TOWARDS SPEECH RATE INDEPENDENCE IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

F. Martínez, D. Tapias and J. Álvarez (e-mail: daniel@craso.tid.es)

Speech Technology Group
Telefónica Investigación y Desarrollo, S.A. Unipersonal
C/ Emilio Vargas, 6; 28043 - Madrid (Spain)

ABSTRACT

In this paper we present a new speech rate classifier (SRC) which is directly based on the dynamic coefficients of the feature vectors and it is suitable to be used in real time. We also report the study that has been carried out to determine what parameters of speech are the best regarding the speech rate classification problem. In this study we analyse the correlation between several speech parameters and the average speech rate of the utterance. Finally, we report a compensation technique, which is used together with the SRC. This technique provides with a word error rate (WER) reduction of a 64.1% for slow speech rate and a 32% reduction of the average WER.

1. INTRODUCTION

It is well known that the performance of LVCSR systems dramatically degrades when the speech rate is different than normal. We have checked out that the WER for slow and fast speech increases up to 2.8 times on average with respect to the WER at average speech rate. This is due not only to the mismatch between the training and the testing conditions but also to other factors (studied in our previous work [1]) like phone elision or weakening, phone duration reduction, aspiration, transient nature of the fast speech spectra, etc. The phenomenon of speech rate variation is very common: we have checked out that the rate of speech can significantly change along a spontaneous speech utterance or even in a read sentence. Furthermore, we have found out that in real applications, when the sentence is misrecognised, users are used to repeat the sentence very slowly to make it more understandable. Besides, there are significant speech rate differences among some of the different dialects of Spanish. Finally, there is also inter/intra-speaker speech rate variability. It is, therefore, necessary to introduce compensation techniques as well as unsupervised speech rate classification methods suitable to be used in real time to overcome this problem.

Several unsupervised speech rate measurements and classification methods have been proposed recently [3][4][5]. The first one is based on estimating phone boundaries by means of a multi-layer perceptron. The second relies on the speech recogniser output and the third directly processes the speech signal by measuring the variation of the energy envelope of speech. All of them provide with promising results on databases like WSJ, TIMIT, OGI and Switchboard. However, these databases were not designed to study the speech rate phenomenon so that they may have few examples of both slow and fast speech (especially at extreme speech rates) and, consequently, typical characteristics of high or low ROS might be poorly represented. Therefore, we strongly believe that it is necessary to have a specific database to study the speech rate phenomenon and get reliable evaluation of both speech rate classifiers and compensation techniques. For this reason, all the experiments and conclusions we report have been carried out using the TRESVEL database [1].

The goal of this research work was to develop a real time speech rate classifier and compensation technique. This goal was reached by first looking for correlations between different speech related parameters and the average speech rate since we also believe [5] the measure of speaking rate should be based on the speech signal rather than on the speech recogniser output or on phone boundaries. Therefore, in Section 2 we show the results we have obtained in this study, which is divided into three parts: (a) analysis of the evolution of the ROS within the utterances, (b) study of the relation of both the average pitch and the pitch discontinuities rate (PDR) with the average speech rate and, (c) analysis of the feature vector dependency on the speaking rate. Section 3 describes the new unsupervised speech rate classification method and its experimental results. Section 4 describes the proposed compensation technique and compares the experimental results with the baseline and the compensation techniques tested in our previous work [1]. Finally, in section 5 we present our conclusions and future work in this area.

2. SPEECH PARAMETERS VERSUS SPEAKING RATE

In our previous work, we presented a new supervised measure for the speech rate and reported the characteristics of slow, average and fast speech as far as phone duration, spectra and phonetic changes is concerned, since our goal was to gather information about the speech rate phenomenon. Our current goal is to obtain a reliable ROS classification method based on the speech signal and suitable to be used in real time. Hence, it will be possible for us to use in real applications either the speech rate compensation techniques reported in [1] or the one reported in this paper. In this section we report the last results of the study that has been done to look for a reliable ROS classification method. The section is divided into three parts: local speech rate analysis, pitch versus speaking rate, and feature vector dependency on the speaking rate.

The experiments were done using the TRESVEL database, which was designed to study, evaluate and compensate the

effect of speech rate on LVCSR systems. It is composed of 9600 utterances (3200 utterances for each speech rate), which were pronounced by 40 speakers (20 female and 20 male). It allows to compare every sentence at the three ROS since each speaker uttered 80 sentences at slow, average and fast speech rate.

2.1.- LOCAL SPEECH RATE ANALYSIS

This analysis was carried out in order to determine the speech rate range of variation within an utterance and its trend, and try to find out whether there is a correlation between these factors and the average ROS.

To compute the local speech rate, we first used forced alignment to determine the phone segmentation and then we measured the local speech rate by using windows of a variable number of frames (five phone windows) in order to make sure that all the windows contained enough phones to compute accurate average ROS values. This measure was repeated every two phones by using the formula reported in [1]. The automatic phone segmentation was obtained using the correct transcriptions in order to get reliable measures of the phone rate. Figure 1 shows an example of the speech rate variation within the utterance together with its linear regression.

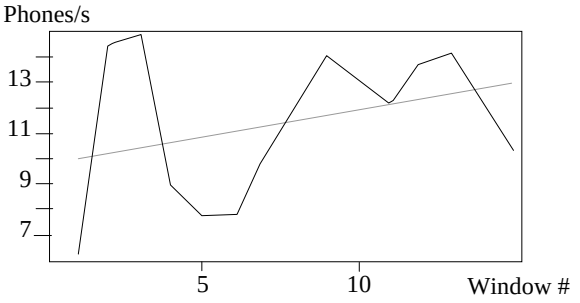

Figure 1: Example of the ROS variation within the utterance

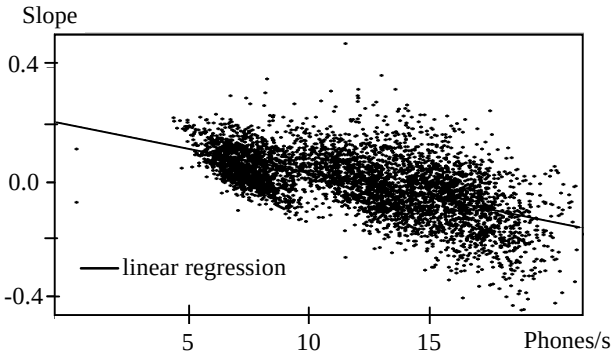

Figure 2: Slope of the ROS variation linear regression versus the speech rate.

The local speech rate analysis shows that the standard deviation of the ROS increases almost linearly with the average speech rate. The analysis also suggests that the range of variation of the ROS is not only related with the average ROS but also with other variables such as the state of mind of the speaker, the intonation and the part of the sentence the speaker wants to emphasize. Nevertheless, we found a correlation between the trend of the speech rate and the average ROS: We computed the linear regression for the local speech rate variation curve of each utterance and then obtained the slope of each line. From this study it turned out that slow speech rates usually have positive slopes, so that the ROS has the tendency to increase along the utterance, while fast speech rates usually have negative slopes and then the ROS tends to decrease along the utterance.

Table 1

% positive slope	slow	average	fast
female	79	43	34
male	64	61	32

Figure 2 represents the slopes of the linear regressions of the ROS evolution within the utterance versus the average speech rate. It also shows its linear regression that takes positive values for slow speech and negative for fast speech. This fact is clearly shown in table 1, which represents the percentage of positive slope occurrence for both female and male speakers at three speaking rates. The correlation between the slope and the average ROS in TRESVEL suggest that it could be used as a confidence measure for the first search pass in order to validate both the hypothesised sentences and the speech rate compensation technique applied in the recognition process.

2.2.- PITCH VERSUS THE SPEAKING RATE

The first experiments tried to find out whether there is a relation between the average ROS and the average pitch in the TRESVEL database. We did it by comparing the average pitch for the same sentences uttered at the three speech rates. The speaker dependent and independent experimental results showed that apparently there is no relation between both parameters.

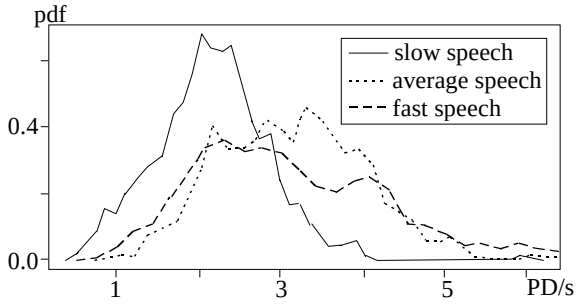

Figure 3:Probability density function of the pitch discontinuities (PD) rate.

We later studied the correlation between the pitch discontinuities rate (PDR) and the average speech rate. The PDR was defined as the number of voiced-unvoiced transitions per unit of time. Figure 3 shows the probability density function of the PDR for each speech rate. The correlation between average ROS and PDR is very low, just 0.16. However, there are just some values of the PDR which are only

reached by slow or fast speech and some ranges of variation that are reached by average and fast speech or by slow and average speech. Consequently, a ROS classifier could not be based on just PDR measures though for some speech rate values it could certainly help to get a reliable confidence measure for the classifier output.

2.3.- FEATURE VECTOR DEPENDENCY ON THE SPEAKING RATE

The goal in this section was to determine the stream of the feature vector which is most affected by the speaking rate since it will be the best parameter set to be used by the speech rate classifier. The feature vector of the speech recogniser is composed of 51 MFCCs which are divided into four streams. Each stream is modeled separately by means of 256 multivariate gaussians and then each SCHMM has four sets of 256 weights.

The study was done by adapting the SCHMMs for slow and fast speech and then evaluating the entire adapted models and each part of them separately. The adaptation of the models was carried out by using a subset of 1000 utterances for each speech rate and applying the Baum-Welch algorithm. The models were later smoothed using deleted interpolation to increase the robustness of the models since they may suffer from sparse data problems. The testing set, of approximately 1300 utterances per speech rate, was randomly chosen from the TRESVEL database and does not contain any utterance of the training set.

Figure 4 shows the experimental results for slow speech: Full adaptation (FA) of all the parameters of the models provides with the largest WER reduction. Cepstral stream adaptation (CA) as well as transition probabilities adaptation (TPA) are the ones which provide with the smallest WER reduction. Just as expected, the adaptation of the dynamic coefficients provides with better results than CA and TPA. In particular, the delta stream adaptation (DA) reduces the WER a 42.5% and the delta-delta stream adaptation (DDA) a 41%. The adaptation of both streams (D&DDA) provides with a 65.4% reduction of the WER. This fact clearly indicates that both streams are the most affected by the speech rate and therefore the most useful to perform the ROS classification.

The experimental results for fast speech show an improvement of the speech recogniser performance especially with the delta and delta-delta streams adaptation method. However, the WER reduction is smaller than for slow speech, mostly at very high speech rates (more than 15 phones/s). There are three reasons for this: (1) some triphone models cannot be properly time-aligned with the speech signal because the duration of some phones is lower than the minimum one allowed by the current HMM topology, (2) the difficulty to accurately predict phonetic phenomena like phone elision, (3) the inappropriate phonetic transcriptions of the dictionary which, in some cases, do not take into account fast speech phonetic phenomena.

The rationale for DA and DDA to behave better than CA and TPA for slow and fast speech is that they capture the dynamic characteristics of the speech spectra. Given that the duration of spectral stable regions decrease with the speech rate [1] (at fast speech rates almost there are no stable regions), the dynamic coefficients will be different for slow, average and fast speech rates. Therefore the adaptation process will correct the mismatch between the average speech rate trained models and the characteristics of slow and fast speech of the testing set.

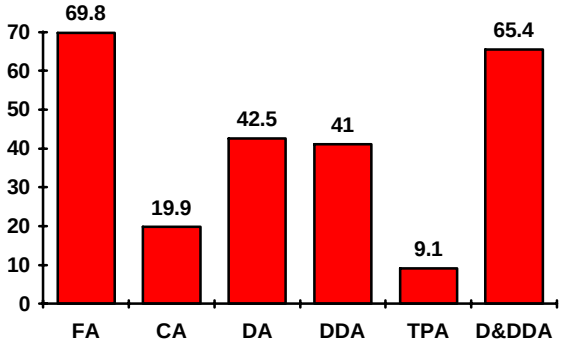

Figure 4: Percentage of WER reduction for slow speech rate

3. SPEECH RATE CLASSIFIER

In the previous section we have shown that the pitch discontinuities per second as well as the slope of the linear regression of the within utterance ROS are slightly related to the average ROS. Hence, they cannot be used to classify the speech rate of the incoming utterance, though they could certainly help to compute confidence measures of the classifier output. The classification procedure we propose in this paper is based on the dynamic coefficients of the feature vector, since we have shown that both delta and delta-delta coefficients are the most affected by speech rate changes and therefore, are the most appropriate to perform the classification. The classification method is based on a gaussian classifier which just need some frames of speech to determine whether the utterance is slow, normal or fast.

The classifier was trained with a subset of the TRESVEL database composed of 3100 utterances approximately. The testing set was randomly chosen from the same database and it is composed of 3800 utterances which do not belong to the training set.

Experimental results show that the delta and delta-delta cepstra based classifiers are able to distinguish fast, slow and average speech from each other with an utterance classification accuracy of 80% for slow speech and 70% for fast speech, what is a good and promising result if we take into account the complexity of the testing set.

4. SPEECH RATE COMPENSATION

A compensation technique has been tested, which is based on the use of speech rate dependent models (SRDM) together with the speech rate classifier (SRC). The experiments have been carried out using the speech recogniser of the ATOS conversational system [2], which vocabulary size is about 4700 words.

Two different scenarios were tested: (1) ideal SRC and SRDM and, (2) our proposed SRC and SRDM. Figure 5 presents the results obtained with scenario (1) and compare them with both the best compensation technique presented in [1], LMPW&TPA, and the baseline system, which was trained for average speech rate. In particular, figure 5.a shows the results for slow speech and figure 5.b the results for fast speech.

We have checked out that the WER of the baseline system for slow and fast speech increases up to 2.8 and 2.5 times respectively with respect to the WER for average speech rate. The baseline as well as both scenarios were evaluated with the same testing sets described in section 3.

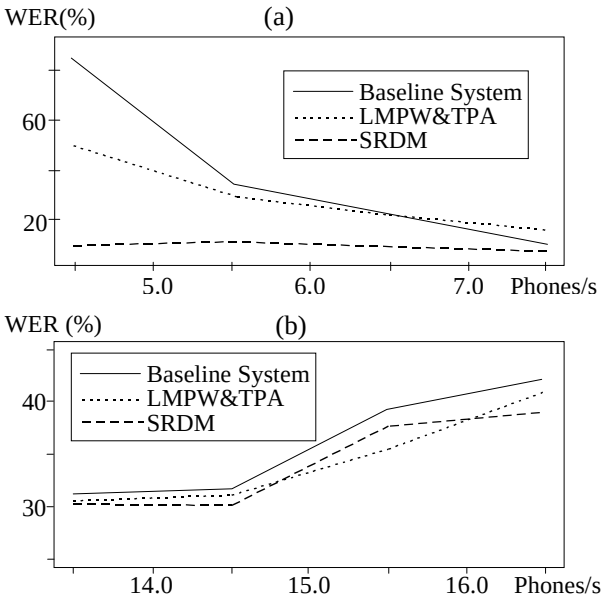

Figure 5: Word error rates for (a) slow speech rate, and (b) fast speech rate.

The results obtained with scenario (2) are a slightly worse than those of scenario (1) for slow and average speech rate: 2.2% reduction of the word accuracy for slow speech and 4.3% for normal speech. Nevertheless, scenario (2) provided with a 8.3% word accuracy improvement for fast speech. The reason for this improvement is that some of the problematic utterances were classified in the normal speech group and then the word accuracy for normal speech slightly decreased.

Despite the worse performance of scenario (2), there is still an important improvement: a 64.1% WER reduction for slow speech and a 19.2% for fast speech at the expense of a small WER increase for normal speech. The average WER for the three speech rates (slow, average and fast) is reduced by a 32% with respect to the baseline system.

5. CONCLUSIONS

In this paper we have reported the relation between different speech parameters and the average speech rate of the utterance. We have shown that a speech rate classifier (SRC) based on the dynamic coefficients provides with very good results and can be used to adapt the speech recogniser to the speech rate with a slight increase of the average word error rate (1.6%) with respect to the ideal case where all the utterances are correctly classified.

The use of speech rate dependent models (SRDM) together with the SRC reduces the word error rate a 64.1% for slow speech and a 19.2% for fast speech at the expense of a small WER increase for normal speech.

Our future work will concentrate on the improvement of both the speech rate classifier and the compensation techniques. We will also focus on the modification of the decoding algorithm to deal with speech rate related phenomena like the phone elision problem.

6. ACKNOWLEDGMENTS

We want to thank Alex Acero (Microsoft) and Pedro Moreno (Digital) for their valuable suggestions in this work. We are also grateful to the Speech Recognition Group of Telefónica Investigación y Desarrollo for their support.

7. REFERENCES

[1] F. Martínez, D. Tapias, J. Álvarez and P. León, “Characteristics of Slow, Average and Fast Speech and Their Effects in Large Vocabulary Continuous Speech Recognition”, In Proc. of Eurospeech’97, Rhodes, Greece, September 1997.

[2] J. Álvarez, D. Tapias, C. Crespo, Y. Cortazar and F. Martínez,” Development and Evaluation of the ATOS Conversational System”, In Proc. of ICASSP’97, Munich, April 1997.

[3] J. P. Verhasselt and J. P. Martens, “A Fast and Reliable Rate of Speech Detector”, In Proc. of ICSLP’96, Philadelphia, USA, October 1996.

[4] M. A. Siegler and R. M. Stern, “On the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition”, In Proceedings of ICASSP’95, Detroit, May 1995.

[5] N. Morgan, E. Fosler and N. Mirghafori, “Speech Recognition Using On-line Estimation of Speaking Rate”, Eurospeech’97, Rhodes, Greece, September 1997.

[6] T. J. Hazen and J. Glass, “A Comparison of Novel Techniques for Instantaneous Speaker Adaptation”, In Proc. Eurospeech’97, Rhodes, Greece, September 1997.