A STATISTICAL STUDY OF PITCH TARGET POINTS IN FIVE LANGUAGES

Estelle Campione, Jean Véronis

Laboratoire Parole et Langage
Université de Provence & CNRS
29, Av. Robert Schuman, 13621 Aix-en-Provence Cedex 1, France
Estelle.Campione@lpl.univ-aix.fr

ABSTRACT

We present the results of a large-scale statistical study of pitch target points in five languages, on a corpus comprising 4 hours 20 minutes of speech and involving 50 different speakers. The entire corpus has been stylized automatically by a technique reducing the F_0 contour to a series of target points representing the significant pitch changes. It was then entirely verified by experts using a resynthesis method, in order to ensure that there was no audible difference with the original. The set of ca. 50000 pitch target points thus obtained was then analyzed from a statistical point of view. In this paper we describe the main results of this study, in terms of frequency distribution of target points, pitch movements and relation of pitch movements to time interval. Our study reveals interesting differences across languages and sex.

1. INTRODUCTION

Large sets of prosodic data on many languages would be a useful resource for theoretical studies, as well as for practical applications such as speech synthesis or speech recognition. However, due to the cost of extracting prosodic information form corpora, and to the lack of robust tools and methods, the data is sparse at the moment and even non existent for many languages. In addition, the data that does exist has been extracted with different approaches depending on the languages, due to methodological differences among schools, and is difficult to compare.

We will present in this paper a statistical study of intonation on a large corpus in five languages. The corpus has been stylized in a homogeneous way by the same, language-independent technique, which enables direct comparison of results. This paper demonstrates that the technique used readily enables a large array of measurements and statistical studies.

We describe the main results of this study, in terms of frequency distribution of target points, pitch movements and relation of pitch movements to time intervals. These results reveal interesting phenomena, and in particular somewhat surprising differences across language and sex. This paper is only intended to present raw results, and will not attempt at theoretical explanations of the observed phenomena.

2. STYLISATION METHOD

The stylization method used in this study was proposed by Hirst and Espesser [7] (see also [6]). Their algorithm (MOMEL, for *MOdélisation MELodique*) reduces the F_0 of the signal to a series of pitch target points that represent the relevant macromelodic movements of the utterance (Figure 1). Once interpolated by a quadratic spline curve, these points generate an F_0 contour which (apart from a few errors that must be corrected by hand) is not distinguishable from the original when fed into a resynthesis technique (PSOLA, [5]). The series of pitch target points thus seems to capture all the relevant macromelodic information of the utterance.

Other stylization methods have been proposed (e.g. [4] [10] [8] [9] [3]). However, the target point representation is particularly simple and economical, and it is well suited to statistical analyses such as the one presented here.

3. CORPUS

The corpus is composed of passages of ca. 20 seconds read by 10 different speakers (5 female, 5 male) in five languages (English, French, German, Italian, Spanish), i.e. 50 speakers altogether [1]. For each language, there are 40 different passages of 5 sentences, but each speaker reads only a subset of them. The total duration is 4 hours 20 minutes. Duration per language ranges from 36.5 minutes (French) to 73 minutes (German).

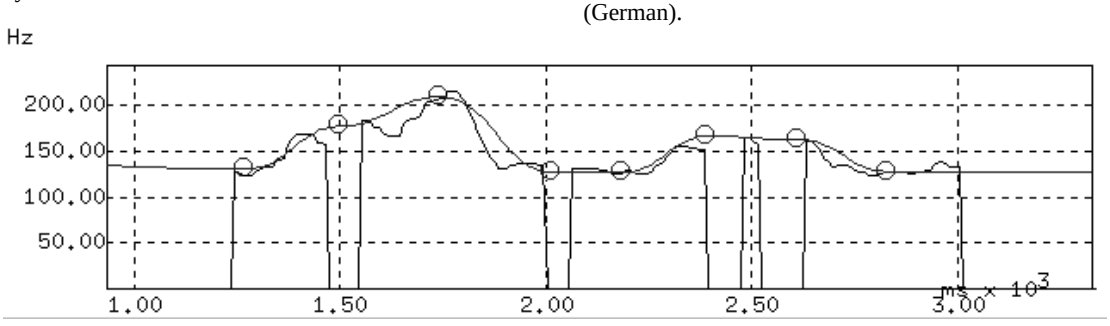

Figure 1. Stylized sentence from French speaker *bf* (*J'ai des problèmes avec mon adoucisseur d'eau*).

The recordings were borrowed from the EUROM 1 database developed in the SAM project [2], and the stylization was done automatically using the MOMEL algorithm described above. The entire stylized corpus was then verified manually and the pitch target points were corrected when necessary (about 5% of cases) so that there was no audible difference between the original and the stylized F_0. Altogether, the corpus contains 50360 pitch target points.

4. FREQUENCY DISTRIBUTION

The mean frequency of target points varies between 89.1 and 97.2 ST for female speakers and between 78.6 and 89.5 ST for male speakers. There are clear differences between languages as shown in Figure 2. The most pronounced difference is between French and German (4.8 ST for female speakers, 4.7 for male).

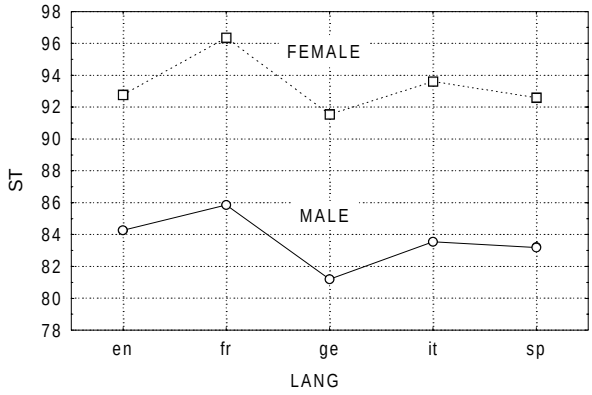

Figure 2: Mean frequency per language and sex

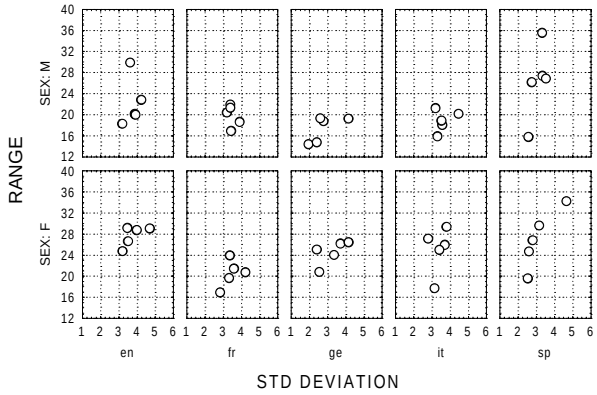

Figure 3: Standard deviation vs range per language and sex

There are also important differences in dispersion between speakers. The standard deviation varies from 1.9 to 4.7 ST, and the speaker's range (highest value minus lowest value) varies from 14.4 to 35.5 ST (Figure 3). Small standard deviations can be associated with large ranges (due to the presence of a few extreme values -- see below). Some languages (e.g. Spanish) seem to show more variability than other (e.g. French).

The distribution of target points for each speaker is approximately normal (Figure 4). However a strict normality assumption (as measured for example by Shapiro-Wilks' *W* test) must be rejected for most speakers.

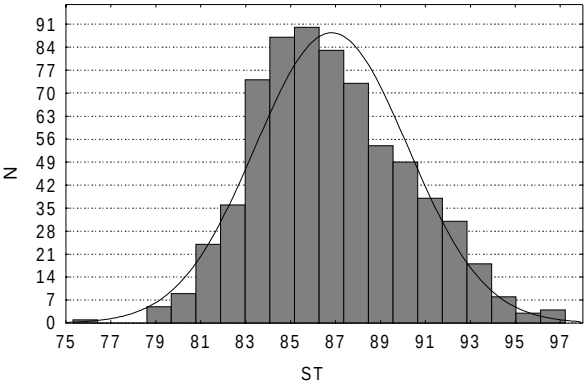

Figure 4: Distribution of target points for French male speaker *bf*

The individual distributions show various degrees of skewness and kurtosis. It is interesting to note that female speakers (apart from the French) have a much greater variability in skewness and kurtosis than male speakers (Figure 5).

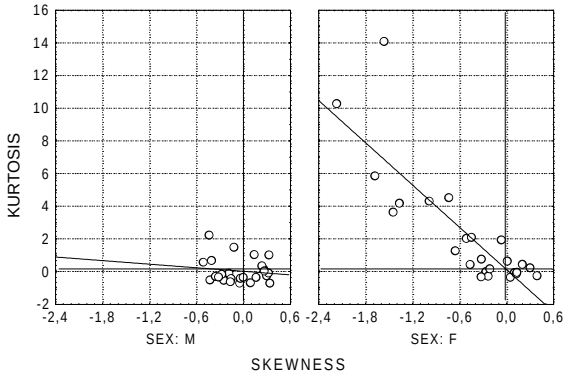

Figure 5: Skewness vs. kurtosis

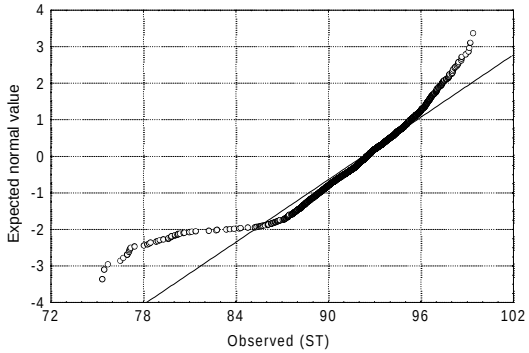

Figure 6: Observed vs. normal values for German female speaker *mi*

Skewness and kurtosis are strongly correlated for female speakers, and closer examination of the data reveals that both are mostly due to the presence of a infra-grave values. Figure 6 shows an example of a German female speaker with a particularly strong excess of extreme values.

5. PITCH MOVEMENTS

Pitch movements (i.e. the difference in ST between two consecutive target points) range between -25.0 and 23.0. However, most movements are of a much more limited amplitude and some languages show more variability than others, as shown in Figure 7.

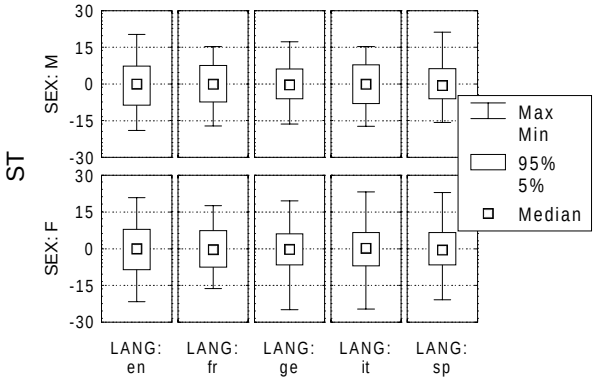

Figure 7: Pitch movements per language and sex.

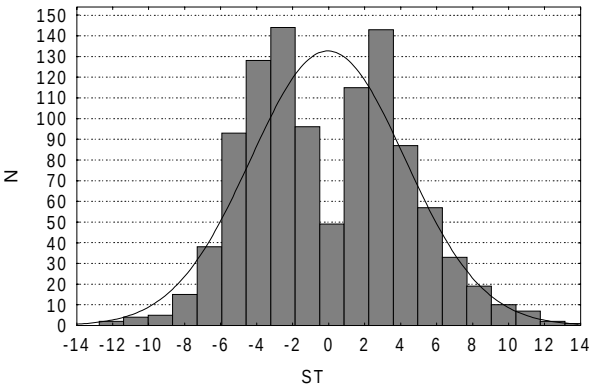

Figure 8: Distribution of pitch movements for Italian male speaker *au.*

The distribution of pitch movements if typically bimodal, with a deficit of values around zero (Figure 8). It clearly differs from the (normal-shaped) distribution that would be expected if consecutive target points were drawn randomly from the speaker's distribution and seems to result from the combination of two distributions, one for ascending movements, one for descending movements. Figure 9 shows the distributions of ascending and descending (*z*-transformed) pitch movements for all French speakers. The distributions can reasonably be well fitted by a Weibull distribution function.

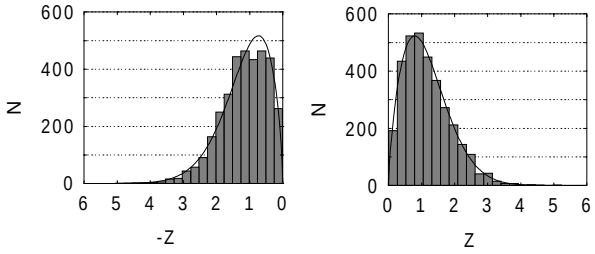

Figure 9: Distribution of *z*-transformed pitch movements for all French speakers fitted by a Weibull law

Each target point is very slightly correlated to the previous one. The correlation coefficient ranges from 0.064 (French) to 0.17 (Spanish). These values are low, but significant at $p < 10^{-4}$, given the large size of the samples. When ascending and descending target points are considered in two separate groups, correlations are more important. They range from 0.57 (German) to 0.62 (English) for ascending points and 0.55 (German) to 0.68 (Spanish). These values are significant at $p < 10^{-4}$. However, they may partly result from the shape of the distribution. Points drawn randomly from a normal distribution for example result in a 0.47 correlation coefficient for categorized movements. We therefore compared the observed coefficients with the values that would be obtained by randomizing the order of points within each passage of the corpus (Table 1 shows the values for French). Overall, when all pitch movements are considered together independently of their direction, the observed correlation is not significantly different from that on randomized points. However, when target points are considered by groups (ascending, descending) the difference with randomized points is highly significant ($p < 10^{-4}$).

	r_{chrono}	r_{random}	**N**	***p***
all	0.0637	0.0458	6940	0.29
H	0.619	0.496	3490	$<10^{-4}$
L	0.599	0.505	3450	$<10^{-4}$

Table 1. Correlation between consecutive target points in normal chronological order vs. randomized order (French).

These findings show that the intonation contours differ significantly from Gaussian noise, and that pitch movements do not constitute a homogeneous population, but are differentiated in two groups, ascending and descending. This is consistent with many theories which consider H-L movements as the basic units of intonation rather than fixes zones of pitch.

6. RELATION OF TIME AND PITCH

Pitch target points (*z*-transformed) are slightly (negatively) correlated with their temporal location (in ms) in the passages: the correlation coefficient ranges from -0.11 (German) to -0.20 (English). The slope of the regression line is comprised between

-0.016 std/s (German) and -0.035 std/s (English), corresponding to an average of -0.048 ST/s to -0.13 ST/s. This indicates a small downdrift effect on the entire duration of the passages (from ca. 1 ST for German to ca. 2.5 ST for English).

The distribution of time intervals between target points is approximately lognormal for all languages (Figure 10), with a geometric mean remarkably constant, ranging between 252 ms (French) and 275 ms (Italian).

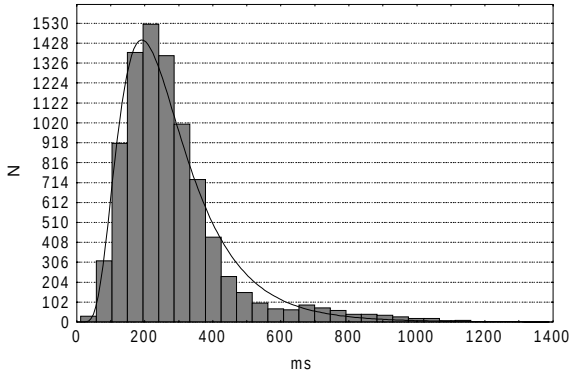

Figure 10: Distribution of time intervals between target points (English)

The correlation between (log-transformed) time intervals and pitch movements is very low, ranging from 0.029 (Spanish) to 0.075 (English) (these values are however significant at least at $p < 0.002$). Correlation by groups (ascending, descending) is much higher since it reaches almost 0.5 for Italian ascending points, and reveals important differences between languages (Figure 11).

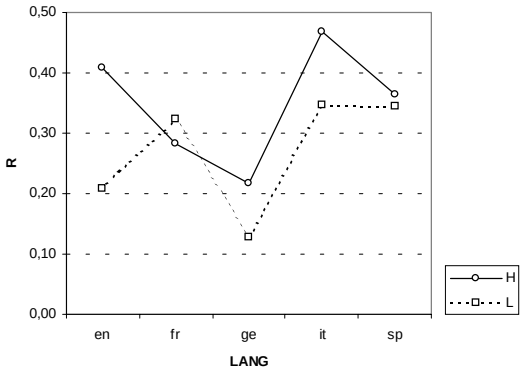

Figure 11: Correlation of pitch movements and time intervals per language.

5. CONCLUSION

This paper demonstrates that stylization of pitch contours by means of target points is a simple and powerful, language-independent technique that readily enables gathering empirical prosodic data on large corpora. The results reported in this study (gathered on a five-language corpus comprising 4 hours 20 minutes of speech and involving 50 different speakers) reveal interesting properties of the frequency distribution of target points, pitch movements and the relation of pitch movements to time intervals. Our study also shows interesting differences across languages and sex. The reasons of these differences remain to be explained: they could be connected to both the specificity of the languages and sociolinguistic differences (age of speakers, etc.) among speaker groups, and require further investigation.

6. REFERENCES

1. Campione, E. and Véronis, J. "A multilingual prosodic database." *ICSLP'98*. Sidney, Australia (these proceedings), 1998.

2. Chan, D., Fourcin, A., Gibbon, D., Granstrom, B., Hucvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C. and Zeiliger, J. "EUROM1 - A Spoken Language Resource for the EU." In *Proceedings of Eurospeech'95*, Madrid, 1, 867-870, 1995.

3. D'Alessandro, C. and Mertens, P. "Automatic pitch contour stylization using a model of tonal perception." *Computer Speech and Language*, 9, 257-288, 1995.

4. Fujisaki, H. and Hirose, K. "Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation." *13th International Congress of Linguists*, 57-70, 1982.

5. Hamon, C., Moulines, E. and Charpentier, F. "A diphone system based on time-domain prosodic modifications of speech." *ICASSP'89*, 238-241, 1989.

6. Hirst, D.J., Di Cristo, A. and Espesser, R. "Levels of representation and levels of analysis for the description of intonation systems." In Horne, M. (Ed.), *Prosody: Theory and Experiment*, Kluwer Academic Publishers, Dordrecht, forthcoming.

7. Hirst, D.J. and Espesser, R. "Automatic modelling of fundamental frequency using a quadratic spline function." *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, 15, 75-85, 1993.

8. Taylor, P. "Automatic recognition of intonation from F_o contours using the Rise/Fall/Connection Model." In *Eurospeech'93*, Berlin, 2, 789-792, 1993.

9. Taylor, P. "The Rise/Fall/Connection model of intonation." *Speech Communication*, 15:1&2, 169-186, 1994.

10. t'Hart, J., Collier, R. and Cohen, A. *A perceptual study of intonation : an experimental-phonetic approach to speech melody*, Cambridge University Press, Cambridge, 1990.