Speaker's Prosodic Strategy for a Large Physical Distance Communication Task

Thibaut FUX [1,2], Véronique AUBERGE [2,3], Gang FENG [2], Véronique ZIMPFER [1]
1 ISL, French-German research institute of Saint-Louis, BP70034, 68301, France
2 GIPSA-Lab, UMR 5216 CNRS/Grenoble INP/ UJF/U. Stendhal, France
[3] LIG -Computer Science Lab, CNRS, LIG, France
{thibaut.fux;veronique.zimpfer}@isl.eu,
{veronique.auberge ; gang.feng}@gipsa-lab.genoble-inp.fr

In a face to face communication situation, the speaker modifies his voice when he notices that the listener goes physically farther. This voice adaptation aims to preserve a good communication condition in spite of environmental constraints. One fundamental question is to know, for a given utterance, which strategies the speaker can process in order to maintain a good level of communication: which part is global "signal processing" and which part is "structural processing" (reinforcing or reorganisation of segmental/prosodic/linguistic materials). The first physical constraint is that, since the sound pressure level decreases when the distance increases, the speaker must compensate this attenuation by increasing his voice production level and by involving specific strategies in order to carry the relevant cues. Thus, the speaker must provide more intense vocal effort so that his voice intensity could increase. Consequently, not only the sound pressure level at lips increases, but also the nature of his voice changes. These modifications of the speech signal, which result from the speaker's vocal effort, are perfectly perceptible for the listener and allow him to estimate his physical distance to the speaker [1].

Even if a lot of studies have shown that the vocal effort affects some properties of the speech signal, the few attempts to modal-to-shouted voice transformation do not seem to be clearly fruitful [2][3][4]. In these studies, the voice transformation was mainly performed by changing mean values of the acoustical parameters and it seems that the prosodic characteristics of the global patterns and salient cues of the shouted utterances were not investigated. However, several authors have shown that the "classical" prosodic parameters (i.e. variation of f0, variation of intensity and timing) are mainly responsible for the perception of vocal effort for high production levels, rather than other acoustical parameters (spectral tilt, formant's frequency and bandwidth) [5][6]; and it is particularly true for f0 [7]. Thus, the present study aims at knowing more precisely which kind of prosodic strategies are involved by speakers in order to ensure the information preservation in such "face to face" vs. "far" interaction.

First, four corpus of non-sense words were recorded. The non-sense words were constructed using all the 17 French consonants and the 3 cardinal vowels [a], [u], [i]. The first two corpus use mono-syllabic words; CV words and CVC words. The last two corpus use bisyllabic words; VCV words and CVCV words. Note that in each word the consonants and the vowels are the same in order to study the influence of the phoneme position on the word. The recording was performed using a protocol simulating the talker-to-listener distance by placing the speaker in a soundproofed room [9]. The listener remained outside but was able to see the speaker through a window. The speaker must adjust his vocal effort until being understood by the listener. This simulated protocol corresponds to a communication distance

of approximately 60 m; evaluated from the inside-to-outside room attenuation (-36dB).

The analysis of this database has shown several differences between spoken and shouted voices:

- The intensity analyses have shown greater variations for vowels than for the consonants for the high voice production level, for the entire four corpus. These results are in agreement, but in different proportion to the variations reported by [8] and may explain the greater intensity dynamics observed in [9]
- About the phoneme's duration we have observed a trend to shortening the consonants' duration and to lengthening the vowels' duration for high voice production levels; especially for the last vowel of the word; which is consistent with [10] and [11].
- The analysis of the f0 has given more interesting and new results (cf. Figure 1).
    - The f0 of the initial and/or the final voiced consonants (in **C**V, **C**V**C**, **C**VCV) increases less than the other phonemes of the word and that their initial f0 value and/or their final f0 values is lower than for vowels. However, the f0 of intervocalic consonants (in V**C**V, CV**C**V) increases in the same proportion than the f0 of vowels.
    - There exists a similar f0 pattern for all the vowels in all the words. We observed an asymmetric rise-fall pattern, with large dynamics (about 4-5 tones) and where the maximum value is around the 2/3 of the vowels length.
    - The last observation concerns the f0 trend line. We observed that the f0 increases along the majority of the shouted words and that the second vowel has higher f0 value than the first one.

These observations suggest a focalization pattern over each vowel, even for the words containing two vowels, but such pattern was not observed on the consonant. Furthermore prevocalic voiced consonants placed at the beginning of the word show a clear increase of the fundamental frequency. This increase seems to be the beginning of the following vowel focus. Thus, rather than a simple increase of the variability of the f0 and the intensity (i.e. an amplification), we observed between spoken and shouted words, a complete prosodic reorganization in order to be hyper-intelligible all over the non-sense words. Furthermore this prosodic organization has been observed during shouted sentences.

The question concerning a possible link from the physical space to the social space is now: could "physically closed" bootstrap analogy transfer with "socially closed" (intimacy), and "physically far" bootstrap analogy transfer with "socially far" (authority attitude), that seems to be shared by many cultures?

[1] **Gardner, M. B., (1969),** "Distance estimation of 0-degree or apparent 0-degree-oriented speech signals in anechoic space", *J. Acoust. Soc. Am.*, Vol. 45(1), pp. 47-53.

[2] **H.A. Cheyne, K. Kalgaonkar, M. Clements, and P. Zurek, (2009),** "Talker-to-listener distance effects on speech production and perception," *J. Acoust. Soc. Am.*, 126(4), pp. 2052-2060.

[3] **Nieto, O. (2008).** "Voice Transformations for Extreme Vocal Effects", *Master's thesis, Pompeu Fabra University*

[4] **Richard G. and d'Alessandro C., (1996),** "Analysis/synthesis and modification of the speech aperiodic component", *Speech Communication 19*, pp.221-244

[5] **Tassa, A. and Liénard, J.-S., (2000),** "A new approach to the evaluation of vocal effort by the PSOLA method", *The European Student Journal of Language and Speech*, http://www.essex.ac.uk/web-sls/papers/00-01/00-01.html (last viewed June, 2011).

[6] **Fux, T., Feng, G. and Zimpfer V., (2010),** "Le rôle de la prosodie dans la perception de l'effort vocal (The role of the prosody in the perception of the vocal effort)", *Actes du 10èmes congrès français d'acoustique*.

[7] **Brungart, D. S., Kordik, A., Das, J. K. and Shaw A. K., (2002),** "The effects of f0 manipulation on the perceived distance of speech", *Proc. of Int. Conf. on spoken language processing*, pp. 1641-1644.

[8] **Fairbanks, G., Miron, M. S., (1957),** "Effects of vocal effort upon consonant-vowel ration within the syllable", *J. Acoust. Soc. Am.*, Vol. 29(5), pp. 621-626.

[9] **Fux, T., Feng, G. and Zimpfer V., (2011), "**Talker-to-listener distance effects on the variations of the intensity and the fundamental frequency of the speech", *Proc. of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, may 22-25,

Prague, Czech Republic.
[10] Rostolland D., (1982), "Acoustic Features of Shouted Voice", *Acustica*, vol. 50, pp. 118-125.
[11] Traunmüller H. and Eriksson A., (2000), "Acoustic effects of variation in vocal effort by men, women, and children", *J. Acoust. Soc. Am*, vol. 107, pp. 3438-3451.

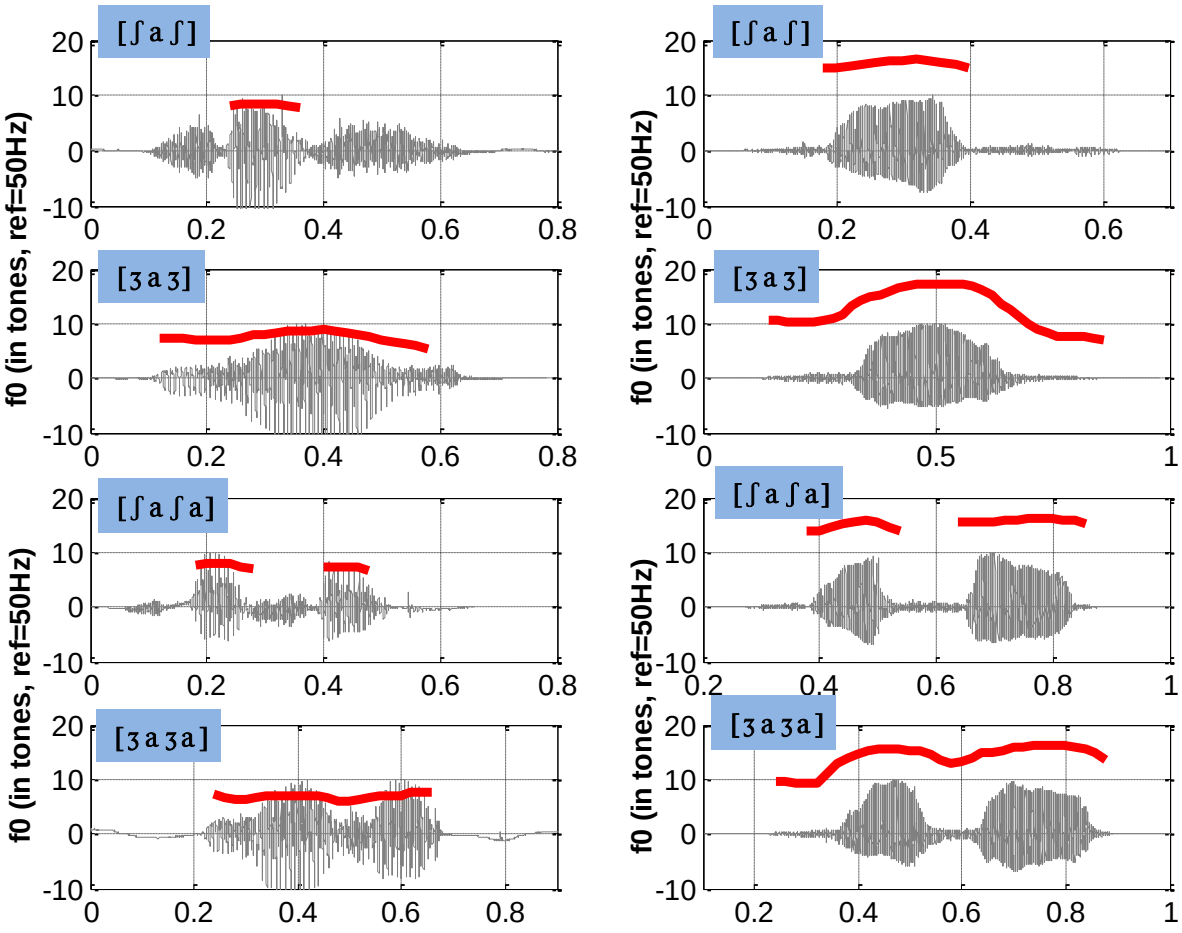

Figure 1: *f0 contour in tones for 4 words* ([ʃ a ʃ], [ʒ a ʒ], [ʃ a ʃ a], [ʒ a ʒ a]) *for spoken (left) an shouted voices (right).*