TOWARDS A REVERSIBLE SYMBOLIC CODING OF INTONATION

Jean Véronis, Estelle Campione

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

ABSTRACT

This paper presents a two-step model for the symbolic coding and generation of intonation. First, the F_0 curve is reduced to a series of pitch target points that capture the macroprosodic information of the utterance. Target points are then converted into a sequence of labels. Generation is achieved through the reverse steps. The model is language independent and requires no prior training on the data. We discuss the influence of the number of categories on the precision of fit, and show, by an evaluation on a large multilingual corpus (4 hours 20 minutes of speech, 50 speakers, 5 languages) that a model composed of three ascending and three descending categories, plus a category for small or null movements enables a regeneration of ca. 99% of points at less than 2 ST than the original. Given that the model is capable of various improvements, it seems a good candidate for practical applications.

1. INTRODUCTION

Several systems of symbolic coding of intonation have been proposed, but so far, automatic labeling and generation from labels are still an open issue. An important feature of coding systems that may not have received much emphasis is reversibility, that is, the possibility of re-generating an F_0 curve perceptually identical to the original from the extracted prosodic labels.

Prosodic coding systems can be categorized in two types: linguistic systems, such as ToBI, which encode events of a linguistic nature, and phonetic systems, such as HLCB [8] or INTSINT [5], which aim only at providing a purely configurational description of the macroprosodic curve without interpretation. It is obvious that the first category poses greater problems for the automatisation, although work is underway in that area ([1] [7]). The second category, easier to implement, is of course less interesting in linguistic terms, but can still contribute to the development of labeled corpora useful for many applications. In addition, it can be seen as a first step towards automatic labeling of systems like ToBI, and can be a help in the development of such systems for languages other than American English. However, even for phonetic systems, labeling, generation and reversibility are far from achieved. Work is underway (e.g. [8]), but large-scale evaluations are not yet available.

In this paper, we present a systematic study of configurational models for encoding intonation contours. In particular we will address the question of the minimal set of labels or categories necessary to achieve reversibility with a good precision. An arbitrary level of precision can be trivially reached by increasing the number of categories, but usable systems must obviously aim at keeping this number minimal.

We use a two-stage model for coding:

1. First, the F_0 curve is stylized by means of target points which capture the macroprosodic information of the utterance [5].
2. Target points are then converted to a sequence of labels constituting a categorization of the pitch movements.

The pitch target points can be extracted automatically from the signal ([6] see also [5]). Once interpolated by a spline curve they produce an F_0 contour undistinguishable from the original (apart from a few detection errors that must be corrected by hand). Other stylization methods have been proposed, but the target point stylization seems particularly simple and economical.

Generation is achieved through the reverse steps: the sequence of labels is converted to target points, which are then interpolated by a spline curve to produce a smooth F_0 curve. Reversibility means that the regenerated F_0 curve should be close to, and if possible not distinguishable from the original.

2. MODEL

In a preliminary study [2], we showed that the distribution of target points (in semi-tones or STs) is approximately normal for a given speaker, although a strict normality assumption (as measured for example by Shapiro-Wilks' W test) must most often be rejected. There is a considerable variation among speakers in terms of skewness and kurtosis, and there is (mostly for female speakers) very often an excess of extreme values, especially in the infra-grave.

However, despite this variability, we will use the normal distribution as a mathematical model because of its simplicity. We will make further simplifications:

1. We will assume that target points are drawn from a normal distribution independently from each other, although there is in fact more correlation between consecutive target points than what would be expected just as a result of the shape of the distribution [2].

2. We will also neglect the relationship between time intervals and pitch movements, although there is some correlation between the two.

3. Finally, we will neglect resetting after short pauses, and downdrift effects.

Due to these simplifications, the results presented here constitute a baseline subject to improvement.

The mathematical model that we propose is fairly simple, and can be extended to any number of desired categories. Of course, increasing the number of categories trivially results in a better fit of the generated F_0 curve. There is therefore a trade-off between the number of symbols and the quality aimed at. In section 3 we will discuss criteria for deciding of such a trade-off.

For any number n of categories C_1, …, C_n, the problem can be stated as follows. Let x_i be the observed frequency of the current point, x_{i+1} the observed frequency of the next point, and C_i the category of the pitch movement between the two points. A generation model consists in a series of functions f_1, …, f_n that predict a frequency for the next point given x_i for each category C_i. For each of these functions there is an error between the predicted value and the observed value:

$$x_{i+1} = f_1(x_i) + \varepsilon_1$$

$$\dots$$

$$x_{i+1} = f_n(x_i) + \varepsilon_n$$

Given a sequence of N target points, the coding that results in the best fit is the sequence of categories that minimizes the mean squared error:

$$\frac{\sum_{i=1}^{N} \varepsilon_i^2}{N} = \frac{\sum_{i=1}^{N} (x_{i+1} - f_i(x_i))^2}{N}$$

It can be shown that, under the assumption that the target points are drawn from a normal distribution independently from each other, there is an optimal generation model for any number of categories. We will start with the simplest case, which uses only two categories, let us say **H** and **L**, coding ascending and descending movements respectively. In the discussion below, we will use z-transformed frequencies, for the sake of simplicity.

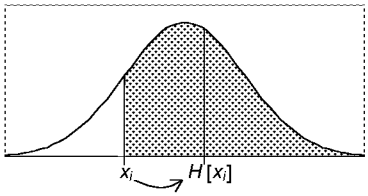

Figure 1. Distribution of ascending points

Let us assume a point x_i. If it is known that the next point is ascending, it must be drawn from the part of the normal distribution on the right of x_i (shaded area on Figure 1).

The density function for this ascending point is:

$$F(x) = \begin{cases} \dfrac{1}{1-\Phi(x_i)} N(x) & \text{if } x > x_i \\ 0 & \text{otherwise} \end{cases}$$

where $N(x)$ is the standard normal law, and $\Phi(x)$ its distribution function.

The error is minimal when for each point x, the next (ascending) point is computed as the expectation value of the density function F, i.e.

$$H[x] = \frac{1}{\sqrt{2\pi}\,(1-\Phi(x))} e^{-\frac{x^2}{2}}$$

If the next point is descending, a similar expectation value can be computed:

$$L[x] = -\frac{1}{\sqrt{2\pi}\,(1-\Phi(-x))} e^{-\frac{x^2}{2}}$$

Figure 2 shows a plot of both functions $H[x]$ and $L[x]$.

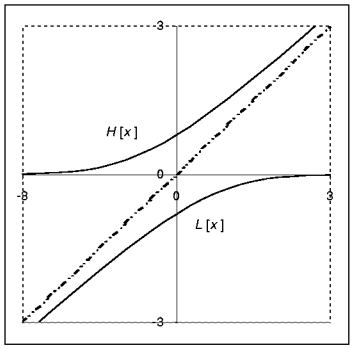

Figure 2. Plot of expectation values $H[x]$ and $L[x]$

The reasoning above can be extended to any number of ascending and descending categories. We will explore a series of even and odd models $\boldsymbol{M}_k$. Even models are models in which $k/2$ categories $\mathbf{H}_1$, … $\mathbf{H}_k$ code the possible target values above any given frequency x, and the same number of categories $\mathbf{L}_1$, … $\mathbf{L}_k$ code the values below x. Odd models are models in which in addition to the categories above, a central band, that we will call **S**, can be used in order to code small or null pitch movements.

Extending the model to an arbitrary number of categories amounts to finding an optimal series of bands that partition the values above and below any given x_i, such that when the next point x_{i+1} belongs to a band C_j, generating it as the expectation value $C_j[x]$ for that band minimizes the error. For lack of space, we cannot detail here the mathematics involved, and simply show in Figure 3 the bands (dotted lines) and expectation values (solid lines) for model $\boldsymbol{M}_7$, comprising 7 categories: $\mathbf{L}_3$, $\mathbf{L}_2$, $\mathbf{L}_1$, **S**, $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{H}_3$.

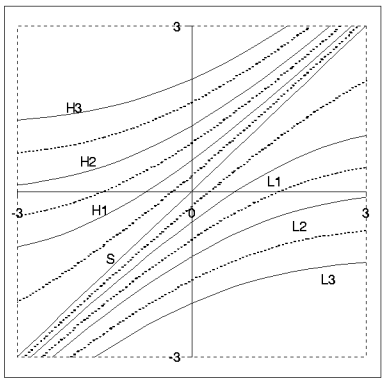

Figure 3. Optimal bands and expectation values (model M_7).

To summarize, for any model M_k, if x_i is the observed frequency of the current point, generating the next point as $C_j[x]$ if it belongs to band C_j minimizes the overall mean squared error. However, when generating a complete sequence iteratively, errors are cumulative, i.e. a coding decision which may be optimal for a subsequence 1, …, *i* can be not optimal for the entire sequence 1, …, *N*, and all the possible sequences should be tried in order to be sure to find the optimal one. In practice, we can simplify the problem and reduce the computations by making coding decisions that minimize the error on the next point without significant loss.

3. EVALUATION METHOD

The successive models were evaluated on a large multilingual prosodic corpus [3]. 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. 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).

The recordings were borrowed from the EUROM 1 database developed in the SAM project [4], and the stylization was done automatically using the MOMEL algorithm ([6] [5]). 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.

Eight models M_2 to M_9 were tested on the entire corpus. The mean and standard variation of each passage were used as parameters for the theoretical normal distribution used in the models.

Two measures were computed. Firstly, we used the mean squared error between the original target points and the regenerated target points in ST, which is a classical measure of quality of fit. However, this measure averages all errors, whereas, from a perceptive point of view, many small errors are less important than a small proportion of large ones, which may change radically the linguistic intention of the utterance. We randomly probed passages from the corpus and determined that when points are generated at less than 2 ST from the original, there was no change of linguistic intention. The largest differences can be audible under careful listening conditions, but are not perceived in the normal speech flow and in any case have no linguistic impact. We therefore use as a second measure the proportion of target points regenerated at less than 2 ST from the original.

4. RESULTS

As expected, the results improve with the number of categories, as shown in Table 1.

	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
MSE	4.54	2.96	1.76	1.28	.92	.74	.58	.48
<2ST	.694	.785	.905	.947	.977	.987	.992	.994

Table 1. Results per model (all languages)

For a given model, some languages are slightly better regenerated than others. The ranking of languages according to performance is the same for all models. Figures 4 and 5 show for example the mean squared error and proportion of points under 2 ST from the original per language for M_7.

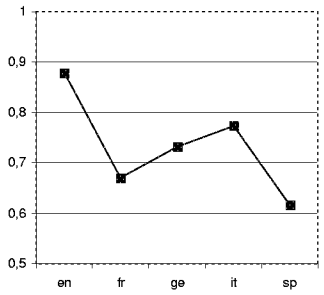

Figure 4. Mean squared error (ST) per language (model M_7).

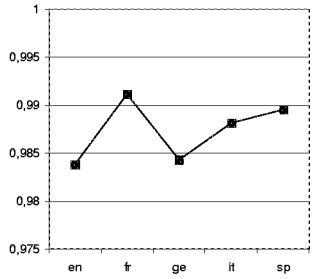

Figure 5. Proportion of points generated at less than 2 ST from the original, per language (model M_7).

The two measures are not completely equivalent; for example, Spanish has the smallest mean squared error, but ranks only second in terms of points under 2 ST from the original.

5. DISCUSSION

It is clear that a model based on only two categories, which places only about 2/3 of points at less than 2 ST from the original, is not usable for any purpose. At the other end of the scale, the model M_9 enables an almost perfect restitution of the original curve, at the expense of a much larger number of categories. It is therefore an empirical question to decide which model to chose for practical applications.

However, we should keep in mind that due to some simplifying hypotheses, the results presented here are a lower bound, and that the models could improve if additional phenomena were taken into account.

The most important factor seems to be the shape of the distribution. We assumed a normal distribution for all speakers, whereas [2] many speakers show various degrees of skewness and kurtosis. The analysis of residuals shows that the mean squared error is strongly correlated to the values of skewness and kurtosis, and that the most badly modeled points are points that are extreme, especially in the infra-grave for female speakers (Figure 6).

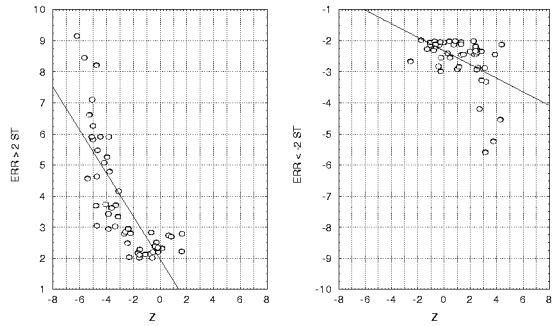

Figure 6. Errors > 2 ST (left) and < -2 ST (right) vs. *z*-transformed frequency (all Spanish speakers, model M_7)

Correcting the model to take into account the shape of the distribution for each speaker is likely to improve the results. For example, in Spanish modeled with M_7, more than half of the errors over 2 ST in absolute value are due to a single speaker, who has the most extreme skewness and kurtosis values. If that speaker were removed, the proportion of points modeled at less than 2 ST from the original would increase from 98.8 % to 99.5%.

Given this possibility of improvement, it seems that the model M_7, which currently reaches ca. 99% of points at less than 2 ST from the original, constitutes a satisfactory compromise between the precision of fit and number of categories. In addition, this model is conceptually simple in terms of categories, since the three categories in each direction can be see as a "normal", medium movement category, in conjunction with categories for "smaller" and "larger" movements (we could use mnemonic labels such as **L+**, **L**, **L–**, **S**, **H–**, **H**, **H+** to reflect these distinctions).

6. CONCLUSION

The model presented in this study enables a reversible symbolic coding of intonation with a satisfactory precision of fit by using three symbols for categorizing each direction of pitch movements, plus a symbol for very small or null pitch movements. The model is language independent and requires no prior training on the data. It has been tested on a large corpus comprising 4 hours 20 minutes of speech in five languages, involving fifty speakers. The precision achieved is around 99%, and the model is likely subject to improvement by taking into account the speaker's pitch target point particular distributions. It therefore seems that the proposed model could be of some use for practical applications such as automatic prosodic labeling of large speech databases.

7. REFERENCES

1. Black, A. and Hunt, A. "Generating F_0 contours from ToBI labels using a linear regression." *ICSLP'96*, Philadelphia, 1996.
2. Campione, E. and Véronis, J. "A statistical study of pitch target points in five languages." *ICSLP'98,* Sidney, Australia (in these proceedings), 1998a.
3. Campione, E. and Véronis, J. "A multilingual prosodic database." *ICSLP'98,* Sidney, Australia (in these proceedings), 1998b.
4. 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." *Eurospeech'95,* Madrid, 1, 867-870, 1995.
5. 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.
6. 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.
7. Ostendorf, M. and Ross, K. "A multi-level model for recognition of intonation labels." In Sagisaka, Campbell and Higuchi (Eds), *Computing Prosody.* Springer, Berlin, 291-308, 1997.
8. Taylor, P. "The Rise/Fall/Connection model of intonation." *Speech Communication,* 15:1&2, 169-186, 1994.