CHARACTERISTICS OF SLOW, AVERAGE AND FAST SPEECH AND THEIR EFFECTS IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

F. Martínez, D. Tapias, J. Álvarez, P. León
Speech Technology Group
Telefónica Investigación y Desarrollo, S.A.
C/ Emilio Vargas, 6 28043 - Madrid (Spain)
Tel. +34 1 337-42-52, FAX: +34 1 337-42-02, E-mail: daniel@craso.tid.es

ABSTRACT

In this paper we report the characteristics of slow, average and fast speech. The study has been done using the TRESVEL Spanish database. It is composed of 3200 sentences uttered at three different speech rates and contains speech material from 20 male and 20 female speakers. This database has been designed to study, evaluate and compensate the effect of speech rate in Large Vocabulary Continuous Speech Recognition (LVCSR) systems. We report a new measure for the rate of speech (ROS). The ROS is normalised using an appropriate set of constants that depends on the expected duration of each phone. We also report the characteristics of slow, average and fast speech. Finally, we report the degradation in performance of a continuous speech recognition system when the speech rate is low and high, and the evaluation of two compensation techniques. Adaptation of the language weight, insertion penalties and HMM state-transition probabilities for slow speech provides a 21.5% reduction of the word error rate (WER).

1. INTRODUCTION

The speech rate within a dialogue varies both globally and locally among speakers due to various factors like emotion [2], emphasis, lexical stress, dialogue status, etc. This variation dramatically affects the performance of LVCSR systems, as double to triple word error rates for fast and slow speakers.

We have found out that in real applications, if the sentence is misrecognized, users are used to repeat the sentence very slowly to make it more understandable. Consequently, as all the components of the system are adjusted to do speech recognition at the average speech rate, the word accuracy dramatically degrades. Therefore, some compensation mechanism has to be used to make speech recognizers robust against variations of the ROS.

Several studies have been done in other research groups [3][4][5] to deal with this problem. All of them propose different methods to measure the ROS as well as approaches to adapt the speech recognizer to the ROS. The databases used to do evaluations in these research works were recorded at normal speech rate (density b of figure 1) and the experiments for unusual rates were carried out extracting from this data those speakers that spoke faster or slower than the average. There is a study for Japanese [6] where a database of normal, slow and fast speech is used, but the database is small and the study was only carried out for male speakers. We strongly believe that it is necessary to have a database containing enough examples of all the ROS in order to determine the characteristics of the sounds at different speech rates (densities a and c of figure 1) and get reliable evaluations of the compensation techniques. For this reason, the first goal of our study was the design and collection of the TRESVEL database, which is described in section 2.

Section 3 describes a new measure of the ROS that has been developed to make the measure independent of the set of phones that compose the sentence. Section 4 describes a summary of the study that was carried out to determine the characteristics of slow, average and fast speech that affect speech recognition accuracy. In section 5 we present the results of the study about phone duration at different speech rates. Section 6 presents the experimental results of the baseline system evaluation and the performance of the compensation techniques. Finally, in section 7 we present our conclusions.

2. THE TRESVEL DATABASE

The TRESVEL Spanish database has been designed to study, evaluate and compensate the effect of speech rate on LVCSR systems. It is composed of slow, average and fast speech. The speakers were asked to speak normally, fast and slow, so that we could: (a) characterise the acoustic properties of sounds like duration, coarticulation effects,... for each speech rate, and (b) determine other aspects like sound relaxation or deletion and speech rate boundaries.

The database is composed of 1400 different sentences containing telephone and driving license numbers, amounts and spontaneous speech sentences. There are 40 speakers (20 male and 20 female) and each speaker uttered 80 sentences at three different speech rates (slow, average and fast), so that there is a total of 9600 sentences (3200 sentences for each speech rate).

Figure 1 shows the probability density function of the speech rate for the three different cases. As it can be seen, the ROS range of variation for each speech rate is

very large and the three density functions overlap, what proves the lack of consensus among speakers on what we subjectively call slow, average and fast speech rate.

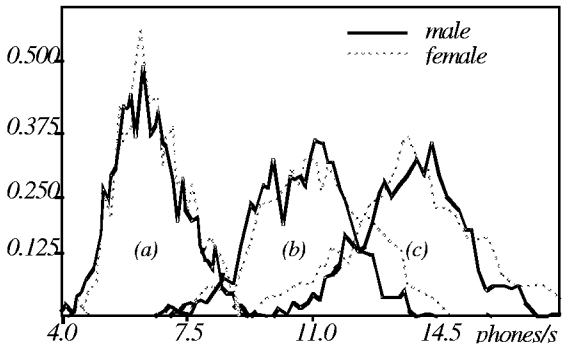

Figure 1: Probability density functions of the speech rate for:
a) low speech rate sentences
b) average speech rate sentences
c) high speech rate sentences

3. MEASURE OF THE RATE OF SPEECH

It has been pointed out that phone rate, excluding silence periods, is a reliable measure for the ROS. In this sense, several measures for the phone rate like Inverse of Mean Duration (IMD) [4] or Mean of Rates (MR) [3] have been reported.

Our experiments show that these measures perform well in many cases, but they have a drawback: They can provide different values of the ROS for two sentences uttered at the same speaking rate. The reason for this is that the expected duration of a phone is different from one phone to another, so that the ROS depends on the set of phones that compose the sentence. We have defined a new ROS measure which is more independent of this fact:

$$r = \frac{N}{\sum_{i=0}^{N-1}\left[\frac{d_i}{E[d_i]}\cdot\overline{d}\right]}$$

where N is the total number of phones, d_i is the duration of the ith phone of the utterance, $E[d_i]$ is the expected duration for the ith phone and $\overline{d}$ is the mean duration of a phone. Both, $E[d_i]$ and $\overline{d}$ are obtained using a large training spontaneous speech database.

4. CHARACTERISTICS OF SLOW, AVERAGE AND FAST SPEECH

The study of the characteristics of slow, average and fast speech has been carried out by listening to the TRESVEL database files and examining their waveforms and spectrograms. In general, it can be stated that slow speech is used to be properly pronounced while average and fast speech are used to have effects like phone elision or weakening, aspiration, assimilation, assibilation, etc.

There are two ways speakers are use to talk when they are asked to speak slowly: (a) increasing the duration of the phones keeping cross-word articulation (no silences between words), and (b) introducing long silences between words and slightly increasing the phone duration. We have checked out that in both cases phones were carefully pronounced.

Phone elision is more common in fast than in average speech though there are cases like the intervocalic /d/ which is used to be either elided or weakened at both speech rates, mostly in the verbal desinence /-ado/. It has been also observed that fast speech is basically a sequence of transitions from one sound to the next sound, so that the percentage of stable regions in the spectrogram is low in comparison with the same percentage at the average speech rate. We believe this transient nature of the spectra is one of the reasons why the acoustic models behave poorly for fast speech. This effect can be observed in figure 2, where it is shown the spectrogram for the [rríaa] sound group at low, average and high speech rate.

Vowel and vowel sequences are also affected by the speech rate; there are cases where a vowel is elided ([oa] > [a]: quiero añadir... > quierañadir), cases where two vowels are transformed into a different one ([ue] > [o]: pues > pos), and cases where are vowel is assimilated ([e] > [e~]: mensaje > mesaje).

We have also observed that very often the affricate [tS] becomes fricative in fast speech while it happens less frequently in average speech and very rarely at low speech rate.

Consonant groups are affected by the speech rate in many different ways. For example, the /ns/ group is used to become /s/ (mensaje > mesaje) and the /sT/ group is used to become /rT/ (doscientos > dorcientos).

We have found out many other cases where phones are affected by the speech rate, but we are not going to report them as it would require a extensive description which would not fit into this paper. Nevertheless, we would like to point out that the important fact of these effects is that some of them cannot be properly modelled by triphones since a particular triphone can be pronounced in different ways depending on its near contexts and the speech rate.

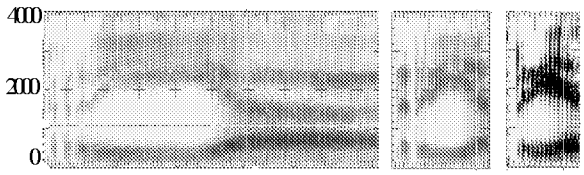

Figure 2: spectrograms of sounds [rríaa] pronounced at slow, average and fast speech rate.

5. PHONE DURATION

Automatic and accurate computation of the phone rate and phone duration requires the correct transcription for the utterance. To compute the phone duration, we first

used forced alignment to determine the phone segmentation and then both phone duration and phone rate are obtained from this segmentation.

The duration of phones is related to the lexical stress, the contexts of the phone, the position of the phone inside the word, its phonetic properties and the speech rate of the sentence.

Figure 3 shows the duration of phone "a" in milliseconds as a function of the speech rate. The duration mean as well as the standard deviation decrease as the ROS increases. Our experiments show that this behaviour is common for all the phones.

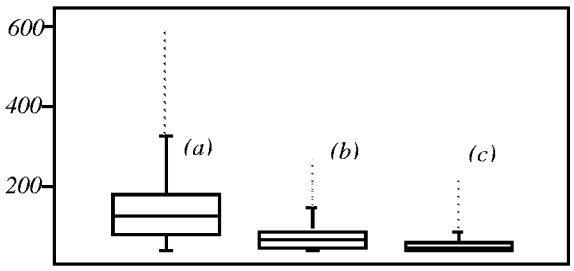

Figure 3: "A" Vowel duration for (a) slow, (b) average, and (c) fast speech.

The phone duration study was done for the Spanish basic phone set, which was later divided into six groups since the average duration for phones belonging to the same group was very similar. The groups that were studied are: vowels, fricatives (fricativ.), voiced plosives (v. plos.), unvoiced plosives (u. plos.), voiced consonants (v. cons.) and affricates (affric.).

Table 1 shows the average duration of each group for slow, average and fast speech rate as well as percentages of duration reduction with respect to the duration of slow speech.

It can be inferred from the analysis of table 1 that there are three kinds of duration behaviour in terms of percentage of duration reduction: (a) Vowels, which are the most affected by the speech rate: 61.6% duration reduction on average for fast speech rate and 47.5% for average speech rate. (b) Fricatives, voiced plosives, voiced consonants and affricates: their duration is reduced about 32% at average speech rate and about 49% at fast speech rate. (c) Unvoiced plosives, which are by far the least sensitive sounds to the speech rate.

We have checked out the results presented in Table 1 by manually segmenting a reduced set of sentences; even though there are slight differences in the average duration of the sounds (mostly for fast speech rate), the percentages related to the duration reduction are basically the same.

6. EXPERIMENTS AND RESULTS

The speech recognition experiments have been carried out using the speech recognizer of the ATOS conversational system [1], which vocabulary size is about 4700 words.

TABLE 1: Average duration for slow, average, and fast speech (ms) and percentage of duration reduction

	SLOW	AVERAGE		FAST	
vowels	135	71	47.5 %	51	61.6 %
fricativ.	138	93	32.6 %	69	49.9 %
v. plos.	90	64	29.3 %	48	46.8 %
u. plos.	112	93	18.1 %	72	36.0 %
v. cons.	123	82	33.8 %	61	50.3 %
affric.	179	116	35.0 %	89	50.5 %

The first experiment evaluated the performance of the baseline system for the three speech rates. The analysis of results showed that 62% of errors for fast speech are due to substitutions, 31% to deletions and 7% to insertions. These results indicate that the model parameters and the HMM topology are not appropriate for fast speech since the percentages of deletions and substitutions are very high. Concerning the case of slow speech, 42% of errors are due to substitutions, while 55.8% are due to insertions and 2.2% to deletions. These results show that even though the parameters of the models are not the most appropriate for slow speech rate, the large amount of insertions together with the lack of balance between insertions and deletions are more important problems.

Two compensation techniques and their combination were tested to reduce the effect of the above mentioned problems:

(a) The first compensation technique, Language Model Penalty and Weights Adaptation (LMPWA), tries to deal with the lack of balance between insertions and deletions for both slow and fast speech: the language weight for each search pass together with the word insertion penalty help to control the percentage of insertions and deletions, so that some experiments were carried out to optimise these parameters for each speech rate. The results show a 15.4% reduction of the word error rate (WER) for slow speech and just a 1.4% reduction for fast speech.

(b) The second compensation technique, Transition Probabilities Adaptation (TPA), modifies the HMM state-transition probabilities to adapt them to fast and slow speech. This idea has been previously tried by other authors [3][4] and our experiments confirm its usefulness. Based on this idea, we have developed two different approaches to do TPA getting an additional 13.6% improvement for slow speech and 5.9% for fast speech in terms of WER reduction.

In the first approach, we adapted the HMM state-transition probabilities to fast/slow speech by reducing/increasing the probability of remaining at the same state (a_{ii}) using the following equations for slow and fast speech rate respectively:

$$a_{ii}^{s} = a_{ii}^{a} + \lambda^{s}\left(1 - a_{ii}^{a}\right),$$

$$a_{ii}^{f} = \lambda^{f} \cdot a_{ii}^{a}$$

where λ^{s} and λ^{f} are constants determined empirically, a_{ii}^{s}, a_{ii}^{f} and a_{ii}^{a} are the probabilities of

remaining in state "i" for slow, fast and average speech rates respectively.
The other state-transition probabilities, i.e. $a_{i,i+1}$ and $a_{i,i+2}$ were adjusted proportionally to their relative importance.
Although this method improved the performance of the baseline system, a more accurate HMM state-transition probabilities adaptation approach was tried. In this second approach, the transition probabilities of each phone are adapted separately. This is carried out by taking into account the phone duration study of section 5. The equations used to get the new state-transition probabilities for slow and fast speech are:

$$a_{ii}^{s,p} = a_{ii}^{a,p} + \lambda_p^s \cdot (1 - a_{ii}^{a,p})$$

$$a_{ii}^{f,p} = \lambda_p^f \cdot a_{ii}^{a,p}$$

where the index "p" is the phone identification.
The constants λ_p^s and λ_p^f are different for each phone and are proportional to their duration reduction at each speech rate.
As before, the other state-transition probabilities, i.e. $a_{i,i+1}$ and $a_{i,i+2}$ were adjusted proportionally to their relative importance.
This approach for TPA performs better than the previous approach for slow speech rate. Nevertheless, it did not work so well for fast speech rate. We believe there are two main reasons for this method to behave poorly at fast speech rates: (1) some triphone models cannot be properly time-aligned with the speech signal since the duration of some phones is lower than the minimum one allowed by the current HMM topology, and (2) the difficulty to accurately predict phonetic phenomena like phone elision, that reduce the duration of the utterance. For these reasons the first TPA approach was used for fast speech rate while the second one for slow speech rate.
(c) Finally, our experiments show that the combination of both compensation techniques, LMPW&TPA, outperforms the results obtained using each method separately: the WER was reduced a 21.5% for slow speech rate and a 7.8% for fast speech rate.
Figure 4, represents the WER reduction for the reported experiments.

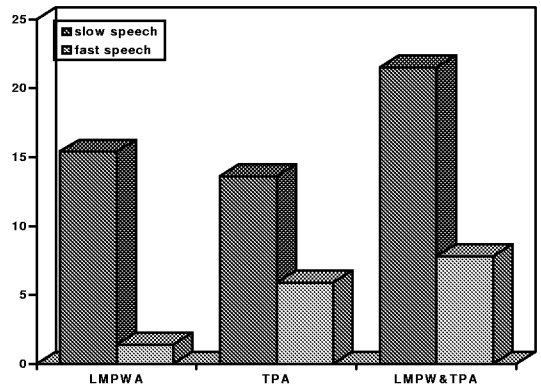

Figure 4: Percentage of the WER reduction

7. CONCLUSIONS

In this paper we have presented the results of the study about the characteristics of slow, average and fast speech, their effects in large vocabulary continuous speech recognition and two compensation techniques.
We have shown that two simple methods can reduce the WER a 21.5% at slow speech rate and a 7.8% at fast speech rate.
There are two main conclusions in this work:
(a) The main reasons why WER dramatically increases at fast speech rates are: the transient nature of the fast speech spectra, the difficulty to accurately predict phonetic phenomena like phone elision as well as the time-alignment mistakes made by the speech recognizer when the duration of sounds is smaller than the minimum duration allowed by the triphone models.
(b) Some of the phenomena observed for both slow and fast speech cannot be properly modelled by triphones since a particular triphone can be pronounced in different ways depending on its near contexts and the speech rate.

ACKNOWLEDGEMENTS

We want to thank Matthew Siegler (CMU) for providing us with a copy of his Master thesis "Measuring and Compensating for the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition".
We are also grateful to the Speech Recognition Group of Telefónica Investigación y Desarrollo for their contributions and suggestions in this work.

REFERENCES

[1] J. Álvarez, D. Tapias, C. Crespo, I. Cortazar and F. Martínez, "Development and Evaluation of the ATOS Spontaneous Speech Conversational System", ICASSP'97, Munich, April 1997.
[2] J. Vroomen, R. Collier and S. Mozziconacci, "Duration and Intonation in Emotional Speech", Proceedings of EUROSPEECH'93, vol. 1, pp. 577-580.
[3] M. A. Siegler, and R. M. Stern, "On the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition Systems", Proceedings of ICASSP'95, pp. 612-615, Detroit, May 1995.
[4] N. Mirghafori, E. Fosler and N. Morgan, "Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes", Proceedings of EUROSPEECH'95, pp. 491-494, Madrid, September 1995.
[5] N. Mirghafori, E. Fosler and N. Morgan, "Towards Robustness to Fast Speech in ASR", Proceedings of ICASSP'96, Atlanta, April 1996.
[6] H. Kuwabara, "Acoustic Properties of Phonemes in Continuous Speech for Different Speaking Rate", Proceedings of ICSLP'96, pp. 2435-2438, Philadelphia, October 1996.