Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification

Antonio Origlia, Iolanda Alfano

LUSI-Lab, Department of Physics, University of Naples "Federico II", Italy
Department of Humanities Studies, University of Salerno, Italy
antonio.origlia@unina.it, ialfano@unisa.it

Abstract

Prosodic research in recent years has been supported by a number of automatic analysis tools aimed at simplifying the work that is requested to study intonation. The need to analyze large amounts of data and to inspect phenomena that are often ambiguous and difficult to model makes the prosodic research area an ideal application field for computer based processing. One of the main challenges in this field is to model the complex relations occurring between the segmental level, mainly in terms of syllable nuclei and boundaries, and the supra-segmental level, mainly in terms of tonal movements. The goal of our contribution is to provide a tool for automatic annotation of prosodic data, the Prosomarker, designed to give a visual representation of both segmental and suprasegmental events. The representation is intended to be as generic as possible to let researchers analyze specific phenomena without being limited by assumptions introduced by the annotation itself. A perceptual account of the pitch curve is provided along with an automatic segmentation of the speech signal into syllable-like segments and the tool can be used both for data exploration, in semi-automatic mode, and to process large sets of data, in automatic mode.

Keywords: prosodic analysis, automatic annotation, segmental/supra-segmental events synchronization

1. Introduction

In order to provide a tool for automatic analysis and visualization of the relations between the segmental and the supra-segmental level, it is necessary to combine into a single approach two different speech analysis algorithms aimed, respectively, at speech segmentation into syllabels and to pitch stylization. We present here a brief review of the approaches that can be found in the literature before presenting, in Section 2., the architecture of the Prosomarker tool along with the chosen approaches.

1.1. Automatic syllabification

Syllable segmentation is important in speech processing because it is connected with the main prosodic factors including rhythm and tempo and also because the opinion that syllables can be used as basic units in speech recognition has been investigated for a long time, see for example (Wu et al., 1997; Jones et al., 1997). At the same time, the definition of *syllable* is still controversial. It depends, mainly, on the observed language, on the phonotactical rules involved in the morpho-phonological description adopted for that language and on some particular phonetic constraints. In the field of articulatory phonetics and phonology some authors link syllables with jaw movement (De Saussure, 1967), some others to chest burst (Stetson, 1951). From the acoustics point of view, energy temporal patterns play a fundamental role: (Jespersen, 1920) was the first one to link syllabification with energy oscillation, observing that syllable nuclei are usually found in correspondence with energy maxima, while syllable boundaries correlate with energy minima. A first attempt to automatically segment a speech utterance into syllabic portions was presented in (Mermelstein, 1975). In this work a loudness function obtained by assigning a weight to each element within a set of spectral bands was used. An algorithm evaluating the shape of the loudness pattern (convex-hull) was then employed to find syllable boundaries.

In (Pfitzinger et al., 1996), the speech signal was processed in three steps: first the authors used a bandpass filter, then they computed the energy pattern using a short term window and finally they low-pass filtered this energy function. The syllable nuclei were found by searching the local maxima of the energy contour. Another important result of Pfitzinger and colleagues is the comparison of the different manual syllabic segmentation that were done by several human labelers. They found an agreement of 96% on nuclei positions, making them assume this value as an upper bound for any automatic segmentation.

Another approach for speech syllabification was proposed by (Jittiwarangkul et al., 1998). Their method was based on energy computation and successive smoothing. They tested various kinds of temporal energy functions for syllable boundary detection. The behavior of their algorithm depended on a number of empirically predefined thresholds.

In (Wu et al., 1997) the analysis method was based on smoothed speech spectra computed by two dimensional filtering techniques. This way the energy changes of the order of 150ms were enhanced while other techniques to emphasize the syllable onsets were used. The average energy over nine critical frequency bands every 10ms was also considered. The resulting vector was concatenated with log-RASTA features and was provided as input for a multilayer perceptron.

In (Greenberg and Kingsbury, 1997) the speech modulation spectrogram, a system for searching invariant features related to frequency portions of the speech spectrum, distributed across critical band-like channels, was introduced. According to Greenberg, invariants are mainly positioned in slowly varying dynamic features of the speech signal. The processing and recognition of speech features involves temporal constants that take two kinds of factors into account: speech rhythm parameters and the auditory temporal

integration of the slowest spectral components.
In (Nagarajan et al., 2003) an automatic syllable segmenter using the minimum phase group delay function was developed. The authors' approach is deterministic in the sense that they don't make use of stochastic evaluations about the signal. In their work they try to face the principal problem of the classic approaches to segmentation using the short term energy function, that is thresholding and energy fluctuations. If we consider the short term energy function as a magnitude spectrum, it can be demonstrated that it is associated to a minimum phase signal. The study of the negative derivative of the short term energy function (that is the *group delay function*, if it was a magnitude spectrum) shows that it has peaks at syllable boundaries which are less sensible to energy fluctuations. This approach tries to find a more reliable reference to establish a decision threshold for syllable boundaries. An error rate of utmost 40ms for the 83% of the syllable segments suggests that this is one of the most powerful approaches found in literature. Continuation of this work was also presented in (Prasad et al., 2004).
Lastly, in (Petrillo and Cutugno, 2003), an algorithm employing energy analysis to set syllable boundaries corresponding to energy minima between two maxima was presented. Additional strategies to refine the initial result were employed to avoid segments containing fricative sounds only and to recompact long stressed vowels that were erroneously splitted. The values used for the set of parameters needed to perform automatic syllabification were obtained by using a number of function minimization techniques like genetic programming (Carnahan and Sinha, 2001) and simulated annealing (Kirpatrick et al., 1983).

1.2. Pitch stylization

The definition of pitch stylization given in (t'Hart et al., 1990, p. 42) states that a stylized pitch curve [...] *should eventually be auditorily indistinguishable from the resynthesized original* and [...] *it must contain the smallest possible number of straight-line segments with which the desired perceptual equality can be achieved*. A stylization algorithm is often considered a filter for microprosodic effects and deals with the difficult problem of *perceptual* equality of two pitch curves given the observation in (t'Hart et al., 1990, p. 25) that [...] *no matter how systematically a phenomenon may be found to occur through a visual inspection of F_0 curves, if it cannot be heard, it cannot play a part in communication.*
Among the first attempts to follow this principle, the work presented in (Hirst and Espesser, 1993) has to be highlighted. In this work the pitch curve was considered as the result of the composition of a micro-prosodic component, intended as perturbations caused by mere articulation, and of a macro-prosodic one. The MOMEL algorithm (Hirst and Espesser, 1993; Hirst et al., 2000) was designed to filter out the micro-prosodic component, therefore retaining the macro-prosodic one by means of a quadratic spline function and has been largely used in the past for prosodic analysis tasks.
In (D'Alessandro and Mertens, 1995), a tonal perception model was employed to perform the stylization task. This tonal perception model was built around the concept of *glissando*, a tone perceived as dynamic by the human ear, which has been heavily investigated in a number of psychoacoustics studies (Sergeant and Harris, 1962; Pollack, 1968; Rossi, 1971; Klatt, 1973; t'Hart, 1976; Rossi, 1972; Schouten, 1985). The algorithm presented in (D'Alessandro and Mertens, 1995) was then used to create the *Prosogram* (Mertens, 2004), which has been used in a number of recent studies on intonation (Patel, 2005; Ioannou et al., 2006; Caridakis et al., 2006; Avanzi et al., 2008).
The definition of stylization has been formalized as an optimization process for the first time in (Ghosh and Narayanan, 2009). In that work a Dynamic Programming algorithm was designed to find the optimal balance between an empirically determined number of segments and the Mean Square Error of the stylized curve with respect to the original one. Following the same route, in (Origlia et al., 2011) a *divide et impera* algorithm designed to balance the Normalized Root Mean Square Error and the number of control points used to produce the stylized curve was presented. This interpretation of the stylization process allows the use of well-established and powerful programming techniques and tries to deal with the difficult problem of defining a mathematical formulation of how *good* a stylization is. However, these approaches are also limited by the fact that perceptual phenomena are either difficult to describe, not only in a mathematical sense, or even not completely understood.

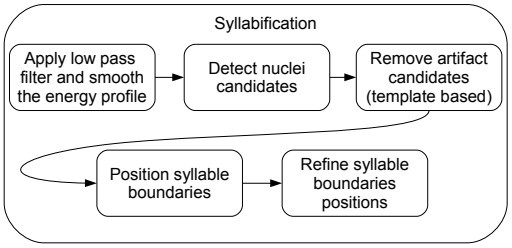

Figure 1: The syllabification process

2. Architecture

The system architecture is composed of two main processes running independently. The first one is dedicated to data extraction from the segmental level. This process extracts the energy profile to detect syllable nuclei and position syllable boundaries accordingly to the energy minima principle (Jespersen, 1920). The syllabification algorithm is a modified version of the method presented in (Petrillo and Cutugno, 2003) integrated with Harmonic Noise Ratio (HNR) analysis to recover isolated syllable nuclei for which the pitch tracking algorithm failed to detect anything. Moreover, artifact peaks caused by phenomena other than syllable nuclei occurrence are removed using a template based detection strategy. The process is summarized in Figure 1. Syllable nuclei length is then estimated by computing the -3db band of energy peaks.
The second process deals with suprasegmental analysis of the speech signal. The tool performs a filtering step designed to remove the effect of microprosody from the pitch

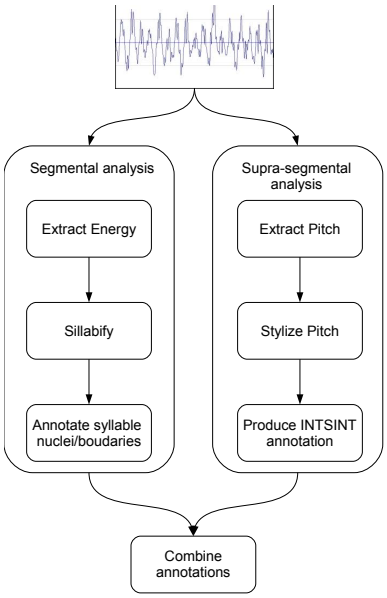

Figure 2: The architecture of the Prosomarker tool

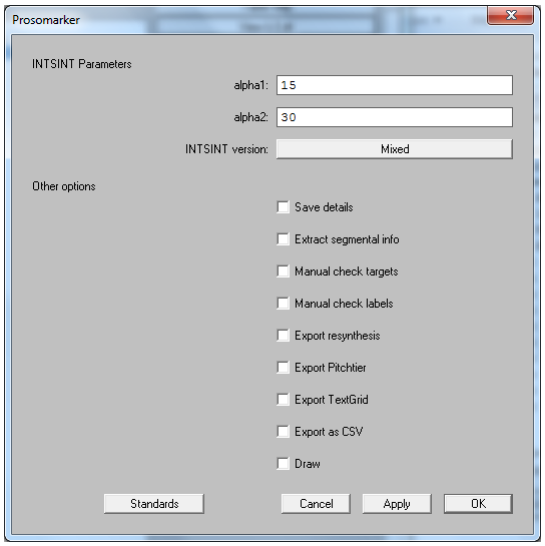

Figure 3: The main interface of Prosomarker

profile by means of the pitch stylization algorithm presented in (Origlia et al., 2011). In this first version of the tool, the INTSINT coding scheme (Hirst et al., 2000) is used to produce the automatic annotation. We chose to employ INTSINT among the various coding schemes because it was specifically designed to annotate the target points of a stylized curve, thus producing a phonetic, rather than phonological, account of intonation. We used the OpS algorithm because it was shown (Origlia et al., 2011) that this algorithm performs similarly to the MOMEL algorithm (Hirst and Espesser, 1993; Hirst et al., 2000) both in terms of perceptual equality and in terms of number of points used while being parameter independent.

While the pitch stylization and annotation process is always performed, segmental analysis and annotation are performed only if the user chooses to visualize this kind of events.

In Figure 2, we summarize the architecture of Prosomarker. The design is modular to let us easily update the tool by working separately on the syllabification algorithm and on the pitch stylization algorithm. Modular independence also leaves open the possibility, in the future, to parallelize the process, thus saving computational time, and to extend the analysis performed. For the implementation, we chose to employ the well known software PRAAT (Boersma and Weenink, 2011) as it contains a large set of primitives to perform phonetic analysis. Also, PRAAT is designed to efficiently handle multilayer annotations in terms of automatic generation, because of the scripting language, in terms of visualization, because of the built-in editors and drawing capabilites, and in terms of compatibility with external software, as the TextGrid format is widely supported.

3. The interface

Prosomarker can provide three different versions of the INTSINT coding scheme: Ampli3, Levels and Mixed (for details, see (Campione et al., 2000)), each one on a different tier. Parameters to control these different versions can be set through the main interface. Since Prosomarker is designed to work on speech corpora, as soon as the user checks the desired options and presses the OK button in the main interface, the tool will ask for the folder in which the audio (WAV) files can be found and, if any of the exporting options is set, it will ask for the folder in which to save results. In Figure 3 we show a screenshot of the interface of the tool.

Prosomarker can run both in automatic and semi-automatic mode: the user can select which steps of the annotation process he/she wishes to check manually and the tool will show the intermediate result waiting for confirmation before proceeding. It is also possible to go back to the target points manual positioning step after visualizing the automatically assigned labels. This allows to check the positioning of the target points, to view and modify the labels and to introduce new tiers in the final TextGrid (for example, to introduce comments). Also, different exporting options are available. Here we summarize all the different options the user can choose to customize how Prosomarker behaves:

- **Save details:** when this option is set, the tool records the actual pitch value (in Hz) corresponding to each target point along with the distance from the preceding target (in seconds) instead of just showing the INTSINT labels.
- **Extract segmental info:** it activates the segmental analysis process. Syllable nuclei and boundaries positions are automatically found and annotated in a separate tier.
- **Manual check targets:** it activates the semi-automatic mode of Prosomarker. After performing the pitch stylization step, the tool will create a Manipulation object and open the corresponding PRAAT editor window in which the user can add target points, remove them or adjust their position.
- **Manual check labels:** it activates the semi-automatic mode of Prosomarker. After performing the automatic

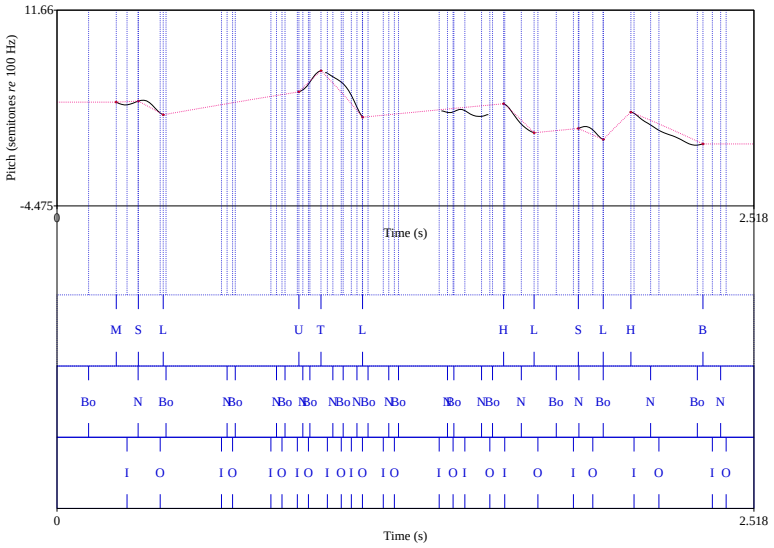

Figure 4: An example of the annotation produced by Prosomarker. First tier: INTSINT labels. Second tier: syllable nuclei (N) and syllable boundaries (Bo). Third tier: extension of syllable nuclei from incipit (I) to offset (O). Syllable nuclei without corresponding pitch were found via HNR analysis

annotation step, the tool opens a PRAAT editor window showing the waveform of the original sound file, its spectrum, pitch and intensity profile along with the produced annotations. Any operation available in PRAAT to manage TextGrids is available at this time. If the *Manual check targets* option was set, the possibility of going back to the target points adjustment step becomes available at run-time.

- **Export resynthesis:** it instructs Prosomarker to generate a resynthesized version of the original sound file in which the stylized pitch curve is substituted to the original one by means of the PSOLA algorithm available in PRAAT. The resulting audio file is saved in the output directory set by the user.

- **Export PitchTier:** it instructs Prosomarker to save the Pitch-Tier object containing the target points used in the stylization process. If the user changes the target points, these changes will be saved. This can be useful to perform further analysis after running the tool.

- **Export TextGrid:** it instructs Prosomarker to save the TextGrid containing the generated annotations. If the user modifies the TextGrid, the changes will be saved. This option is useful to transport data coming from the Prosomarker tool into other software supporting the TextGrid format.

- **Export as CSV:** it instructs Prosomarker to export a Comma Separated Values (CSV) file containing labels and time of occurrence of each label in the final TextGrid. This can be useful to import data into software not supporting the TextGrid format.

- **Draw:** it instructs Prosomarker to draw the original pitch curve along with its stylization and with the aligned TextGrid. In Figure 4 we show an example of the automatic annotations Prosomarker produces generated with this option set.

4. Applications and future development

Prosomarker is an application designed to represent data coming from two algorithms dealing with different linguistic levels. The integrated visualization of these levels is proposed as a framework to provide researchers dealing with prosody an objective account of the occurrence of segmental and suprasegmental events along with their synchronization. Running in semi-automatic mode, the tool can be used both for fast data exploration and as a valid support to a prosodic analysis based on a phonetic approach: labels associated to target points not only provide a coherent description of global prosodic patterns, but they are also related with segmental events in such a way they can reveal linguistic regularities in the relationship between prosodic events and segmental string. Approaching speech from a perspective that tries to account for segmental and prosodic events simultaneously, but independently from each other, applies equally to quantitative and qualitative research strategies and offers the possibility to support a prosodic analysis considering different levels of detail within different theoretical frameworks. Depending on the specific research issues, Prosomarker can be used to analyze prosodic realizations related with linguistic modalities, pragmatic functions or emotion expressions, for instance.

The examples in Figure 5 aim to show how Prosomarker can constitute a valid support in the interpretation of significant differences: they deal with the realizations of the same Italian question “E’ alzata?” (“Is it standing up?”) produced by a native Italian speaker and by a nonnative speaker in which we can observe how Prosomarker’s annotations highlight differences in the comparison between L1 and L2 productions. The native realization is indeed char-

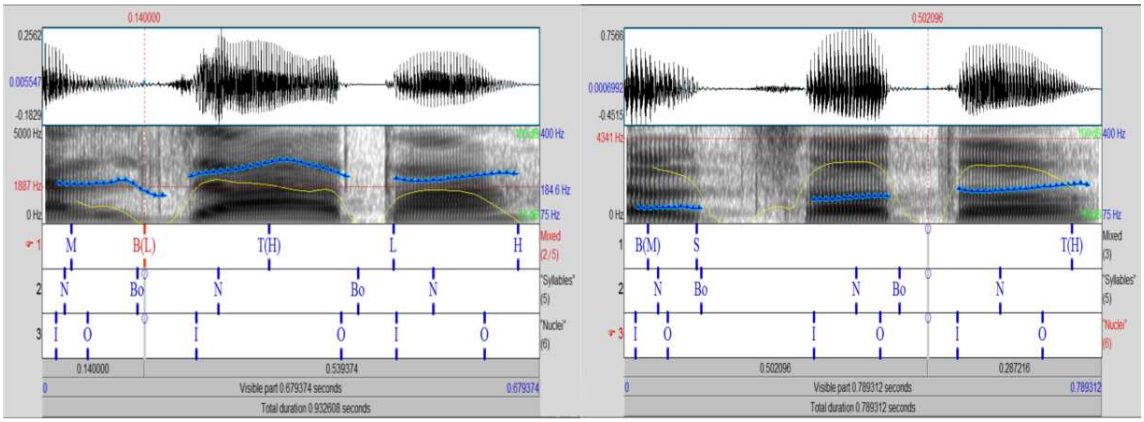

Figure 5: The production of a native Italian speaker (on the left) compared with the production of a nonnative speaker (on the right). On the first tier: annotation labels; on the second tier: f0 differences of each target point compared with the previous one; on the third tier: duration increase; on the fourth tier: syllable nuclei (N) and syllable boundaries (Bo); on the fifth tier: extension of syllable nuclei from incipit (I) to offset (O)

acterized by a rising-falling contour (M - B(L) - T(H) - L - H) in which the maximum f0 value is aligned with the nucleus of the stressed vowel and with an important increase (almost 80 Hz), while the nonnative production presents a rising contour (B(M) - S - T(H)) in which f0 value increases progressively, reaching its maximum value at the end of the utterance.

In automatic mode, the tool can be used to rapidly process large sets of data for subsequent statistical analysis. In particular, the opportunity to produce an abstract representation of intonation involving different linguistic levels at the same time, but keeping them well separated is suitable to perform machine learning tasks. This will be one of the first testing grounds on which we intend to apply the tool. The two modules Prosomarker is composed of are independent so that we can continue work on them separately and obtain an update of the tool each time an update to these subsystems is produced. Modularity can also be exploited to perform parallelization and reduce computational time. Lastly, the automatic detection and representation of events both on the segmental and on the suprasegmental levels will be extended in the future.

5. References

M. Avanzi, A. Lacheret-Dujour, and B. Victorri. 2008. Analor. a tool for semi-automatic annotation of french prosodic structure. In *Proc. of Speech Prosody*, pages 119–122.

P. Boersma and D. Weenink. 2011. Praat: doing phonetics by computer [computer program]. version 5.2.40.

E. Campione, D. Hirst, and J. Véronis. 2000. Stylisation and symbolic coding of f0: comparison of five models. In A. Botinis, editor, *Intonation: analysis, modeling and technology*, pages 185–208, Dordrecht. Kluwer.

G. Caridakis, L. Malatesta, L. Kessous, and A.and Karpouzis K. Amir, N.and Paouzaiou. 2006. Modeling naturalistic affective states via facial and vocal expression recognition. In *Proc. of ICMI*, pages 146–154.

J. Carnahan and R. Sinha. 2001. Nature's algorithms [genetic algorithms]. *IEEE Potentials*, 20(2):21–24.

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

F. De Saussure. 1967. *Course de linguistique generale*. Laterza, payot. italian edition by t. de mauro edition.

P. K. Ghosh and S. Narayanan. 2009. Pitch Contour Stylization Using an Optimal Piecewise Polynomial Approximation. *IEEE Signal Processing Letters*, 16(9):810–813.

S. Greenberg and B. E. Kingsbury. 1997. The modulation spectrogram: In pursuit of an invariant representation of speech. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, pages 1647–1650.

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

D.J. Hirst, A. Di Cristo, and R. Espesser. 2000. Levels of representation and levels of analysis for the description of intonation systems. In M. Horne, editor, *Prosody: Theory and Experiment Studies*, Dordrecht. Kluwer.

S. Ioannou, L. Kessous, G. Caridakis, K. Karpouzis, V. Aharonson, and S. Kollias. 2006. Adaptive on-line neural network retraining for real life multimodal emotion recognition. In *Proc. of ICANN*, pages 81–92.

Jespersen. 1920. *Lehrbuch der Phonetic*. B.G. Teubner, Leipzig e Berlin.

N. Jittiwarangkul, S. Jitapunkul, S. Luksaneeyanavin, V. Ahkuputra, and C. Wutiwiwatchai. 1998. Thai syllable segmentation for connected speech based on energy. In *Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems (APCCAS'98)*, pages 169–172.

R. J. Jones, S. Downey, and Mason. J. S. 1997. Continuous speech recognition using syllables. In *Proceedings of Eurospeech '97*, pages 1171–1174.

S. Kirpatrick, C.D. Gelatt, and M.P. Vecchi. 1983. Optimization by simulated annealing. *Science*, pages 671–680.

D. H. Klatt. 1973. Discrimination of fundamental fre-

quency contours in synthetic speech: implications for models of pitch perception. *Journal of the Acoustical Society of America*, 53:8–16.

D. Mermelstein. 1975. Automatic segmentation of speech into syllabic units. *Journal of Acoustical Society of America*, 54(4):880–883.

P. Mertens. 2004. The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In *Proc. of Speech Prosody*.

T. Nagarajan, H. A. Murthy, and R. M. Hegde. 2003. Segmentation of speech into syllable-like units. In *Proceedings of Eurospeech 2003*, pages 2893–2896.

A. Origlia, G. Abete, C. Cutugno, I. Alfano, R. Savy, and B. Ludusan. 2011. A divide et impera algorithm for optimal pitch stylization. In *Proc. of Interspeech*, pages 1993–1996.

A. D. Patel. 2005. The relationship of music to the melody of speech and to syntactic processing disorders in aphasia. *Annals of the New York Academy of Sciences*, 1060:59–70.

M. Petrillo and F. Cutugno. 2003. A syllable segmentation algorithm for English and Italian. In *Proc. of Eurospeech*, pages 2913–2916.

H. Pfitzinger, S. Burger, and S. Heid. 1996. Syllable detection in read and spontaneous speech. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP)*, pages 1261–1264.

I. Pollack. 1968. Detection of rate of change of auditory frequency. *Journal of experimental psychology*, 77:535–541.

V. K. Prasad, T. Nagarajan, and H. A. Murthy. 2004. Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communication*, pages 429–446.

M. Rossi. 1971. Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica*, 23:1–33.

M. Rossi. 1972. Interactions of intensity glides and frequency glissandos. *Language and Speech*, 21:384–394.

H. E. M. Schouten. 1985. Identification and discrimination of sweep tones. *Perception and psychophysics*, 37:369–376.

R. L. Sergeant and J. D. Harris. 1962. Sensitivity to unidirectional frequency modulation. *Journal of the Acoustical Society of America*, 34:1625–1628.

R. Stetson. 1951. *Motor Phonetics*. North Holland, Amsterdam.

J. t'Hart, R. Collier, and A. Cohen. 1990. *A Perceptual Study of Intonation: An Experimental-Phonetic Approach*. Cambridge University Press, Cambridge.

J. t'Hart. 1976. Psychoacoustic backgrounds of pitch contour stylization. Technical report, IPO – annual progress report.

S. Wu, M. L. Shire, S. Greenberg, and N. Morgan. 1997. Integrating syllable boundary information into speech recognition. In *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, pages 987–990.