# INTERPOLATIVE VARIABLE FRAME RATE TRANSMISSION OF SPEECH FEATURES FOR DISTRIBUTED SPEECH RECOGNITION

[1]Huiqun Deng, [1]Douglas O'Shaughnessy, [2]Jean Dahan, [2]William F. Ganong

[1]INRS-EMT, University of Quebec, Montreal, Canada
[2]Nuance Communications Inc.

## ABSTRACT

In distributed speech recognition, vector quantization is used to reduce the number of bits for coding speech features at the user end in order to save energy for transmitting speech feature streams to remote recognizers and reduce data traffic congestion. We notice that the overall bit rate of the transmitted feature streams could be further reduced by not sending redundant frames that can be interpolated at the remote server from received frames. Interpolation introduces errors and may degrade speech recognition. This paper investigates the methods of selecting frames for transmission and the effect of interpolation on recognition. Experiments on a large vocabulary recognizer show that with spline interpolation, the overall frame rate for transmission can be reduced by about 50% with a relative increase in word error rate less than 5.2% for clean and noisy speech.

**Index Terms**: Data compression, speech coding, speech recognition, interpolation.

## 1. INTRODUCTION

Modern communication technologies provide resource-limited mobile users with the ability to access remote speech recognizers and/or multilingual translation engines. In distributed speech recognition, speech features are extracted at user ends (e.g., cell phones, PDAs, etc.), and the feature information is transmitted over error-protected data channels to remote speech recognizers. A typical speech recognizer is designed to decode speech-feature vector streams at a fixed frame rate of 100 frames/second. A speech-feature vector consists of 12 MFCCs (mel-frequency cepstral coefficients) extracted from a 20-ms windowed speech signal, and each MFCC is coded in 1 byte. If a speech-feature stream is not compressed, then a data flow at a bit rate of 9.6 kb/second needs to be sent to a remote recognizer. Reducing the speech feature transmission rate is especially important for resource-limited devices to save transmission energy, and is also helpful to reduce data traffic congestion in communication networks. The challenge is to reduce the bit rate of transmission without significant loss in the speech recognition accuracy. Standard feature compression algorithms [1] for distributed speech recognition reduce the number of bits used for coding each frame's features by employing vector quantization, while keeping the frame rate of the feature stream fixed. Such compression can be viewed as lossy within-frame compression, or lossy static compression. Extracting speech features at a rate of 100 frames/second is to capture rapidly changing speech features. However, it is known that the maximal rate of phonemes produced by a human is about 12 phonemes/second. If each phoneme is represented by a 3-state HMM with each state corresponding to one frame, then a speech feature stream at an average rate of about 36 frames/second contains enough information for a speech recognizer to decode speech signals. The expected saving in frame rate is from stationary speech sounds (such as vowels and fricatives), which exhibit repetitive spectra over several consecutive frames and result in redundancy in the speech feature stream. It is also reported that if speech features are extracted at a frame rate of 100 frames/second, the average number of frames over a phoneme segment is approximately eight [2]. This means that for an HMM-based speech recognizer to decode a speech feature sequence, an average frame rate of about 37.5 frames/second (i.e., 3/8 =37.5% of 100 frames/second) is sufficient. Therefore, it is possible to transmit speech features at a variable frame rate with an average rate as low as about 37.5 frames/second to remote recognizers, without causing the recognizer to lose recognition accuracy.

Eliminating feature stream redundancy exhibited in the time domain is a dynamic compression approach. Its essential idea is to transmit more frames in fast changing regions than in relatively stationary regions, and the frame transmission rate is variable. This strategy has been used in video compressions and narrow-band LPC speech coding, where dropped frames are reconstructed by interpolation in order not to cause perceptual notice of human receivers. It is reported that if the frame rate is below 37 frames/second, distortions in the speech quality to human ears can be noticed [3]. In [4], an optimal interpolation based on frame transitional probability achieved an average frame rate of 31.1 frames/second (with end-point detection to eliminate silence segments) and obtained very high intelligibility to human listeners. It can be seen that for both human perception and machine recognition of speech signals, the minimum average frame transmission rate without significant degradation in quality of service is about 31 to 37.5 frames/second.

In automatic speech recognition, variable frame rate (VFR) has been previously introduced for two motivations: to save computation cost [5] and to capture rapidly changing speech features [6] [7]. In [5], speech features are extracted at 100 frames/second, while in [6][7], features are extracted at 400 frames/second to capture rapidly changing speech features. A feature difference between the current frame and the previous selected frame is calculated according to certain criteria such as Euclidian distance, feature derivatives, entropy, etc. If the difference is within a given threshold, the current frame is dropped for the recognition, otherwise it is selected. It is reported that if frames are selected based on feature derivatives and the recognizer is also trained using such selected frames and their derivatives (i.e., their derivatives are obtained in the compressed time domain), then the half frames can be left out without causing significant loss in recognition of isolated digits [5]. In the VFR approach to capture rapidly changing speech features, it is reported that no significant improvement is obtained for the recognition of clean speech but such occurs for noisy speech recognition [8]. Now, we introduce VFR to reduce the average frame rate in the transmission of speech features to remote speech recognizers. This paper focuses on our strategy for selecting frames for the transmission and the interpolation of frame features for recognition. Section 2 presents our frame selection method with the constraint of interpolation errors at the remote recognizer. Section 3 reports the results of applying our method on a large vocabulary speech recognizer, and compares linear interpolation with spline interpolation. Section 4 concludes the present work.

## 2. SELECTION OF FRAMES FOR TRANSMISSION

Observing the sequences of MFCCs, we found that their evolution with time exhibits relatively stationary segments interleaved with rapidly changing transients. Stationarity implies predictability and redundancy in the signal. This observation motivates us to approximate MFCC sequences using piece-wise linear or spline functions, and view frames over a time interval as redundant frames if their MFCCs can be interpolated using linear or spline functions within a threshold error. Thus, to the remote recognizer, we only need to transmit boundary frames (anchor frames), the time interval between the boundaries, and the parameters of the interpolation functions. The frames not transmitted can be interpolated from the transmitted anchor frames at the remote speech recognizer if it is trained to process MFCC streams at a fixed frame rate. This paper assumes that the recognizer is trained from feature streams of 12-dimensional MFCCs and their 1st and 2nd derivatives at a fixed frame rate 0f 100 frames/second. Selection of frames for the transmission should ensure that interpolation errors do not cause significant degradation in speech recognition rate. To that end, we select/drop frames according to interpolation errors.

Let feature vectors (MFCCs) of frames at $t_1$ and $t_1+M$ be selected for the transmission. For the receiver to linearly interpolate the 12 MFCC sequences at $t_1+1$, ……, $t_1+M-1$, the following formula can be used:

$$y^*{}_i(t_1+t) = y_i(t_1) + \frac{y_i(t_1+M) - y_i(t_1)}{M}t, \qquad 1 \le t \le M-1 \quad (1)$$

where $y_i(t_1)$ and $y_i(t_1+M)$ are the original MFCC coefficients in dimension $i$ at frames $t_1$ and $t_1+M$, respectively; $y^*{}_i(t_1+t)$ is the interpolated MFCC in dimension $i$ at frame $t_1+t$.

For the spline interpolation, the MFCCs of a dropped frame at $t_1+t$ can be interpolated by the receiver using:

$$y^*{}_i(t_1+t) = \alpha_i t^2 + \beta_i t + \gamma_i, \qquad 1 \le t \le M-1 \quad (2)$$

where $t$ is the time difference from the interpolated frame $t_1+t$ to $t_1$. The receiver needs to know the parameters $\alpha_i$, $\beta_i$, and $\gamma_i$ for 12 splines. To that end, we let the $i^{th}$ spline pass the two points (0, $y_i(t_1)$) and ($M$, $y_i(t_1+M)$). Then, the following equations hold:

$$\gamma_i = y_i(t_1) \quad (3)$$

$$\alpha_i M^2 + \beta_i M + \gamma_i = y_i(t_1+M). \quad (4)$$

Clearly, the value of $\alpha_i$ or (not and) $\beta_i$ must be provided by the sender for the receiver to construct the $i^{th}$ spline. The value of $\alpha_i$ should be determined by the sender such that the mean squared error of the interpolation is minimized, i.e.,

$$\alpha_i^* = \min_{\alpha_i} \sum_{t=1}^{M-1} [\alpha_i t^2 + \beta_i t + \gamma_i - y_i(t+t_1)]^2. \quad (5)$$

Substituting $\beta_i$ and $\gamma_i$ in Eq. (5) with expressions in terms of $a_i$ given in Eqs. (3) and (4), then Eq. (5) becomes:

$$\alpha_i^* = \min_{a_i} \sum_{t=1}^{M-1} \{\alpha_i t^2 + (\frac{y_i(t_1+M) - y_i(t_1)}{M} - \alpha_i M)t + y_i(t_1) - y_i(t+t_1)\}^2. \quad (6)$$

Then, the optimizing $a_i$ can be determined by setting the derivative of the mean squared error with respect to $\alpha_i$ to zero:

$$\sum_{t=1}^{M-1} \{\alpha_i t^2 + (\frac{y_i(t_1+M) - y_i(t_1)}{M} - \alpha_i M)t + y_i(t_1) - y_i(t+t_1)\}(t^2 - Mt) = 0. \quad (7)$$

Thus,

$$\alpha_i^* = \frac{-\sum_{t=1}^{M-1}\{\frac{y_i(t_1+M) - y_i(t_1)}{M}t + y_i(t_1) - y_i(t+t_1)\}(t^2 - Mt)}{\sum_{t=1}^{M-1}(t^2 - tM)^2}. \quad (8)$$

The interpolation error for both the linear interpolation and the spline interpolation can be measured as:

$$e_i(t_1+t) = y_i^*(t_1+t) - y_i\,(t_1+t). \quad (9)$$

Ideally, the frames transmitted should ensure that the interpolated features are within the same class as their original features. We need to find the threshold error for selecting the frames for the transmission to ensure each feature pattern after interpolation is in the same class as the original one. It is known that each speech feature pattern in the recognizer is represented using 12 MFCCs with each being normalized (scaled and offset) and coded in 256 quantization levels (8 bits). We consider that a number of interpolation errors greater than a few quantization levels can degrade the classification significantly. We set the threshold value of interpolation errors as $e_{TH}$ (in quantization level) and the number of interpolation errors greater than this threshold as $N$ over the interpolation interval. To catch transient speech features, the number of the interpolated MFCCs that have errors greater than the threshold $e_{TH}$ should be limited by:

$$N = \sum_{t=1}^{M-1}\sum_{i=1}^{I} u\left(\left|y^*_i(t_1+t) - y_i(t_1+t)\right| - e_{TH}\right) \le N_{TH} \qquad (10)$$

where $u(x)$ =1 when $x$ is greater than zero, and $u(x)$ =0, otherwise; $N_{TH}$ is the threshold of the number of interpolated MFCCs with large errors; the value of $I$ is set to 4 in this paper, because we notice that MFCCs in higher dimensions are sensitive to noise disturbance. If the value of $N$ exceeds $N_{TH,}$ $M$ should be reduced to satisfy Eq. (10). There is trade-off between frame rate reduction and recognition accuracy. Smaller values of $e_{TH}$ and $N_{TH}$ lead to less reduction in frame rate and less degradation in classification. The values of $e_{TH}$ and $N_{TH}$ are determined from experiments that yield satisfactory trade-offs between frame rate and recognition accuracy. Given $e_{TH}$ and $N_{TH,}$ the procedure for selecting frames for transmission according to the errors of linear or spline interpolation is described in pseudo code below. Lines D and E are excluded for linear interpolation.

```
Line A:  send the frame $t_1$=1
Line B:  set $M$=2 for linear interpolation, set $M$=3 for spline
Line C:  $N$=0
         if $t_1+M>$ number of total frames
Line D:      if $M>=4$, send $\alpha^*$ for interval $t_1+1$ to $t_1+M$-2
Line E:      if $M<=3$, send frame $t_1+1$
              send the last frame
              send interval value $M$-1 (not for spline $M$=3)
              stop processing
         end
         if $t_1+M <=$ number of total frames
             for $t$=1 to $M$-1
                 for $i$=1 to $I$
                     calculate $y^*_i(t_1+t)$
                     if $|y_i(t_1+t)- y^*_i(t_1+t)|>e_{TH}$
                       $N=N+1$
                     end
                 end
             end
             if $N > N_{TH}$
               do Line D and Line E for spline interpolation
               send the frame $t_1+M$-1
               send interval value $M$-1 (not for spline M=3)
               update $t_1=t_1+M$-1
               go to Line B
             end
             if $N <= N_{TH}$
               $M=M+1$
               go to Line C
             end
         end
```

The process of frame rate reduction and frame recovery for distributed speech recognition is illustrated in Fig.1, where Y(.) represents 12-dimensional MFCCs, ΔY(.) and Δ ΔY(.) represent 1st and 2nd derivatives of the MFCCs, and $Y^*(.)$ represents an interpolated MFCC vector.

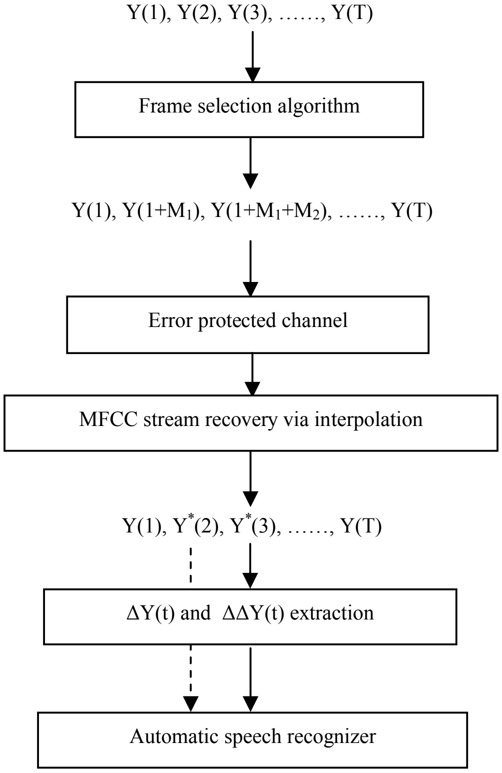
Y(1), Y(2), Y(3), ……, Y(T)
Frame selection algorithm
Y(1), Y(1+M1), Y(1+M1+M2), ……, Y(T)
Error protected channel
MFCC stream recovery via interpolation
Y(1), Y*(2), Y*(3), ……, Y(T)
ΔY(t) and ΔΔY(t) extraction
Automatic speech recognizer

Fig. 1. The frame rate reduction and frame feature recovery for distributed speech recognition.

## 3. EXPERIMENTAL RESULTS AND DISCUSSION

The effect of the above frame rate reduction and recovery via linear and spline interpolations on the final speech recognition rate is tested on a speech recognizer for clean and noisy speech signals. The clean speech signals were recordings of 10 female and 10 male talkers, and the noisy ones were telephone recordings of 7 talkers in noisy shopping malls. For each speech signal, the speech features (MFCCs) were extracted at a fixed rate of 100 frames/second. The above frame selection algorithm is then applied to the MFCC sequences to select frames for transmission. The MFCCs of the selected frames and the intervals between the selected frames are used to interpolate the MFCCs of unselected frames according to Eq. (1) or (2). Given the recovered MFCC sequences, which have 100 frames/second, the 1st and 2nd derivatives of the MFCCs are calculated from a few adjacent frames using the method as the recognizer does for regular MFCC sequences. Finally, the recovered 36-dimensional MFCC sequence is decoded by the recognizer. For different $e_{TH}$ and $N_{TH}$, the transmission frame rates determined by linear interpolation and the relative increase in word error rates produced by a large vocabulary speech recognizer are summarized in Tables I and II, for clean speech and noisy cell phone speech, respectively. For both clean and noisy speech, the relative increase in word error rate increases as frame rate decreases, and it is less than 10% when the average frame rate is about 43.9 frames/second.

An example result of frame selection and linear interpolation using the algorithm in section 2 is illustrated in Fig. 2. The frames selected for transmission are indicated using vertical lines; the original 12 MFCC sequences are plotted using (red) dots; the MFCC sequences linearly interpolated from the selected frames are plotted in solid (blue) lines. As can be seen, intervals between selected frames (i.e., the interpolation intervals) are longer for segments where the spectrum is relatively stationary than for segments where the spectrum changes quickly. It is noted that MFCC sequences in silence (noise) regions are unpredictable and the linear interpolation errors are generally large, resulting in frequent transmission of frames for unwanted silence segments. Therefore, if end point detection is combined with the frame selection, more savings in frame rate for transmission can be obtained.

A problem of the linear interpolation of MFCC sequences is that there are many zeros in the 2nd derivatives of the interpolated MFCC sequences; however, the current speech recognizer was trained from all frame features extracted at a rate of 100 frames/second without dropping and interpolation, and the 2nd derivatives used for the training are generally non-zero. Therefore, there is some mismatch between the linearly interpolated data and training data, and the recognition could be degraded due to this mismatch. To reduce the mismatch, retraining the recognizer using interpolated data is needed. In contrast, spline interpolation can provide non-zero 2nd derivatives and the mismatch can be alleviated. We test the recognition rate and overall transmission frame rate for spline-interpolated speech-feature sequences. The overall transmission frame rate has taken into account speech features (12 MFCCs coded in 12 bytes per selected frame) as well as the 12 spline parameters $\alpha$'s for 12 dimensions in each interpolation interval. Each optimizing $\alpha$ *is* coded in 1 byte, and transmitting 12 spline parameters takes 12 bytes, which we count as one transmission frame. Tables III and IV show the overall transmission frame rates and relative increase in word error rates for spline-interpolated clean and noisy speech. Results show that spline interpolation offers better

Table I. Relative increase in word error rate and transmission frame rates of linearly interpolated clean speech features
(averaged over 3048 utterances of 20 talkers)

| $e_{TH}$, $N_{TH}$ | Frame rate (frames/sec.) | Relative increase in WER (%) |
|---|---|---|
| 2, 3 | 55.98 | 4.93 |
| 3, 3 | 51.09 | 8.68 |
| 5, 3 | 43.92 | 7.89 |
| 8, 3 | 37.87 | 19.72 |

Table II. Relative increase in word error rate and transmission frame rates of linearly interpolated noisy speech features
(averaged over 256 noisy utterances of 7 talkers)

| $e_{TH}$, $N_{TH}$ | Frame rate (frames/sec.) | Relative increase in WER (%) |
|---|---|---|
| 2, 3 | 53.98 | 7.17 |
| 3, 3 | 49.9 | 10.04 |
| 5, 3 | 42.16 | 7.53 |
| 8, 3 | 32.73 | 19.71 |

Table III. Relative increase in word error rate and transmission frame rates of spline interpolated clean speech features
(averaged over 3048 utterances of 20 talkers)

| $e_{TH}$, $N_{TH}$ | Frame rate (frames/sec.) | Relative increase in WER (%) |
|---|---|---|
| 3, 3 | 70.64 | 3.35 |
| 4, 3 | 63.25 | 4.14 |
| 5, 3 | 57.32 | 3.55 |
| 5.4 | 52.62 | 3.94 |
| 5,5 | 50.18 | 5.13 |
| 5,7 | 45.64 | 11.05 |
| 6,5 | 45.86 | 9.66 |

Table IV. Relative increase in word error rate and transmission frame rates of spline interpolated noisy speech features
(averaged over 256 utterances of 7 talkers)

| $e_{TH}$, $N_{TH}$ | Frame rate (frames/sec.) | Relative increase in WER (%) |
|---|---|---|
| 3, 3 | 70.59 | 0.36 |
| 4, 3 | 63.00 | 0.36 |
| 5, 3 | 56.50 | 0.72 |
| 5.4 | 51.84 | 3.23 |
| 5,5 | 49.20 | 1.08 |
| 5,7 | 44.63 | 3.94 |
| 6,5 | 44.47 | 3.94 |

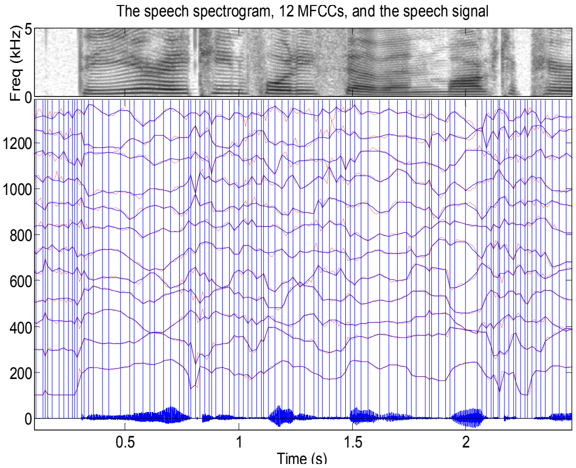

Fig. 2. The clean speech signal (bottom) of "The year eighteen forty seven marked ……" produced by a female talker, the original 12 MFCC sequences (red dots), the selected frames (vertical lines) and the interpolated MFCC sequences (solid blue curves) determined by the linear interpolation with $e_{TH}$=3, $N_{TH}$ =3.

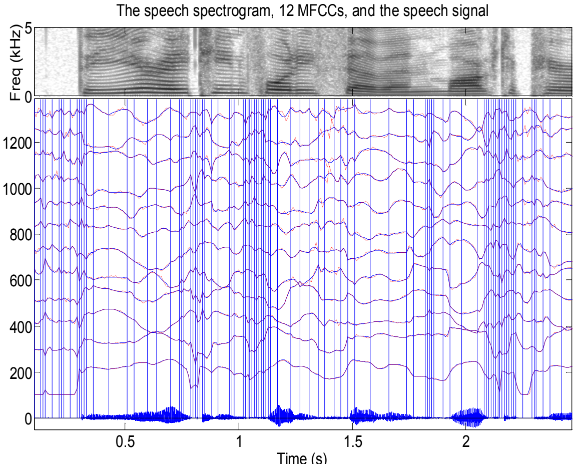

Fig. 3. The same as Fig. 2, except that the interpolated MFCC sequences (blue solid curves) and selected fames (blue vertical lines) are determined by the spline interpolation with $e_{TH}$=3, $N_{TH}$=3.

trade-offs between recognition rate and transmission frame rate than linear interpolation does. An example of spline interpolated MFCC sequences are displayed together with their original ones in Fig. 3. It is interesting to note that the spline interpolation led to much denser frame selection for transient segments than for silence and other speech segments. Such a distinction is not so obvious for linear interpolation. The overall bit rate for transmitting speech feature streams is determined by the bit rate for transmitting MFCCs, interpolation parameters, and interpolation intervals. It is found that each interpolation interval can be coded in 4 bits. Then, the overall bit rate for transmitting speech feature streams at an overall frame rate of 50 frames/second is equal to (for linear interpolation) or less than (for spline interpolation) 50×(12×8+4)= 5 kb/second.

## 4. CONCLUSION

For distributed speech recognition, the transmission frame rate reduction and recognition rate determined by linear interpolation and spline interpolation are investigated. Spline interpolation leads to better trade-offs between recognition rate and frame transmission rate than the linear interpolation does. At about half the standard frame transmission rate, the relative increase in error rate is less than 3.23% for noisy speech, and is 5.13% for clean speech. Future work will be to combine dynamic compression with static compression to achieve an even lower bit rate without significant loss in recognition rate.

## 5. REFERENCES

[1] ESTI ES 202 212 V1.1.2 (2005-11), Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended front-end feature extraction algorithm: Compression algorithms; Back-end speech recognition algorithm.

[2] Wu, T., Compernolle, D. V., Duchateau, J. and Hamme, H. V. "Single Frame Selection for Phoneme Classification," Proceedings of Interspeech, Pittsburgh, Pennsylvania, Sept. 17-21, 2006, pp. 641-644.

[3] Viswanthan, V. R., Makhoul, J. and Schwartz, R., "Variable Frame Rate Transmission: a Review of Methodology and Application to Narrow-Band LPC Speech Coding," IEEE Transactions on Communications, Vol. 30, No. 4, Apr. 1982, pp. 674 – 686.

[4] Chung, C. and Chen, S., "Variable Frame Rate Speech Coding Using Optimal Interpolation," IEEE Transactions on Communications, Vol. 42, No.6, June 1994, p. 2215-2218.

[5] Le Cerf, P. and Van Compernolle, D., "A New Variable Frame Rate Analysis Method for Speech Recognition," *IEEE Signal Processing Letter*, Vol. 1, No. 12, pp.185-187, Dec. 1994.

[6] Zhu, Q. and Alwan, A., "On the use of variable frame rate analysis in speech recognition," IEEE ICASSP 2000, Vol. 3, pp. 1783 – 1786.

[7] You, H., Zhu, Q. and Alwan, A., "Entropy-based variable frame rate analysis of speech signals and its application to ASR," IEEE ICASSP 2004, vol., pp. 549-52.

[8] Macias-Cuarasa, J., Montero, J. M., Ferreiros, J., Cordoba, R. and D'Haro, L. F., "Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition," Eurospeech 2003, Geneva, pp. 1809-1812.