ON THE CHARACTERISTICS & EFFECTS OF LOUDNESS DURING UTTERANCE PRODUCTION IN CONTINUOUS SPEECH RECOGNITION

D. Tapias, C. García , C. Cazassus[1]
Telefónica Investigación y Desarrollo, S.A. Unipersonal
C/ Emilio Vargas, 6
28043 - Madrid - SPAIN

ABSTRACT

We have checked out that, in speech recognition based telephone applications, the loudness with which the speech signal is produced is a source of degradation of the word accuracy if it is lower or higher than normal. For this reason, we have carried out a research work which has reached three goals: (a) get a better understanding of the Speech Production Loudness (SPL) phenomenon, (b) find out the parameters of the speech recognizer that are the most affected by loudness variations, and (c) compute the effects of SPL and whispery speech in Large Vocabulary Continuous Speech Recognition (LVCSR). In this paper we report the results of this study for three different loudnesses (low, normal and high) and whispery speech. We also report the word accuracy degradation of a continuous speech recognition system when the speech production loudness is different than normal as well as the degradation for whispery speech. The study has been done using the TRESVOL Spanish database, that was designed to study, evaluate and compensate the effect of loudness and whispery speech in LVCSR systems.

1 . INTRODUCTION

The SPL varies both locally and globally among utterances: during the production of a sentence, there are significant changes in the amplitudes of the sounds, depending on the phrasal and lexical stress. Additionally, depending on the kind of communication channel, the background noise and the level of confidentiality of the information (PIN, credit card number, etc.) speakers tend to increase or decrease the SPL. Finally, there are speakers that usually speak louder than others by phone. These facts are responsible for three effects: (a) spectral shape and envelope changes, (b) pitch variations and (c) signal to noise ratio (SNR) variations. We have checked out that these effects dramatically affect the performance of LVCSR systems since our experiments show that the word error rate (WER) increases up to 2 times for slow SPL with respect to the WER at the normal SPL.

The study we are presenting in this paper is of interest to design methods for accounting for the sources of variability in LVSCR systems in two directions: detection and compensation of the SPL variation phenomenon.

We have found several studies related to SPL and other voice parameters [1][2][3][4][5]. The first one is based on the idea that loudness patterns are closer to t[1]he human perception of sound waves than spectrograms and presents a method to compute loudness patterns which is later used to characterize the subjective performance of a codec. The second focuses on vocal effort and type of phonation and perform a series of perception experiments to prove the importance of factors like spectral damping from the perceptual point of view. The third shows that the sources of sound pressure level variation are the subglottal pressure and glottal configuration depending on the vowel that has the syllable (full vowels and reduced or non-nuclear full vowels respectively). The fourth performs perception experiments to prove that loudness together with duration are cues in the perception of stress. Finally, the fifth studies the acoustic characteristics of lexical stress since it has been pointed out that stress might be useful for improving speech recognition performance [8]. All this studies are related to human perception and automatic stress detection and provide with very useful results for speech analysis and synthesis, but they are not directly related to the detection and compensation of the loudness variations in speech recognition. Our study has three goals: (a) get a better understanding of the SPL phenomenon and whispery speech, (b) find out what parameters of the speech recognizer models are the most affected by loudness variations and (c) compute the effects of SPL and whispery speech in LVCSR.

Since we strongly believe that it is necessary to have a specific database to study the SPL phenomenon and to get reliable evaluations of both SPL classifiers and compensation techniques, the first step of this research work was the design and collection of the TRESVOL database, which is described in section 2. Section 3 presents the characteristics of the SPL as far as speech rate, pitch, spectrum and SNR is concerned. Section 4 describes the speech recognition experiments and the results. Finally, in section 5 we present our conclusions and future work.

2. THE TRESVOL DATABASE

The TRESVOL Spanish database has been designed to study, evaluate and compensate the effect of SPL in LVCSR systems. It is composed of utterances produced at low, normal and high loudnesses and also of whispery speech utterances. The speakers were asked to utter each sentence at the four above mentioned ways, so that we could: (a) characterize the acoustic

[1] ENST (Bretagne), currently at Telefónica I+D with a ERASMUS grant.

properties of sounds like energy, pitch, duration,... for each speaker and case and (b) determine the effects of SPL and whispery speech in continuous speech recognition.

The database is composed of 1200 different sentences containing telephone and driving license numbers, amounts and spontaneous speech sentences. There are 30 speakers (15 male and 15 female) that were carefully selected as in [6][7], and the total number of utterances is 9600 (2400 utterances for each SPL and 2400 for whispery speech).

Figure 1 shows the probability density function (pdf) of the square root of the energy for the four cases. It can be observed that the range of variation for the energy is very large and the four density functions overlap, what shows the lack of consensus among speakers on what we subjectively call low, normal and high SPL. Concerning whispery speech, its pdf is very close to the one corresponding to low SPL; there are even cases were the energy of whispery speech is higher than the energy of speech produced at normal loudness.

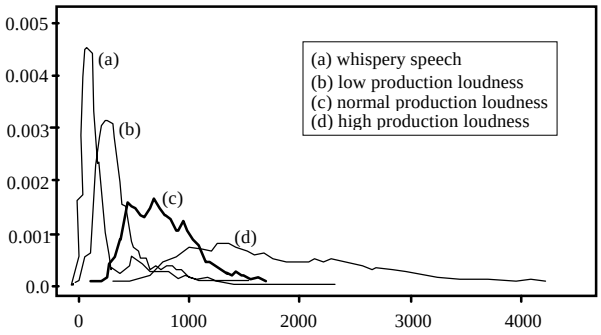

Figure 1: Probability density function of the square root of the energy.

3. CHARACTERISTICS OF THE SPL

In this section we report the characteristics of the SPL phenomenon as far as speech rate, pitch, spectrum and SNR is concerned. All the experiments and measures were done using the TRESVOL database.

3.1 SPEECH RATE

Automatic and accurate computation of the phone rate requires the correct transcription of the utterance and a measure of the rate of speech (ROS) that does not depend on the set of phones that compose the sentence [6].

The speech rate was computed automatically in two steps. First, the utterances were forced aligned to determine the phone segmentation. Then, the speech rate was obtained by using the measure of ROS described in [6]. Part of the resulting segmentation of the speech material was checked manually to verify the reliability of the measures.

Figure 2 shows the results: the probability density functions of the speech rate for each speech production loudness are very similar. The sample mean of the speech rate for low SPL is slightly larger than the one for normal SPL and the sample mean of the speech rate for high SPL is slightly lower than the one for normal SPL. Hence, there are no substantial differences between them. Additionally, by examining sentence by sentence at the three different SPLs we concluded that there is no relation between speech rate and SPL and therefore a variation of the SPL does not imply a variation of the speech rate.

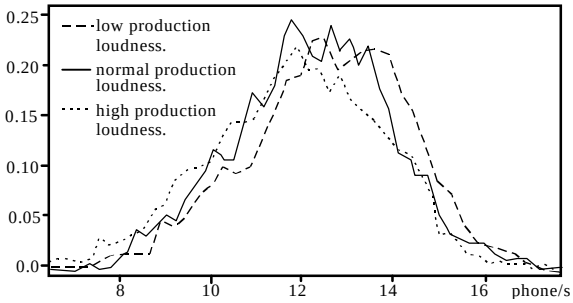

Figure 2: Probability density function of the speech rate for each production loudness.

3.2 PITCH

The first experiments tried to find out whether there is a relation between the average pitch and the SPL. We did it by comparing the average pitch for each sentence and speaker uttered at the three SPLs. The results show that the average pitch increases with the SPL in 83% of the utterances. There are some cases where the average pitch for normal or high SPL is slightly lower than for low or normal SPL respectively, though this is not the general behavior and it is usually related to utterances for which the increase of the SPL is small.

The second experiment was oriented to obtain the average increase of the average pitch for the speakers of the database. We checked out that the average increase of the average pitch is larger for female speakers than for male. We also checked out that the largest increase is produced when a speaker goes from normal to high SPL (31 Hz for female speakers and 29 Hz for male).

Finally, in figure 3 we plot the scattergram of the square root of the energy versus the average pitch, where a point represents the square root of the energy and the average pitch of a sentence of the database. This plot shows that there is a general relation between both parameters for both female (upper part of the plot) and male speakers (lower part of the plot).

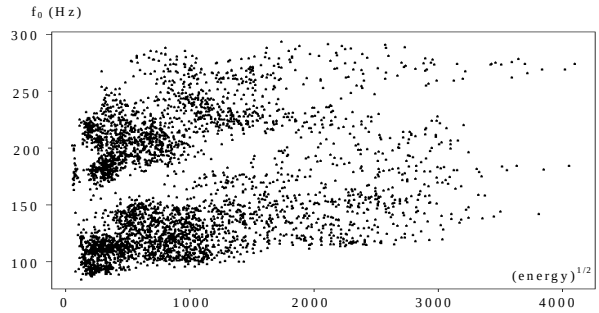

Figure 3: Speech production loudness versus pitch.

3.3 SPECTRUM

In [4] it is stated that energy differences between stressed and unstressed syllables are concentrated in the higher part of the spectrum, whereas intensity differences below 500 Hz are negligible. We have observed the spectrum of different phones in both stressed and unstressed syllables uttered at the three SPLs in order to find out whether this rule can be also applied to differences in the production loudness. We have checked out that, in general, there are two main differences between voiced sounds uttered at different SPLs: (a) the energy difference in the higher part of the spectrum, (b) the periodicity of the spectrum.

Figure 4 shows both differences for the stressed vowel "e": The upper spectrum corresponds to the high SPL, the lower one corresponds to the low SPL and the one in the middle corresponds to the normal SPL. It can be observed that periodicity of the spectrum is high for both high and normal SPL while is lower for low SPL (in the higher part of the spectrum). Additionally, the largest energy differences are concentrated in the higher part of the spectrum.

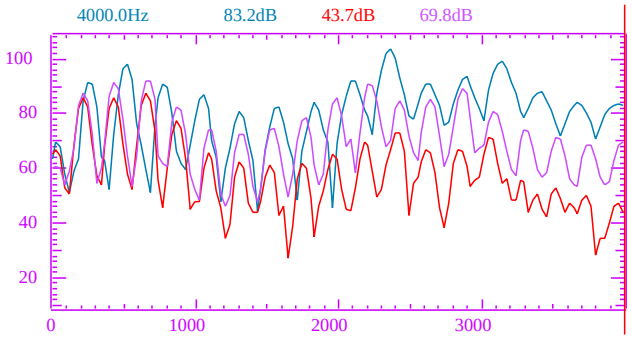

Figure 4: Spectrum of vowel "e" at three different SPLs.

Figure 5 shows the spectrum for the non-stressed vowel "o" at the end of an utterance, what implies that the energy of the three SPLs is low compared to the previous case. The energy differences are more balanced than in figure 4 in the sense that they affect all the frequencies of the spectrum a similar way. Additionally, only the high loudness utterance has a fully periodic spectrum.

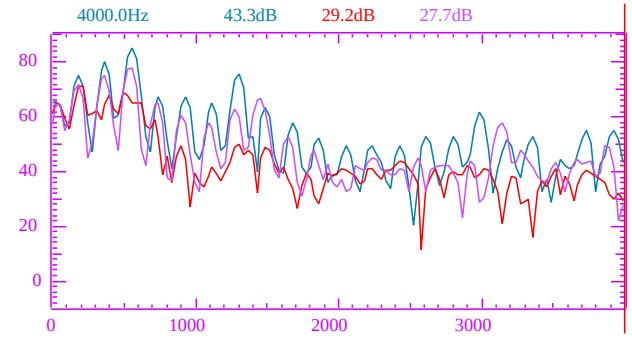

Figure 5: Spectrum of vowel "o" at three different SPLs.

Concerning other phones different than vowels, we have observed the energy variation of fricatives is low and seems to affect to all parts of the spectrum in a similar way. It seems that fricatives that belong to non-stressed syllables are the most affected by SPL changes, though a more careful study is being carried out to determine whether this is a rule. Nasals, for example, are affected by the production loudness like the vowels, even though the energy variation in lower and is more balanced along the frequency axis. Table 1 shows the percentage of average energy reduction with respect to the energy of high SPL case for the Spanish sounds. The energy was measured in dB.

	Normal SPL	Low SPL
vowels	10.4 %	18.9 %
fricatives	8.7 %	15.6 %
plosives	9.9 %	17.7 %
nasals	8.3 %	16.7 %
v. consonants	9.7 %	18.3 %
affricates	10.1 %	18.5 %

TABLE 1: percentage of energy reduction.

3.4 SIGNAL TO NOISE RATIO

It is well known that background noise is a source of degradation of the word accuracy [9]. This is not only due to the mismatch between the training and the testing conditions but also to the end-point detector errors, that increase when the SNR decreases.

The TRESVOL database was recorded in a special recording room which has always the same kind and level of background noise, so that the SNR of the resulting database is proportional to the SPL. We have checked out that this also occurs in real telephone applications.

Figure 6 shows the probability density function of the SNR for each subset of the database: the sample mean of the SNR is 38 dB for high SPL and 31 dB for normal SPL while it is 20 dB for low SPL and just 14 for whispery speech. Hence, even though the background noise level was very low, it dramatically affects the last two cases and, consequently, this will be one of the reasons why the recognition WER will increase in both cases. It can also be observed that the variance for low SPL is larger than the variance of the other density functions.

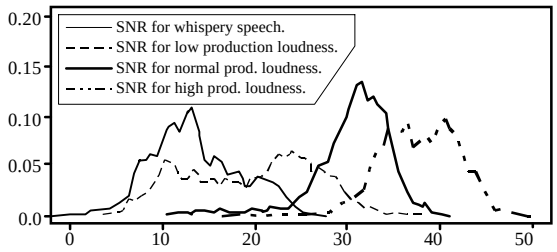

Figure 6: SNR probability density function.

We have evaluated the end-point detector and have checked out that in the case with the lowest SNR, that corresponds to whispery speech, it made errors that cannot be solved by the speech recognizer in 17% of the files.

4. EXPERIMENTS AND RESULTS

The goal of the experiments was to determine the performance of our baseline system for the three different speech production loudnesses as well as for the whispery speech.

The experiments have been carried out using the speech recognizer of the ATOS conversational system [10], which vocabulary size is around 4700 words.

Figure 7 shows the results. As expected, the lower WER is obtained, for normal SPL. This result is reasonable given that the training database is composed of files with similar SNR and SPL. In spite of that, the WER for normal SPL is still high for this vocabulary size. This is due to the fact that 50% of the testing database is composed of connected numbers and amounts, which is a very difficult task because of the large variation of the speech rate, the difficulty to include all the possible pronunciations in the dictionary and the language model, that does not help too much in this task.

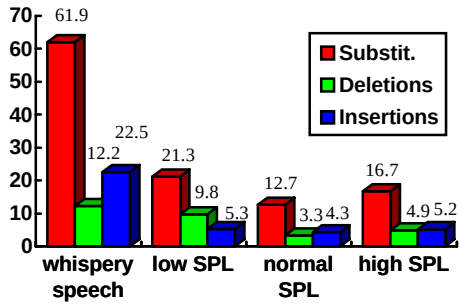

Figure 7: WER for different SPLs and whispery speech.

The WER for high SPL is higher than for normal SPL, but the percentages of insertions and deletions are still similar. In the cases of whispery speech and low SPL there is a lack of balance between insertions and deletions, and the WER is much higher than in the two previous cases.

The lack of balance between insertions and deletions suggests that by adjusting the language weight and the word insertion penalty, like in [6], we could easily compensate part of the effect of low SPL and whispery speech, even though the large amount of substitution errors indicates that the model parameters (mean, variances and gaussian weights) should be adapted.

The state transition probabilities could remain as they are since the rate of the speech does not change (section 3.1) and therefore they would not require adaptation.

5. CONCLUSIONS

In this paper we have reported the characteristics of the SPL phenomenon as far as speech rate, pitch, spectrum and SNR is concerned. We have shown that while SPL is not related to speech rate changes, it is related to pitch variations, spectral shape and envelope changes and SNR variations. We have also analyzed the recognition errors and checked out that just for low SPL and whispery speech there is a lack of balance between insertions and deletions and therefore part of the errors could be compensated by adjusting the language weight and the word insertion penalty. Finally, it can be observed that the means, variances and gaussian weights should be adapted to do recognition at SPLs different than normal.

We are currently focusing our research work on three different areas: (a) continuing the study of the SPL variation phenomenon to derive rules that allow us to predict spectral changes, (b) developing compensation techniques based on our current knowledge and, (c) developing a SPL and whispery speech detector/classifier, that will help us to apply the best compensation technique for each utterance.

6. BIBLIOGRAPHY

[1] M. Hauenstein, "A Computationally Efficient Algorithm for Calculating Loudness Patterns of Narrowband Speech", In Proc. ICASSP'97, Munich, Germany, April 1997.

[2] G. Klasmeyer, "The Perceptual Importance of Selected Voice Quality Parameters", In Proc. ICASSP'97, Munich, Germany, April 1997.

[3] H. M. Hanson, "Vowel Amplitude Variation During Sentence Production", In Proc. ICASSP'97, Munich, Germany, April 1997.

[4] A. M.C. Sluijter, V. J. van Heuven, "Intensity and Vocal Effort as Cues in the Perception of Stress", In Proc. Eurospeech'95, Madrid, Spain, September 1995.

[5] D. van Kuijk, L. Boves, "Acoustic Characteristics of Lexical Stress in Continuous Speech", In Proc. ICASSP'97, Munich, Germany, April 1997.

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

[7] F. Martínez, D. Tapias, J. Álvarez, "Towards Speech Rate Independence in Large Vocabulary Continuous Speech Recognition", In Proc. ICASSP'98, Seattle, Whashington, USA, April 1998.

[8] A. Waibel, "Recognition of Lexical Stress in a Continuous Speech System - A Pattern Recognition Approach", In Proc. ICASSP'86, Tokyo, Japan, April 1986.

[9] P. J. Moreno, "Speech Recognition in Noisy Environments", PhD. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, 1996.

[10] J. Álvarez, D. Tapias, C. Crespo, I. Cortazar, F. Martínez, "Development and Evaluation of the ATOS Spontaneous Speech Conversational System", ICASSP'97, Munich, Germany, April 1997.