Revisiting Scenarios and Methods for Variable Frame Rate Analysis in Automatic Speech Recognition

J. Macías-Guarasa, J. Ordóñez, J.M. Montero, J. Ferreiros, R. Córdoba and L.F. D'Haro

Speech Technology Group. Dept. of Electronic Engineering. Universidad Politécnica de Madrid
ETSI de Telecomunicación. Ciudad Universitaria s/n. 28040-Madrid SPAIN
macias@die.upm.es

Abstract

In this paper we present a revision and evaluation of some of the main methods used in variable frame rate (VFR) analysis, applied to speech recognition systems. The work found in the literature in this area usually deals with restricted conditions and scenarios and we have revisited the main algorithmic alternatives and evaluated them under the same experimental framework, so that we have been able to establish objective considerations for each of them, selecting the most adequate strategy.

We also show till what extent VFR analysis is useful in its three main application scenarios, namely "reduction of computational load", "improve acoustic modelling" and "handling additive noise conditions in the time domain". From our evaluation on a difficult telephone large vocabulary task, we establish that VFR analysis does not significantly improve the results obtained using the traditional fixed frame rate analysis (FFR), except when additive noise is present in the database and specially for low SNRs.

1. Introduction

Every contemporary automatic speech recognition systems has a feature preprocessing stage, which aims at reducing the inherent redundancy of the speech signal and extracting a sequence of feature vectors, each of them summarising the necessary temporal and spectral behaviour of a short segment of the acoustical speech input. The ultimate goal is to estimate the sufficient statistics to discriminate among different phonetic units while minimising the computational demands of the classifier [4].

In most cases, the input signal is first windowed into frames, with a certain overlapping between adjacent windows. Regarding this process, there are two different approaches when deciding the frame shift used:

- Fixed frame rate analysis (FFR): The frame shift between two adjacent frames is always the same (this is the traditional approach and most commonly used method in preprocessing modules)
- Variable frame rate analysis (VFR): The frame shift between adjacent windows varies and thus, the frame rate is not uniform

VFR techniques allow us to adjust the frame rate according to the value of some specific metric (typically related to the level of spectral variations detected), computing more feature vectors in those regions in which spectral information varies faster.

In the bibliography, we can find several references to VFR analysis applied to speech recognition systems. The major drawback of the published works is that they describe and evaluate specific algorithmical approaches in limited experiments without a clearly specified applicability criterion. We can classify the VFR-related literature as oriented towards three main objectives:

- Reduction in the number of processed frames without performance loose, in [4, 1], for example. This approach is attractive as it allows savings in computational load for subsequent modules
- Better acoustic signal modelling in regions with fast spectral changes by using lower values of the frame shift used, as in [6].
- Increased immunity to performance loss in tasks suffering from additive noise in the time domain, as in [6, 2].

In this paper we want to provide an overview of the scenarios in which variable frame rate is useful for automatic speech recognition, giving new points of view and comparing the different methods for VFR under a common experimental environment.

2. Variable frame rate methods overview

All VFR-based strategies discussed in the literature, have a common algorithmic sequence in order to get the feature vectors:

1. Frame vectors are computed using a fixed frame rate which determine the minimum shift time between adjacent frames.
2. The spectral change level is calculated according to a well defined metric.
3. A comparison between the measured spectral change and a pre-calculated threshold is done. Frames with a spectral change level under the selected threshold are discarded, while keeping them otherwise.

However, the methods to evaluate the spectral change level are different:

- In [4], they calculate the distance between the current feature vector and the last one that was not rejected
- In [1], they calculate the norm of the first derivative cepstrum vector
- In [6], they calculate the Euclidean distance between adjacent feature vectors, applying a log-energy weighting (the objective is to give more importance to frames with

high energy and thus, less affected by noise). A frame is selected if the cumulative weighted distance since the last not rejected vector is higher than the threshold.

We have done an exhaustive preliminary experimentation in order to select the best approach, and found this to be the one described in [6]. We also found some drawbacks in the method described in [4] (which will be referred to as the "classical method") and in [1] (which will be referred to as the "derivative method"). In figure 1 we show the spectrogram for the Spanish word *tapias* and, for each method, the evolution of the calculated metrics and the place where frames were accepted (indicated by vertical lines). In this case, we have used a fixed frame shift of 10 ms. and we trained the threshold in order to achieve an equivalent frame shift of 22.5 ms, as it will be done in section 4 where we aim at reducing the computational load). In figure 1 we can see that very few frames are kept in vowels regions (even less than in the silence region) and results achieved in our experiments are not good.

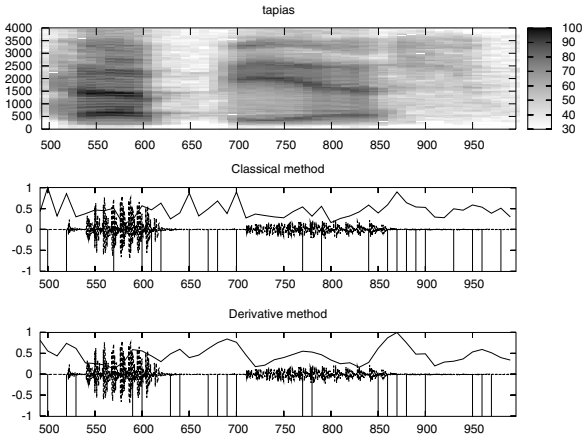

Figure 1: Frames selected for classical and derivative methods

In figure 2 we have depicted the frames that will be chosen when the method described in [6] is used (cumulative weighted distance). In addition to that, we show the same information when the first derivative is used instead of the Euclidean distance. In both cases, the selected frames distribution is very similar, and due to the weighting and cumulative processes, frames are spread out in a reasonable way: few frames in silent regions, a few more in stationary regions with high energy (vowels) and even more in the transient regions, where spectral information changes pretty fast.

3. Experimental setup

Experiments have been carried out using part of VESTEL, a realistic isolated word telephone speech database, captured using the Spanish PSTN and composed of 9,790 utterances. We have used the *leave one out* method with ten subgroups in order to increase the statistical significance of the results.

The dictionary used in this paper is composed of 1,946 words. The feature vector is composed of 10 cepstral MFCC coefficients, 10 cepstral derivatives, log energy plus their first derivative. The speech recogniser uses 45 allophones and context dependent HMMs (the detailed architecture is described in [3]. Recognition rates are shown including a confidence interval for 95% statistical significance.

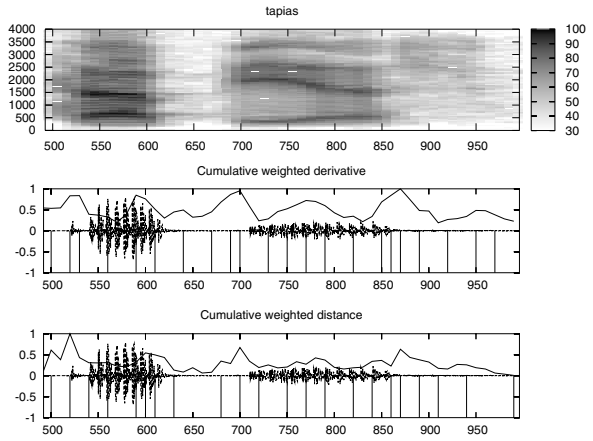

Figure 2: Frames selected for cumulative weighted distance and derivative methods

4. Experimental results

In this section we will show some of the experimental results we obtained using VFR analysis in the three main application scenarios described above.

4.1. Reduction of computational load

Experiments in [4] and [1] aim at a reduction of computational load by discarding calculated frames. In order to do so, they start using a frame shift of 10 ms., and discard feature vectors using the corresponding strategy described above. In both papers it can be seen that discarding vectors using VFR techniques, it is possible to keep the recognition rate and even improving it. VFR is useful in this scenario as the discarding process is done in a non uniform way, so that the non-discarded vectors are the ones really involving important spectral changes.

However, in the mentioned papers they only make a comparison between the VFR achieved rate and the one obtained using a FFR with a frame shift of 10 ms. In our opinion, and in order to offer a fully objective evaluation, a comparison should be made also with the result we would get when using a FFR strategy which calculates exactly the same number of frames. So, we start also with a 10 ms. frame shift, and discard vectors until we get an equivalent frame shift of 22.5 ms (which corresponds to a reduction of 55.56% in the number of calculated vectors). The point we want to stress here is the following: Is the observed improvement due to the efficiency of VFR methods or we could get the same results using FFR analysis with higher frame shift values?

In figure 3 we show the recognition rates achieved by FFR analysis and frame shift of 10 ms. and 22.5 ms. and the ones obtained by VFR analysis with an equivalent frame shift of 22.5 ms. using the cumulative weighted distance and derivative. As it can be seen, even though we have discarded more than half of the frames, the recognition rate is still higher than the baseline (using FFR), which is consistent with the results found in the literature, but the difference is not statistically significant when compared to the result with a fixed frame shift of 22.5 ms. Of course, we have to take into account that we are using an isolated speech database, and these results will have to be further validated on continuous speech tasks.

To summarise, the reduction in computational load is significant when using VFR, but there are no real advantages as

compared with a system with the same computational demands (same number of processed feature vectors) using FFR analysis.

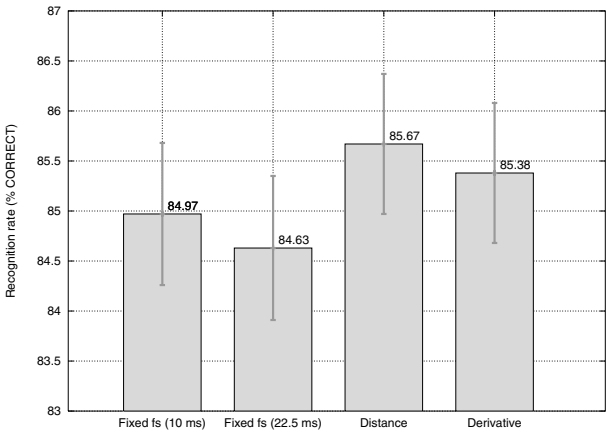

Figure 3: Recognition rate comparison FFR vs. VFR

We finally characterise the system behaviour for a wide range of frame shift values, results which are summarised in figure 4, in which we can observe that a too low frame shift leads to very poor results. This is due to the increase of insertions produced in the acoustic decoding process, as the HMM is obliged to "swallow" too many frames (we have not changed the number of states per model). The most interesting observation is that, for a broad range of frame shift values for which we can get important reductions in computational demands, the recognition rate performance in not affected significantly.

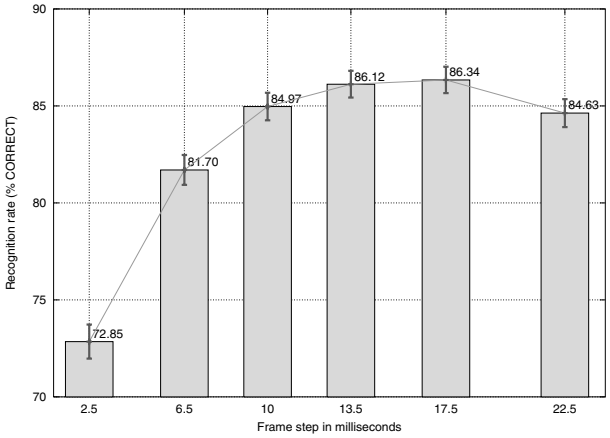

Figure 4: Recognition rate for FFR analysis

Finally, we also want to stress that (under this scenario), even though VFR seems to be superior to FFR, the differences are not statistically significant, so that depending on the discarding rate, it's not always useful to apply VFR.

4.2. Experiments to improve acoustic modelling

In [6], VFR analysis is applied using frame shifts lower than 10 ms. The objective is improving the acoustic modelling by paying more attention to speech segments in transient regions where spectral changes are faster.

In clean speech conditions, VFR analysis was better than FFR when they evaluated a database composed of nasal consonants but didn't work when TIDIGITS was used.

In this scenario, we wanted to evaluate the proposed strategy in a much more complex task: the VESTEL database and a dictionary of 1,946 words. In our experiments, we started from a 2.5 ms frame shift and discarded frames till we got an equivalent frame shift of 10 ms (using the weighted cumulative distance method). We got a recognition rate of 84.97%$\pm$0.71 using the FFR strategy with a 10 ms. frame shift, while we got 84.57%$\pm$0.72 using the VFR method. As it can be seen, VFR does not improve the result achieved with fixed frame rate in our experimental conditions, showing that VFR does not achieve any improvement when the acoustic conditions of the evaluation database are very general (which is specially the case in VESTEL). The improvements obtained in the nasal database are probably due to the very limited acoustic environment involved.

4.3. Experiments to deal with additive noise

In [6], VFR analysis is used to reduce the effect of noise in system performance. There, a speech shaped noise was used, obtaining reasonably good results.

In this scenario, we wanted to validate those results in a much more complex task (VESTEL + 1,946 words dictionary) along with different types of noise (white Gaussian noise and Volvo noise, both of them belonging to the NOISEX database [5]). We contaminated the database using the standard procedures and estimating the speech signal level using the P.56 recommendation of ITU-T, and, for each type of noise, we have computed the recognition rate at a given SNR in the following conditions:

- FFR analysis with 10 ms. frame shift, with both the training and testing sets contaminated with additive noise. Of course, this is an unreal situation as in real-world tasks it is not possible to know in advance which type and power of noise we will find. This is, however, a good way to estimate the maximum achievable recognition rate
- FFR analysis with 10 ms. frame shift, with training performed in clean conditions and testing done in noisy conditions. The results obtained under these conditions are supposed to achieve the minimum recognition rate and it will be the rate to beat using VFR analysis.
- VFR analysis using cumulative weighted distance. We start from a 2.5 ms. frame shift and discard feature vectors until getting an equivalent frame shift of 10 ms. Training is done with clean data and testing is done with noisy speech.

In figure 5 we show the results obtained when we add white noise to the speech signals with SNR's of 15 and 20 dB. For both SNRs, the VFR strategy gets statistically significant improvements when compared with the FFR case in the same conditions (clean training, noisy testing).

In figure 6 we show the recognition rates achieved when Volvo noise is added. In this case we have used SNR's of 15 and 3 dB (lower than the ones we used when white noise was added because Volvo noise is less harmful). For high SNR's (15 dB in the case of Volvo noise) using VFR analysis doesn't improve the recognition rate, but when SNR decreases down to 3 dBs, VFR analysis clearly surpass FFR preprocessing (under the same conditions) in a statistically significant way.

The important improvements obtained are due to the fact that the VFR method preferably discards the feature vectors

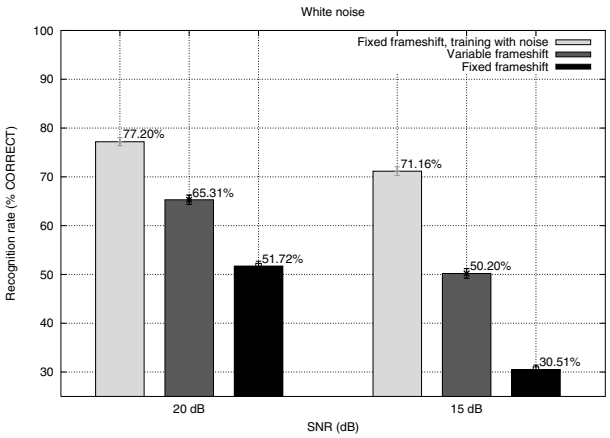

Figure 5: Recognition rate in additive white noise conditions

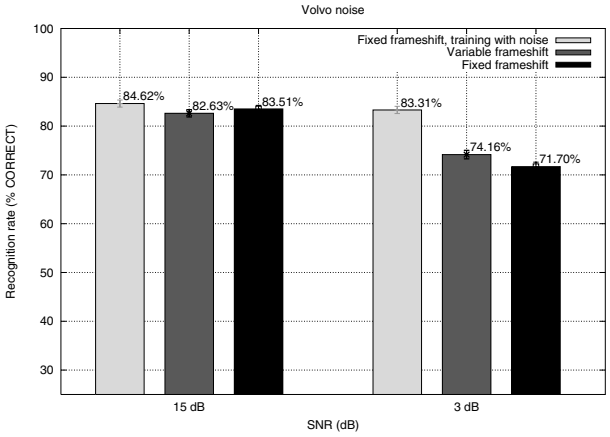

Figure 6: Recognition rate in additive Volvo noise conditions

more affected by noise (those with little impact in what respect to spectral change and with lower energy, as we are using the log-energy weighting)

To summarise, when using VFR analysis in additive noise conditions, the lower the SNR, the better the VFR performs compared with FFR methods.

5. Conclusions and future work

In this paper we've done a brief revision and evaluation of some of the methods used in VFR analysis applied to speech recognition.

We have revisited the main algorithmic alternatives and evaluated them under the same experimental framework, so that we have been able to establish objective considerations for each of them.

From our preliminary experimentation under this homogeneous experimental framework, we have been able to determine that the use of energy weighting and distance accumulation achieves the best results (either using Euclidean distance between vectors or the first MFCC derivative).

We have also studied till what extent VFR analysis is useful in its three main application scenarios:

- We have used VFR analysis to discard feature vectors without a significant impact in recognition performance. The obtained results are similar to the ones presented in the literature and our contribution is defining a more objective baseline condition for the comparison between VFR and FFR. We have shown that when discarding up to 50% of the feature vectors, the VFR methods results do not surpass those obtained by FFR analysis.
- We have also used VFR to improve the acoustic modelling ability of the speech recogniser, calculating more feature vectors in rapidly changing speech segments. We have shown that unless a very restricted database (in terms of spectral characteristics) is used, no improvements can be expected.
- We have used VFR analysis to deal with additive noise conditions (in the time domain). We have shown that, even for a complex recognition task, the VFR techniques surpass, in a statistically significant way FFR analysis, specially for low SNRs. However, the results obtained are still far from those when matched conditions are present in training and testing.

So, to summarise, we can say that VFR analysis seems to have little room for improving speech recognition systems when applied to difficult tasks, except for the case of facing noise conditions with low SNRs. The improvements shown in the literature seem to be not statistically significant when compared with FFR analysis achievements in the same conditions.

Our main interest in this area follows the fact that VFR seems to work very well in additive noise conditions. Right now we are starting to evaluate this approach in a very difficult fast spontaneous speech recognition task: conversations between air traffic controllers and commercial plane pilots. We expect VFR to lead to some improvements given its ability to also adapt to spectral variation changes.

6. References

[1] Philippe Le Cerf and Dirk Van Compernolle. A new variable frame rate analysis method for speech recognition. *IEEE Signal Processing Letters*, 1(12), 1994.

[2] Xiaodong Cui, Markus Iseli, Qifeng Zhu, and Abeer Alwan. Evaluation of noise robust features on the aurora databases. In *Proc. ICSLP '02*, pages 481–484, Los Angeles, 2002. University of California.

[3] Javier Macías-Guarasa. *Arquitecturas y métodos en sistemas de reconocimiento automático de habla de gran vocabulario (Architectures and methods in large vocabulary speech recognition systems)*. Phd. thesis, ETSIT UPM, 2001.

[4] K.M. Ponting and S.M. Peeling. The use of variable frame rate in analysis in speech recognition. *Computer Speech and Language*, 5(2):169–179, April 1991.

[5] A. Varga, H.J.M Steenneken, M. Tomlinson, and D. Jones. The noisex-92 study on the effect of additive noise on automatic speech recognition. Documentation included in CD-ROMS of NOISEX-92, 1992.

[6] Qifeng Zhu and Abeer Alwan. On the use of variable frame rate analysis in speech recognition. *ICASSP*, pages 3264–3267, 2000.