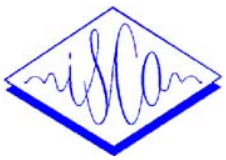

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

Down-sampling speech representation in ASR

Hynek Hermansky[1],[2], **Pratibha Jain**[1]
[1]**Oregon Graduate Institute of Science and Technology**, Portland, Oregon, USA.
[2]**International Computer Science Institute**, Berkeley, California, USA.

10.21437/Eurospeech.1999-22

Abstract

Features for automatic speech recognition (ASR) are typically sampled at about 100 Hz (10 ms analysis step). Recent experiments indicate that the most efficient components of the modulation spectrum of speech for ASR are up to about 16 Hz [1]. Consequently, RASTA processing attenuates modulation frequencies higher than 16 Hz and should in principle allow for a subsequent down-sampling of the features.

It has been shown earlier that in a Gaussian mixture model based speaker recognition system(which uses single state HMM, thus not requiring any time alignments of the incoming speech) one could down-sample the speech representation after RASTA filtering without any significant loss of performance [2]. However since ASR uses Viterbi time alignment, reduced number of time samples due to down-sampling, although justified by Nyquist criteria after the low-pass filtering, could create problems.

In this paper we experimentally show that the down-sampling of features after RASTA filtering is feasible and could result in considerable computational or at least storage/transmission savings.

1 Temporal processing

Speech contains many source of information such as information about the linguistic message, about the speaker of the message, and about the communication channel used for the recording and transmission of the speech signal. For a given task, it is helpful to retain relevant source of information in extracted features while suppressing the irrelevant ones.

In ASR, the task is to decode the linguistic message This linguistic message is coded in the movements of the vocal tract. The speech signal reflects these movements. The rate of change of the non-linguistic components in speech often lies outside the typical rate of change of the vocal tract shape. The RASTA [3] and LDA [4] techniques take advantage of this fact and bandpass filter time trajectories of speech feature vectors.

1.1 Exploiting the bandpass property

RASTA filters out the fast (and slow) changes of spectral components over time. Since fast changes (high modulation frequency components) are eliminated by RASTA filtering, the Nyquist criterion would suggest that the RASTA filtered features could be sampled at a sampling rate slower than that of original non-filtered features(Fig 2).

1.2 RASTA-PLP

In our case (RASTA-PLP) [5] the RASTA filtering is performed on trajectories of log critical-band spectral energies. After RASTA bandpass filtering, the trajectories are band-limited to around 12 Hz. PLP cepstral features are extracted from the resulting representation by an all-pole modeling which fits the spectral frames to yield autoregressive coefficients. For ASR, these are then typically converted to the model cepstra coefficients.

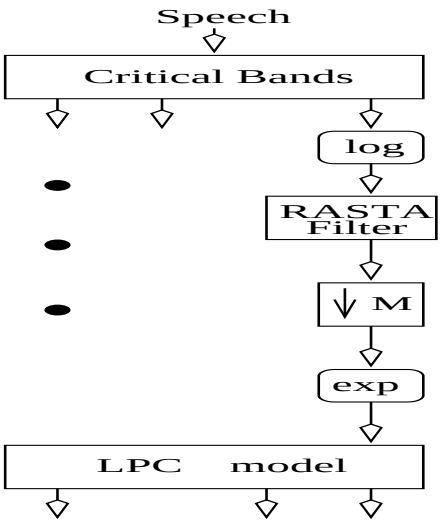

Figure 1: Proposed System

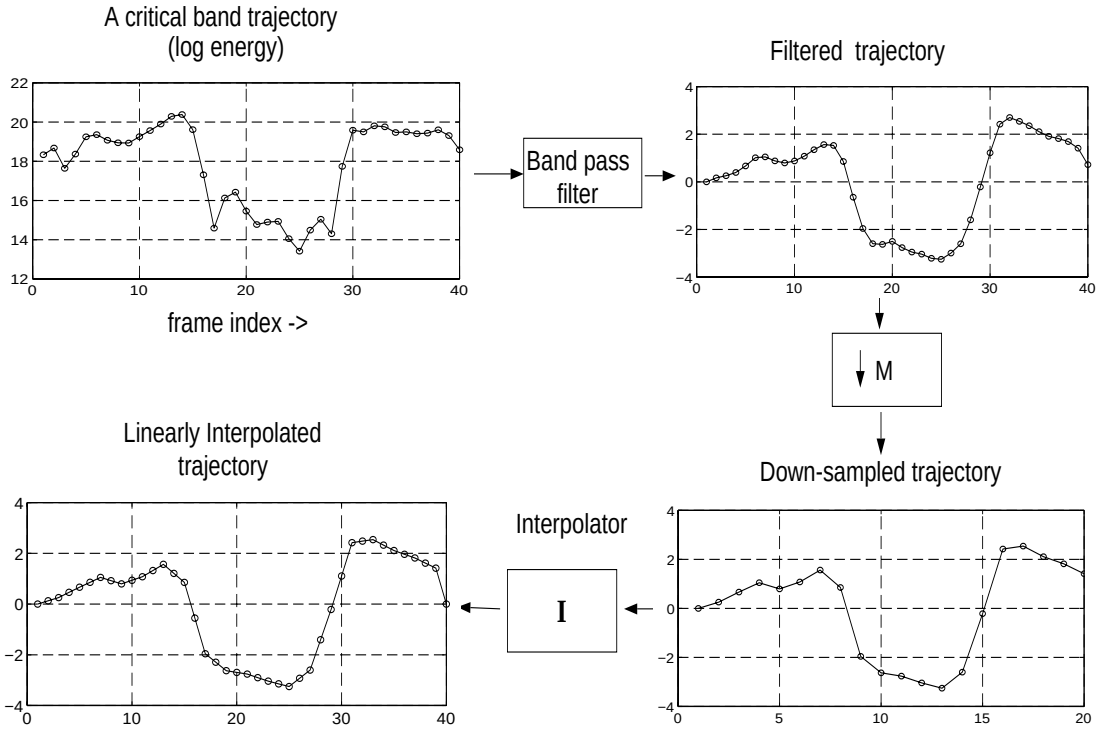

Figure 2: Filter bank Trajectories

2 Thesis

According to Nyquist criteria if

$$f_s \geq 2.w_s$$

where f_s is the sampling frequency and w_s is the bandwidth of the signal, the signal can be reconstructed from its sampled version. Thus, we should be able to use the down-sampled RASTA filtered features without any loss in information (Fig 1).

3 Experiment

We have used a subset of OGI Numbers corpus [6] for the recognition experiments. It consists of continuous spoken digits with utterances one to seven digits long. The training set consists of 2546 utterances and the test set contains 2156 utterances. Eight RASTA-PLP cepstral features along with delta and delta-delta were used at 10ms frame rate with and without down-sampling. For down-sampling case we dropped every alternate frame during training. As we were mainly interested in comparisons of the original and down-sampled features, a simple HMM recognizer used for these experiments consisted of 23 context-independent phoneme models.

The experiments were performed for 3 states 5 mixture and 5 state 3 mixture models using HTK toolkit [7].

4 Results

4.1 Simple three-state phoneme models

For these simple models we have used the down-sampled features directly and observed practically no degradation in performance (Table 1).

Frame rate	Word error rate (%)
$10ms$(without down-sampling)	9.07
$20ms$(with down-sampling by 2)	9.08

Table 1: With 3 state 5 mixture models

4.2 More complex five-state phoneme models

However, in some situations the more complex models may be desirable. Thus, when we increased the complexity of our context independent phone models to five

states, word error rate decreases as seen in table 2. Unfortunately, in such a case some phones may contain fewer frames than the number of states in the phone model and (as was in our case after down-sampling) may fail to train. In this case our solution was to explicitly interpolate the available down-sampled representation prior to Viterbi search. We have observed that even the very simple linear interpolation yields the same performance as with original features (Table 2).

Frame rate	Word error rate (%)
$10ms$(without down-sampling)	8.01
$10ms$(with Interpolation)	8.01

Table 2: With 5 state 3 mixture models

Of course, in this case, we lose the advantage of faster Viterbi search. Still, this solution may be attractive when the speech features need to be stored or transmitted, since one only needs to store or transmit the down-sampled representation.

5 Conclusions

As predicted by Nyquist criterion, the band-pass characteristic of RASTA filters allows for down-sampling the feature representation. This has been previously found useful in Gaussian mixture model based speaker recognition which does not explicitly use any temporal alignments [2].

In the current work we experimently verified that the RASTA-processed down-sampled features can also be used in ASR. For phoneme models with simple temporal structures, the down-sampled features can be used directly. Models with larger number of states may require some minimum number of feature frames per phoneme. In this case one could still store or transmit the down-sampled features and interpolate the features in the classifier.

References

[1] Hynek Hermansky Noboru Kanedera, Takayuki Arai and Misha Pavel, "On the importance of various modulation frequencies for speech recognition," in *Proc. EUROSPEECH*, Greece, 1997, vol. 2.

[2] S. van Vuuren and Hynek Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *Proc ICSLP*, Australia,Sydney, 1998, vol. 2.

[3] Hynek Hermansky and Nelson Morgan, "Rasta processing of speech," in *IEEE Transactions on Speech and Audio Processing*, 1994, vol. 2, pp. 587–589.

[4] S. van Vuuren and H. Hermansky, "Data-driven design of rasta-like filters," in *Proc. of EUROSPEECH*, Greece, 1997, pp. 409–412.

[5] Nelson Morgan Hynek Hermansky H. Guenther Hirsch Joachim Koehler and Grace Tong, "Integrating rasta-plp into speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, speech and signal Processing*, Australia, 1994, pp. 421–424.

[6] T. Lander R.A. Cole, M. Neol and T.Durham, "New telephone speech corpora at cslu," in *Proceedings of EUROSPEECH*, 1995, pp. 821–824.

[7] Steve Young, "Htk toolbox," in *HTK book*, Cambridge University, 1997.