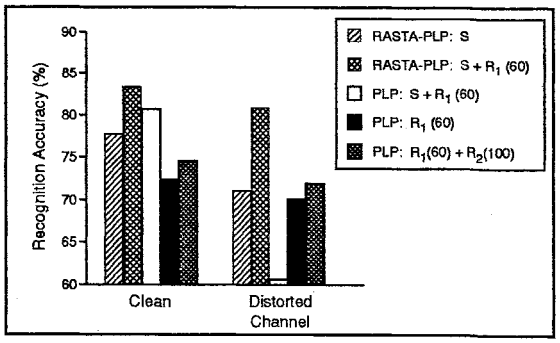

FIGURE 3. Letter recognition accuracy obtained with various 8th-order RASTA-PLP and PLP-based feature sets.

Figure 3 emphasizes the fact that RASTA-PLP is more than simply a derivative effect and that significant gain in performance is obtained by augmenting the RASTA static feature with its first derivative. Additional experiments, not plotted in Figure 3, showed that adding the RASTA second time derivative feature decreases the recognition accuracy by approximately 2%. Static features from RASTA-PLP yield generally better results

than those obtained from the PLP-based R_1 regression feature alone. This is due to the differences in the filter frequency responses of the two techniques. While it is possible to match the initial slope of the frequency response of both filters by adjusting the time constant of the RASTA filter and the frame length and number of frames of the first derivative calculation, it is not possible to closely match the frequency responses of the filters to provide very similar filtering effects. The main difference between the RASTA filter and the first derivative filter is that the RASTA filter has a broader passband. The large improvement provided by RASTA processing on clean speech is due to the fact that we are dealing with telephone speech.

To evaluate the interaction between RASTA processing and the window length used in the calculation of the first time derivative, we assessed the influence of the length of the frame shift on the recognition accuracy. Figure 4 shows the results obtained for two different lengths of the frame shift: 10 and 16 msec (the analysis window was always twice the frame shift).

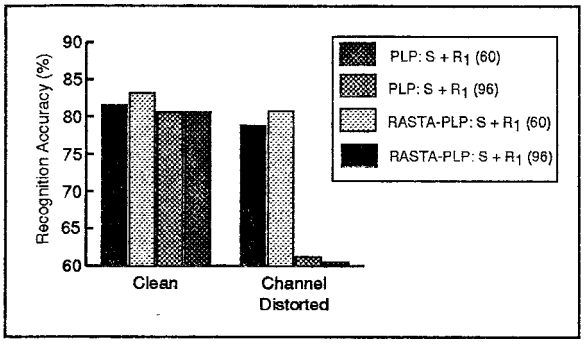

FIGURE 4. Letter recognition accuracy with two different lengths of the frame shift (10 and 16 msec) of PLP- and RASTA-PLP-based feature sets.

It can be seen that, while increasing the length of the frame shift can lead to an improvement of the recognition performance (however small and only for the distorted channel case, but more significantly for the MFCC case not represented in the figure), it is detrimental to RASTA spectral parameter filtering. The increase of the length of the frame shift seems to slightly improve the PLP-based feature set, but it increases the effective time constant of the RASTA filter leading to a decrease in recognition accuracy for RASTA-PLP.

To remove the slowing varying parts of the MFCC-based spectral representation, we assessed two CMN techniques based on either a short-term computation of the cepstral mean over several frames of speech data, or a long-term cepstral mean computation on the whole utterance. Results are shown in Figure 5.

It can be seen that long-term cepstral mean normalization combined with the MFCC analysis and the three feature set ($S+R_1+R_2$) gives the best results. Compared to the best RASTA-PLP results, letter recognition accuracy was improved by almost 3%. Short-term cepstral mean normalization, suitable for real-time purposes, mainly improved the recognition accuracy when the mismatch between training and testing is important. For the short-term cepstral normalization, different window lengths gave similar results. Additional experiments, where the long-term cepstral mean was computed over the previous speech sentence (corresponding to a different call), yielded a recognition accuracy lower than when the long-term cepstral mean is computed on the target test sentence (decrease of about 3%).

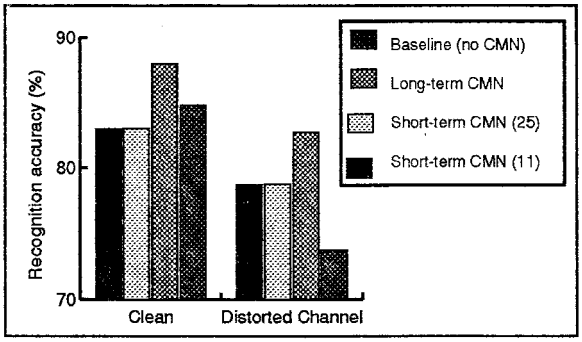

FIGURE 5. Influence of different cepstral mean normalization techniques on $S+R_1+R_2$ derived from the MFCC coefficients. The number of frames used to compute the short-term cepstral mean is indicated between parentheses (the frame length was 16 msec).

CMN was also used in combination with RASTA processing. Results are presented in Figure 6. In contrast to the results presented in [6], the combination of RASTA processing and CMN does not provide any improvement for clean telephone speech. However, a slight improvement is obtained in case of a mismatch between training and testing. A possible difference in the amount of mismatch between training and testing in our experiments compared to that of [6] for the clean speech case may explain our different results. Even if both RASTA filtering and CMN tend to remove the long-term average log spectrum, our results indicate that it may be useful to combine RASTA processing and CMN to better equalize the mismatch between training and testing.

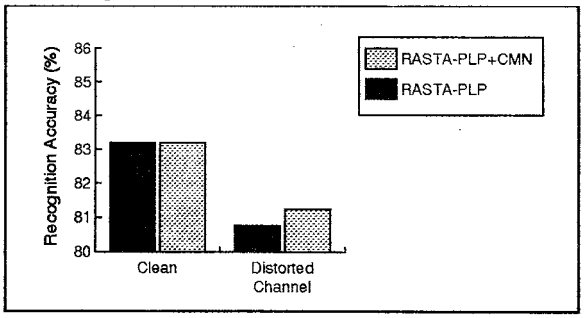

FIGURE 6. RASTA processing and cepstral mean normalization on PLP-based coefficients (static and first derivative).

IV. INFLUENCE OF THE TRAINING DATA ON RECOGNITION ACCURACY

By using spectral parameter filtering techniques, which suppress phonetically irrelevant variability, it should be possible to decrease the amount of training data while still maintaining the same level of performance. One way to evaluate this hypothesis is to decrease the amount of training data and to assess the effect on recognition performance. Results from such experiments are presented in Figure 7. It can be seen that the LP-based MFCC analysis (without CMN) is more

sensitive to the reduction of training data (the reduced training data corresponds to a reduction of 4) than the RASTA-PLP analysis. When the MFCC analysis was combined with CMN, we did not observe this behavior, confirming the effectiveness of CMN in reducing the variability.

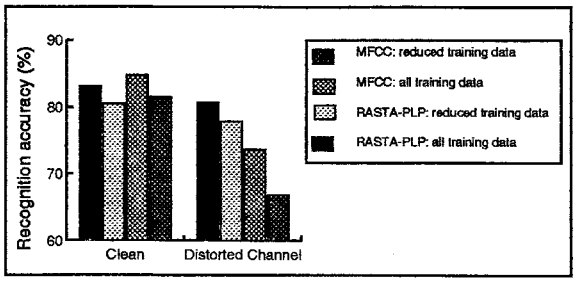

FIGURE 7. Influence of the size of the training data on MFCC ($S+R_1+R_2$) and RASTA-PLP ($S+R_1$) feature sets.

V. DISCUSSION AND CONCLUSIONS

All the methods studied can be interpreted in the framework of spectral parameter filtering. The major differences between all these techniques lie in the frequency responses of the filters which differ in the amount of spectrum equalization and accuracy of the estimation. The common goal of these techniques is to eliminate the low modulation frequencies. In this paper, we evaluated experimentally the effects of parameter filtering and the interaction between various filters in the case of telephone speech and in the presence of a mismatch between training and testing.

We experimentally observed that there is a complex interaction when these techniques are combined and that, among the various spectral parameter filtering techniques studied, a combination of MFCC static, first and second regression features with long-term CMN provides the best results. The normalization effects of these techniques permit a reduction in the amount of training data necessary to obtain a certain level of accuracy.

For real-time purposes the low dimensionality of the speech parametrization yielded by the PLP analysis and the ability to compute RASTA filtering in real-time makes RASTA-PLP still attractive, even if better performance can be obtained with other techniques.

As spectral parameter filtering techniques need to be combined to yield an accurate signal estimation while reducing the effects of mismatch between training and testing, trade-offs between temporal resolution and estimation errors become quite apparent.

As there is not a single feature robust to distortions and accurate enough to give a good representation of the speech signal, features and enhancement techniques need to be combined. However, as the interaction between the different features is difficult to control, Linear Discriminant Analysis (LDA) may help deal with feature correlation and the complex interaction already mentioned. We performed several preliminary experiments to assess the effect of LDA on our best combination of features ($S+R_1+R_2$ obtained for LP-based mel cepstrum analysis combined with CMN). Our first results indicated that LDA did not yield any improvement for clean telephone speech. However, better robustness in the distorted channel case was observed.

The best results of our study were obtained when long-term cepstral mean normalization was combined with the feature set $S+R_1+R_2$ derived from the MFCC analysis. Letter recognition accuracy was improved by almost 3% and name retrieval recognition increased from 95.3% to 96.7% on a name retrieval evaluation with a 3,388 name dictionary.

REFERENCES

[1] T.H. Applebaum and B.A. Hanson. Robust speaker-independent word recognition using spectral smoothing and temporal derivatives. In *EUSIPCO*, pages 1183–1186, 1990.

[2] T.H. Applebaum and B.A. Hanson. Tradeoffs in the design of regression features for word recognition. In *EUROSPEECH*, pages 1203–1206, 1991.

[3] R. Cole, K. Roginski, and M. Fanty. English alphabet recognition with telephone speech. In *EUROSPEECH*, pages 479–482, 1991.

[4] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. ASSP*, ASSP-34:52–59, 1986.

[5] H. Hermansky, B.A. Hanson, and H. Wakita. Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain. *Speech Communication*, 4(1-3):181–187, 1985.

[6] H. Hermansky and N. Morgan. RASTA processing of speech. In *ASR IEEE Workshop*, pages 91–92, 1993.

[7] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing*, 2(4):578–589, October 1994.

[8] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In *EUROSPEECH*, pages 1367–1370, 1991.

[9] H.G. Hirsch, P. Meyer, and H.W. Ruehl. Improved speech recognition using high-pass filtering of subband envelopes. In *EUROSPEECH*, pages 413–416, 1991.

[10] J-C. Junqua, S. Valente, D. Fohr, and J-F. Mari. An N-best strategy, dynamic grammars and selectively trained neural networks for real-time recognition of continuously spelled names over the telephone. In *ICASSP*, pages 852–855, 1995.

[11] C. Mokbel, P. Pachès-Leal, D. Jouvet, and J. Monné. Compensation of telephone line effects for robust speech recognition. In *ICSLP*, pages 987–990, 1994.

[12] H. Murveit, J. Butzberger, and M. Weintraub. Reduced channel dependence for speech recognition. In *DARPA Workshop Speech and Natural Language*, pages 280–284, February 1992.

[13] C. Nadeu and B-H. Juang. Filtering of spectral parameters for speech recognition. In *ICSLP*, pages 1927–1930, 1994.

[14] A.E. Rosenberg, C-H. Lee, and F.K. Soong. Cepstral channel normalization techniques for HMM-based speaker verification. In *ICSLP*, pages 1835–1838, 1994.