AUTOMATIC DETECTION OF PROSODIC PROMINENCE BY MEANS OF ACOUSTIC ANALYSES

Fabio Tamburini

Abstract: Prosodic prominence is commonly regarded as the perceptual salience of a linguistic unit relative to its environment. However, we are far from having a consensus on how it is measured subjectively and how it relates to objectively measurable acoustic events or linguistic structures such as lexical stress, prosodic focus, etc. Here we will concentrate mainly on the identification of prominence by means of acoustic parameters and automatic techniques. Considering this topic, some questions are still open in the community: (a) How can we reliably define and portray prosodic prominence? (b) What is the best prominence domain in acoustics? (c) Is prominence a continuous or a discrete phenomenon? (d) What are the acoustic parameters that support it and how can we combine them to reliably identify prominence? (e) To what extent are acoustic parameters language specific? Can we identify universals across languages? (f) What is the best paradigm for the automatic identification of prominence: Rule-Based or Machine Learning Systems? (g) How can we evaluate automatic systems in the right way? This contribution will briefly address these points.

Keywords: automatic detection, evaluation, prominence, prosody.

1. INTRODUCTION

Speakers use prominence at syntactic, semantic and pragmatic level to draw listener's attention on specific utterance segments, to express their emotion or attitude about the topic being discussed, to indicate the focus of an utterance, to disambiguate between different readings, to mark the introduction of new topics, to indicate the information status of a unit (new or given), to change speaking style, etc.

In natural-language applications, automatic prominence detection has been found to be very important for spoken language understanding, especially for Automatic Speech Recognition and for the production of Dialogue systems, and to improve Text-To-Speech systems naturalness and intelligibility (Windmann et al. 2010).

This contribution is mainly devoted to review the main problems we have to face for the automatic prominence identification. We will not delve

into the different phonological theories describing this phenomenon, but we concentrate our discussion at the phonetic/acoustic level.

1.1 How can we reliably define and portray prosodic prominence?

A careful examination of the literature on prosodic prominence reveals remarkable terminological problems. Various studies (Bertinetto 1981; Jensen 2004; Spencer 1996; Taylor 1992; Wightman & Ostendorf 1994) emphasized that the terminology used to describe such phenomena is quite heterogeneous. Each single term tends to refer to different linguistic parameters in different works. Thus, it seems appropriate to define clearly prominence and all related concepts.

One of the most quoted definitions of prominence is due to Terken (1991: 1768):

> Prominence is the property by which linguistic units are perceived as standing out from their environment.

In other words, it is a phenomenon that assigns some degree of saliency to specific units. Jensen (2004: 27), in his definition, emphasized similar properties:

> The term Prominence […] generally refers to the degree to which something stands out from its surroundings. It may be used about specific properties, such as pitch prominence, […], or more generally, as perceived prominence, about the overall degree of emphasis (or de-emphasis) of a certain item.

Mertens (1991: 218; see also Ladd et al. 1994) put the accent on the continuous nature of prominence:

> A syllable is prominent when it stands out from its context due to a local difference for some prosodic parameter. Prominence is continuous (not categorical) and contributions of multiple parameters interact.

The two following definitions reflect the mentioned terminological problems:

> What many phoneticians and linguists have called stress, and what most laymen readily understand under this term, refers to nothing more than the fact that in a succession of spoken syllables or words some will be perceived as more salient or prominent than others. (Couper-Kuhlen 1986: 19)

> The term sentence-accent refers to the perceptual salience of some words over others in utterances, […] (Kohler 2006: 749)

Kohler refers to prominence as SENTENCE-ACCENT, while Couper-Kuhlen notes that traditionally the term *stress* is used with the same meaning of the term *prominence*, often without giving a precise reference to lexical stress or sentence stress, two distinct concepts even if related in some way.

Summarising, we can say that:

> prosodic prominence is a perceptual phenomenon, continuous in its nature, emphasizing some linguistic and segmental units with respect to their surrounding context, and supported by a complex interaction of prosodic and phonetic/acoustic parameters.

1.2 What is the best prominence domain in acoustics?

Most prosodic prominence definitions rely on a generic notion of SEGMENTAL UNIT to define the domain for this phenomenon. Thus it seems relevant to our discussion trying to understand what could be the best segmental unit supporting the automatic analysis of prominence from a phonetic/acoustic point of view.

There is a large agreement among scholars to consider the syllable as the prominence-bearing unit in connected speech. Unfortunately, defining the syllable in a phonetic domain is often misleading (Kopecek 1999; Noetzel 1991; Pfitzinger et al. 1996; Wu et al. 1997) and, consequently, the automatic segmentation of the utterance into syllables is a challenging task.

However, a lot of studies have made clear that the main contribution of prominence to syllable is concentrated in the nucleus (Silipo & Greenberg 1999; Tamburini & Caini 2005; van Kuijk & Boves 1999), thus the relevant conclusion for this discussion is that, from an acoustical point of view, we can identify the syllable nucleus as the relevant unit for prominence detection as in (Jenkin & Scordilis 1996; Waterson 1987).

If the corpus data are segmented, we can use manual segmentations to derive reliable syllable boundaries totally avoiding such kind of problems, but, in most real applications, syllable or nuclei segmentations are not available, and devising accurate procedures to acoustically estimate syllable segmentation is often required. In this direction, some studies, devoted to the automatic identification of pseudo-syllable or phonetic syllables in speech (Mermelstein 1975; Origlia et al. 2014), propose segmental units that can be successfully used also for the automatic prominence identification.

1.3 Should we introduce categories in prominence perception?

In principle, acoustic features supporting prominence perception, being physical quantities, are intrinsically continuous and then prominence can be

considered as a continuous phenomenon, at least from a physical/acoustical point of view. But, how perceptual prominence is categorised? How many classes should we use in the linguistic resources usually employed for training/testing automatic prominence detection systems?

Although perceptual categorisation has been studied in cognitive sciences for years (Cohen & Lefebvfre 2005; Goudbeek et al. 2005; Holt & Lotto 2010), the process of *parsing* the multidimentional continuous space of acoustic features into discrete prominence categories still contains several open issues.

The adoption of binary scale annotations (prominent vs. non-prominent) is certainly the most widespread approach. On the one side, it is quite easy for annotators to identify prominence on a binary scale, but, on the other side, we get a coarse-grained annotation and we usually miss most of the complexity in prominence perception.

We can find a variety of multilevel scales proposed in literature involving different number of levels (we cite here only one reference per type): 3 levels (Goldman et al. 2010), 4 levels (Jensen & Tøndering 2005), 11 levels (Turk & Sawusch 1996), 31 levels (Wagner 2005). Most of these works introduce a scale type and measure the inter-annotator agreement obtained by using such scale, but do not really define it by setting up proper perceptual experiments. In order to solve such problem, Mehrabani et al. (2013) presented an interesting study to induce the optimal number of classes using psychologically/perceptually defined Just-Noticeable Differences in the acoustic parameters concluding that a 4-point scale for prominence annotation is the best choice for resource annotation. Multilevel annotation scales are certainly harder for annotators, but their use enable the possibility to perform interesting measures on prominence profiles, especially for 11 or 31-points-scales (Arnold et al. 2012; Wagner et al. 2012). Specific studies comparing the different scales often lead to contradictory results (Grover et al. 1997; Jensen & Tøndering 2005), but the work in (Arnold et al. 2011) presented some results that seems to favour the choice of multilevel scales composed by a large number of classes.

Typical inter-annotator agreement for prominence annotation is in the range 0.7-0.8 (Pearson rho) for multilevel scales and over 80/85% of matchings for binary scales.

Defining the right number of levels seems quite critical in order to build phonological theories for prosodic prominence, but, from the acoustical point of view, prominence can be safely considered on a continuous or a multilevel scale consisting of a large number of classes.

1.4 What are the acoustic parameters that support prosodic prominence and how can we combine them to reliably identify it?

One of the most pressing questions on this topic, widely studied in a large set of studies, regards the possibility to find measurable linguistic/prosodic phenomena supporting prominence perception. To this extent, we would like to refer primarily to the work of Klaus J. Kohler, for its clarity:

> The category of sentence accent [prominence] is a separate prosodic level from intonation, controllable independently from rhythm, syllabic and segmental structuring, on a scale from 1 to 3. Although it shares F0 as a physical property with intonation, it is not entirely determined by it, but also depends on syllable and segment duration, intensity, and possibly other features. (Kohler 2003: 2930)

> …, it became clear that beside the accent category that is principally signalled by F0 excursion and may therefore be called pitch accent, another type of accent has to be recognised that is primarily related to non-pitch features, viz. acoustic energy, based on phonatory and articulatory force, and may therefore be called force accent. (Kohler 2005: 99)

In this view two main "actors" join their contributions in supporting perceptual prominence (or sentence accent), as interacting and mutually reinforcing phenomena at linguistic-prosodic level (see also Ladd 1996). The first, PITCH ACCENT, is similar to the concept introduced by Bolinger (1958) and concerns specific configurations in F0 profile. The second, FORCE ACCENT, is completely independent from intonational information inside the utterance and it is connected with acoustic phenomena such as intensity, segmental durations and possibly others.

Assuming this view, we can try to cast a first, partial formalisation of prominence, where the two accentual typologies suggested by Kohler both contribute to support sentence prominence. These relationships can be mathematically described by the equation $Prom^i=FA^i+PA^i$, where *FA* and *PA* are respectively the contributions of force accents and pitch accents, both referred to the *i*-th segmental unit inside the utterance.

One of the major challenges in predicting syllable prominence is the disentangling of the various acoustic sources of influence, such as fundamental frequency excursions, duration, intensity related acoustic parameters and listeners' linguistic expectancies. At the acoustic level, various studies (Anastakos et al. 1995; Bagshaw 1994; Heldner 2003; Sluijter & van Heuven 1996, 1997; Streefkerk 1996) suggested, also in an interlinguistic perspective, a dependence of force accents to unit duration and spectral emphasis, while pitch accents would be supported mainly by pitch movements

and by the global, wide-band, intensity inside a particular segmental unit. In the past, we led some experiments confirming such relations for some languages (Tamburini 2003, 2005a, 2006).

1.5 To what extent are acoustic parameters language specific? Can we identify universals across languages?

Jun (2005) proposed a phonological model of prosodic typology based on Autosegmental-Metrical models that considered two different aspects of variation (this view is supported by other scholars, for example by Fitzpatrik 2000). The first dimension is PROMINENCE which classifies languages into four categories: (1) stress-accented, (2) lexical pitch-accented, (3) non stress-accented and non lexical pitch-accented and (4) tonal. This classification is fairly uncontroversial, but it is often seen as a continuum between the categories. The second dimension regards RHYTMIC PATTERN. The traditional classification of languages into three classes, (a) stress-timed, (b) syllable-timed and (c) mora-timed is more controversial and the concept of isochrony is often seen as problematic. There are experimental studies that strongly support this view (Ramus et al. 1999; Low et al. 2001) and others that raise critic judgments against this classification of rhythmic patterns in languages (Pamies Bertran 1999; Warner & Arai 2001). More recent studies tend to consider rhythm as an auditory phenomenon connected with prominence patterns instead of durational measures of segmental units (Russo 2010).

We would like to avoid any consideration about the second dimension proposed by Jun, namely the rhythmic pattern, concentrate our efforts on the prominence dimension and propose a different perspective for building a prosodic typology framework based mainly on phonetic/acoustic parameters instead on phonological theories.

The limited space for this contribution does not allow to properly describe the large number of studies on the relevance of the various acoustic parameters in the different languages. What we can briefly say is that in the literature emerged a clear tendency to identify a small number of acoustic parameters that, to some extent, can have an influence to prominence perception (specific configurations in pitch profile, segmental units duration, intensity, energy measures in different spectral bands, as discussed also in Section 1.4). Languages tend to assign a different level of importance to these parameters, but, more or less, all these acoustic measures play a role in inducing the perception of a prominent unit cross-linguistically.

2. AUTOMATIC PROMINENCE DETECTION

As we have seen in the previous sections, the perception of prosodic prominence is supported by a complex interaction of multiple acoustic and prosodic features. In order to build a computational model for solving a general classification problem we have to design a formal model for feature combination, a common task in speech processing applications.

2.1 What is the best paradigm for the automatic identification of prominence: Rule-Based or Machine Learning Systems?

In Natural Language Processing two radically different approaches can be applied to formalise and define a computational model to solve a specific problem: RULE-BASED and MACHINE-LEARNING methods. The first approach requires the linguists to provide the formal "rules" able to solve the problem, while the latter set of methods can derive a computational model automatically by examining real data manually annotated by experts with the different problem classes. Table 1 outlines the main advantages and disadvantages of both approaches.

RULE-BASED METHODS	MACHINE-LEARNING METHODS
PROS: - allow linguists to combine features using previous results in literature; - allow for linguistic explanations of the models; - allow for total control on the algorithm behaviour; - allow for the creation of a multilingual detector; - do not require large annotated corpora for building the model; CONS: - apply a deductive approach, they do not use any data for building the model; - typically, produce less accurate systems.	PROS: - apply inductive approaches, the model is learned from data; - allow for fast prototyping; - typically produce high performance detectors/classifiers. CONS: - require large annotated corpora for learning; - produce systems bounded to the specific corpus or language variety used during the learning phase; - produce models that are not "readable", we typically have no way to extract useful linguistic information from the learned model.

TABLE 1. RULE-BASED AND MACHINE LEARNING METHOD COMPARISON.

Considering the automatic prominence detection, there are a lot of studies for the construction of high performance classifiers that rely on both methodologies: Abete et al. (2010), Goldman et al. (2012), Ludusan et al. (2011), Tamburini (2006, 2007, 2009) and Vainio et al. (2013) present some recent examples of rule-based system, while in Arnold et al. (2013), Brenier et al. (2005), Cutugno et al. (2012), Li et al. (2011), Tamburini et

al. (2014) some machine-learning techniques were successfully applied to prominence detection in connected speech.

Let us revise briefly the main issues of these methodologically different approaches when applied to our task, starting from machine-learning models.

2.1.1 Machine-Learning Systems (MLS)

Adopting the definition proposed in Section 1.1, we consider prominence as a phenomenon establishing precise syntagmatic relations with respect to the neighbouring syllables. Its identification requires MLS able to properly model sequences of events, because the immediate context information, both in the feature sequence of the input and in the label sequence of the output, are crucial for the correct identification of syllable prominence. We cannot assert that a segmental unit is prominent without comparing it with the immediate context. Thus, the appropriate point of view to handle such kind of problem is considering the speech stream as a sequence of events and build models able to capture the complex relationships between neighbouring units.

Classical machine-learning methods (Naïve Bayes, Support Vector Machines, classical Artificial Neural Networks, etc.) do not handle sequences of events properly. They make the classification decision only considering a concatenation of input features and excluding the contribution of the previous decisions in the global model.

On the contrary, Probabilistic Graphical Models (PGM) (e.g. Conditional Random Fields, Conditional Neural Fields, etc.) taking advantage of discriminative stochastic models, can successfully handle recognition problems that heavily depend on sequences. PGM are powerful frameworks for representation and inference in multivariate probability distribution. They use a graph-based representation as the basis for compactly encoding a complex distribution over a high-dimensional space representing the conditional dependence structure between random variables. The work presented in (Cutugno et al. 2012; Tamburini et al. 2014) shows a clear advantage of this class of models, in terms of global recognition performances, when classifying prosodic prominence at syllable level.

2.1.2 Rule-based Systems

In typical rule-based systems the developer must explicitly formalise all relationships among the speech features involved in the studied phenomenon. To exemplify this approach we will briefly present the rule-based system described in Tamburini (2006, 2007, 2009). In order to build such model we refer to the discussion about the Kohler's work presented in a previous section and extend this model to include acoustic features.

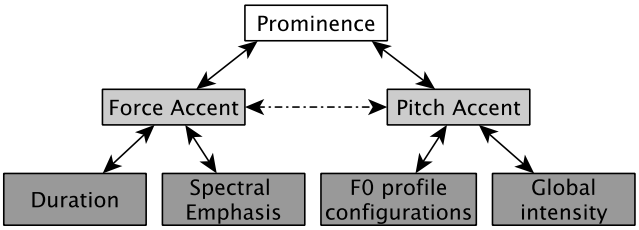

FIGURE 1. RELATIONS BETWEEN PERCEIVED PROMINENCE, LINGUISTIC-PROSODIC PARAMETERS AND ACOUSTIC PARAMETERS AS PROPOSED IN THE RULE-BASED SYSTEM BY TAMBURINI (2006, 2007, 2009).

ACOUSTIC FEATURE	DESCRIPTION
Nucleus Duration	Duration of the syllable nucleus normalised w.r.t. mean and variance duration of the syllable nuclei in the utterance (z-score), as based on the manual segmentation available in the database or detected automatically as described in (Tamburini, 2006).
Spectral emphasis	Normalised SPLH-SPL parameter (Fant *et al*. 2000) (z-score).
Pitch movements	Computed as the product of A_{event} and D_{event} parameters of the TILT model representation (Taylor, 2000) of pitch movements. The raw pitch contour is the median of three pitch tracking algorithms (Tamburini, 2013): RAPT (Talkin, 1995), SWIPE' (Camacho, 2007) and YAAPT (Zahorian, Hu, 2008). The raw pitch profile was stylised by using a quadratic spline function, interpolating the control points derived from the OpS algorithm proposed in (Origlia *et al*. 2013).
Overall intensity	RMS energy computed in the frequency band 50-5000 Hz, normalised using mean and variance (z-score).

TABLE 2. DETAILS OF FEATURE COMPUTATION FOR THE RULE-BASED SYSTEM PRESENTED IN TAMBURINI (2006, 2007, 2009) AND THE SUCCESSIVE SYSTEM DEVELOPMENTS.

Considering the relationship between prominence and acoustic parameters seen in Section 1.4, we can propose a hierarchical model of prominence perception based on acoustic measures inside the utterance. Figure 1 outlines the relations between perceived prominence, linguistic-prosodic parameters and acoustic parameters as proposed in Tamburini (2006, 2007, 2009) and Table 2 shows the computational procedures used to extract the relevant acoustic features from the utterance waveform.

Starting from these acoustic parameters and considering the relationships outlined before we can introduce a prominence function which is able to assign a continuous prominence level to each syllabic nucleus, using only acoustic information:

$$\begin{aligned} \mathrm{Prom}^{i} \;=\; & W_{FA} \cdot \left[SpEmph^{i}_{SPLH-SPL} \cdot dur^{i} \right] + \\ & W_{PA} \cdot \left[en^{i}_{ov} \cdot \left(A^{i}_{event}(at_{M}, at_{m}) \cdot D^{i}_{event}(at_{M}, at_{m}) \right) \right] \end{aligned}$$

where $SpEmph_{SPLH\text{-}SPL}$ is the spectral emphasis, *dur* is the nucleus duration, en_{ov} is the overall energy in the nucleus and A_{event} and D_{event} are the parameters derived from the TILT model as a function of the maxima alignment type - at_M - and the minima alignment type - at_m (see Figure 2). All parameters are referred to the generic syllable nucleus *i*.

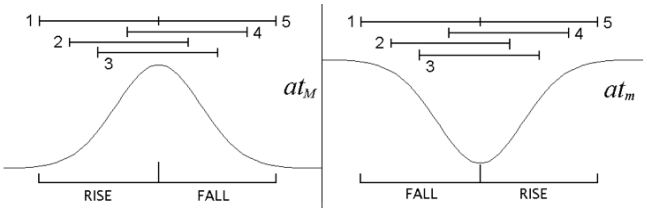

Figure 2. Alignment type parameters between pitch accents and syllable nuclei.

In our model, W_{FA} and W_{PA} weigh the contribution of the two different accent types, while at_M and at_m model the different pitch accent alignments specific to each language. For example, if at_M=1 and at_m=3 the rise section of the maxima and the center of the minima in the F0 profile will be taken as reference points to assign the pitch accent to the corresponding nucleus.

The body of the function *Prom* contains nine parameters, five of them can be considered as supporting the prominence phenomenon from a cross-linguistic point of view ($SpEmph_{SPLH\text{-}SPL}$, *dur*, en_{ov}, A_{event} and D_{event}), while the other four (W_{FA}, W_{PA}, at_M, at_m) can be seen as language specific. It would be intriguing testing such hypothesis on different languages, but the lack of reliable, shared and annotated resources prevent from setting up similar large-scale experiments. We did some work following this idea (Tamburini 2005b, 2009), but it is still largely incomplete.

The *Prom* function provides a prominence measure assigning to each syllable a value in a continuous domain. In general, as we said before, the binary classification into "prominent" or "not prominent" classes cannot be carried out, at least in an optimal way, if the context of the neighbouring syllables is neglected. Thus, considering the syntagmatic nature of the prominence phenomenon, identifying prominent syllables implies a search for the local maxima of the *Prom* function. In this rule-based classifier the prominence value of each syllable nucleus is compared with the two neighbours and, if it represents a maximum, then the corresponding syllable is considered prominent (some corrections are made to deal with special cases).

Figure 3 shows a computed prominence profile compared with the manual annotation for an utterance taken from the TIMIT corpus (Garofolo et al. 1993).

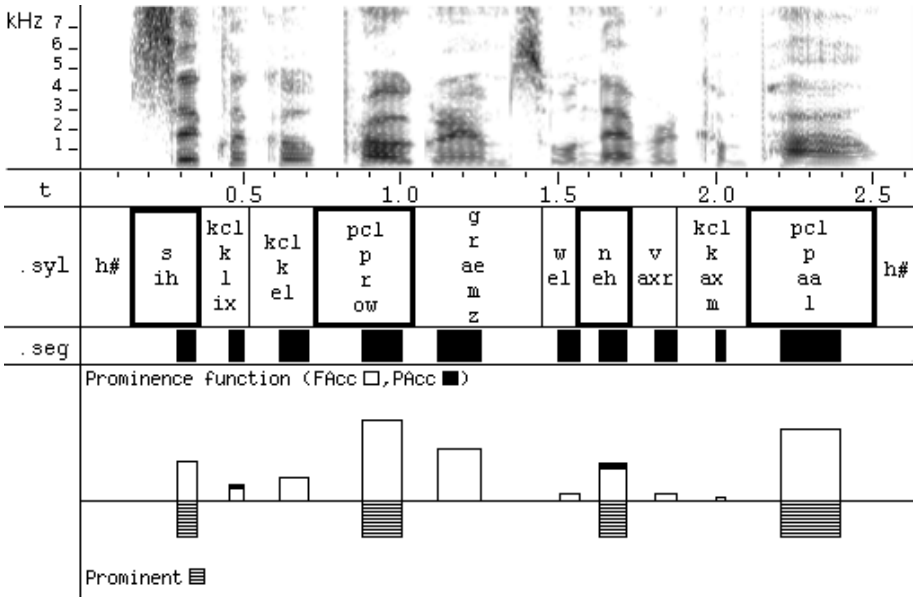

FIGURE 3. PROSODIC PROMINENCE FUNCTION VALUES FOR THE UTTERANCE *CYCLICAL PROGRAMS WILL NEVER COMPILE* (DR1/FDAW0/SX146) FROM THE TIMIT CORPUS. PROCEEDING FROM THE TOP, WE HAVE: THE SPECTROGRAM PLOT, THE SYLLABLE SEGMENTATION (ONLY FOR COMPARISON PURPOSES), THE SYLLABLE NUCLEI AS DETECTED BY THE AUTOMATIC SYSTEM (A BLACK BOX IN THE ".SEG" TIER), AND FINALLY THE PROMINENCE VALUES OF THE FUNCTION *PROM* FOR EVERY NUCLEUS IDENTIFIED BY THE SEGMENTATION PROCEDURE (ABOVE THE AXIS). THE PROMINENT NUCLEI, AS IDENTIFIED BY THE AUTOMATIC SYSTEM, ARE MARKED BELOW THE AXIS (WITH STRIPED BOXES), WHILE PROMINENT SYLLABLES, AS CLASSIFIED BY A HUMAN LISTENER, ARE INDICATED BY A THICK BOX IN THE SYLLABLE SEGMENTATION TIER (".SYL").

3. EVALUATION OF THE AUTOMATIC SYSTEMS

Once we have created an automatic detection/classification system, it is advisable to set up proper experiments in order to carefully evaluate the performances of the new system. This procedure must precede any presentation of the system as it validates the results and the methodological choices taken during system design. For these reasons it is very important to carefully devise the evaluation experiments, the choice of the test data, the evaluation metrics, etc.

Automatic prominence detection systems must follow the same "production cycle" and proper evaluation experiments have to be designed in order to validate the algorithm behaviour, but, depending on the different annotation schema, we have to set up different evaluation procedures.

When handling with prominence annotations, we have three layers of prosodic prominence (Wagner et al. 2012): subjective perceptual, objective acoustic, expectancy based. They are at least partly independent of each

other, however, still many models make the implicit assumption that they should be mirror images especially when setting up evaluation procedures. Interpreting a mismatch between two levels as a "mistake" in one level or the other could led to wrong interpretations, for example a mismatch between subjective prominence perception and an objective acoustic prominence model does not necessarily mean that we have been looking at the wrong acoustic measures of prominence. It could be the case that subjective perception was based primarily on linguistic expectations or that the annotation design, i.e. the subjective perceptual model, was inadequate.

3.1 Binary annotations

In case of binary prominence annotations, general metrics to measure system performances such as Accuracy/Error Rate or Precision/Recall/F-score have been widely used even if the former metrics are not optimal for this kind of problem because of the typical imbalance between the two prominence classes (in real data we have usually less prominent syllables than non-prominent syllables).

3.2 Multilevel or continuous annotations

In Wagner et al. (2012) we considered the problem of evaluating prominence perception annotated, manually or automatically, by continuous profiles. We treated prominence as a continuous variable, not inherently limited to a predefined number of levels. In order to compare two prominence profiles, seen as continuous functions over discrete values, we have to define a specific metric function able to capture the kind of comparison linguists have in mind when judging prominence profiles. We are interested in developing a measure able to (a) verify that the local maxima in the profiles are located on the same syllables and (b) the different heights of these local maxima in the two profiles draw similar pictures. Considering that a local maximum is a point where we perceive a prominence, requirement (a) regards the concept that prominent syllables in the two profiles should match. Requirement (b) ensures that the relative importance of these maxima is respected, evaluating their relative height and prominence strength. These two constraints are very different and require different approaches for measuring the degree of congruence between two profiles: the first asks for a local measure, while the second implies a global measurement and comparison of the two profiles.

We will try to develop this idea by using the examples in Figure 4, where the reference profile A is compared with other profiles. Qualitatively, the linguist would expect that B1, when compared to A, will obtain a me-

dium score, because the two maxima are in the same position but have different heights. B2 should obtain a very low score, because there is no correspondence between the maxima at all, while B3 should get a high score because there are only slight differences in the two profiles.

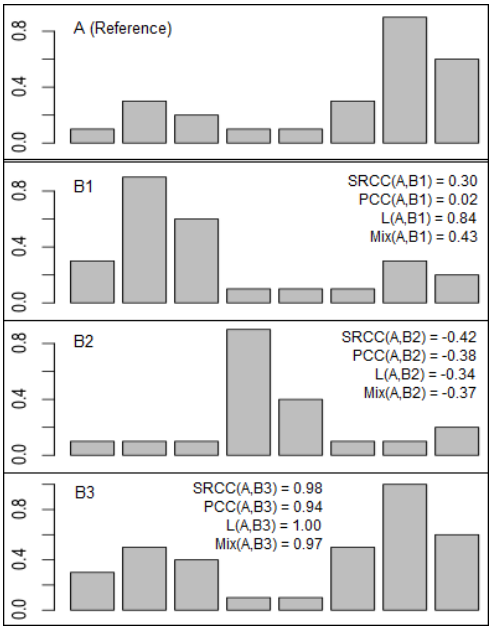

FIGURE 4. SOME PROMINENCE PROFILES AND THE VARIOUS EVALUATION MEASURE VALUES.

There are various methods for comparing two functions that span over continuous values, the most common ones certainly being correlation coefficients. In particular, the Pearson Correlation Coefficient (PCC) and Spearman Rank Correlation Coefficient (SRCC) have been used in various studies for comparing prominence profiles (Heuft et al. 2000; Wagner 2002; Tamburini & Wagner 2007). Unfortunately, as shown in Figure 4, these measures are influenced by the distribution of values in the whole utterance and fail to capture the local correspondence of maxima in the two profiles; they represent good measures of the global matching of the profiles and the degree of matching between the heights of the local maxima. Comparing profile A with B1, we expect, despite the difference in height of the maxima, to have a medium value of correlation, but both SRCC and PCC return low correlation values. The other two examples, namely B2 and B3, behave as expected, providing low values for B2 and high scores for B3.

We therefore consider the PCC as a measure of the global similarity between two profiles A and B and define a local similarity measure (L) by averaging the contribution of the different portions of utterances, compared

through a sliding window. Given the two measures of local and global matching between profiles we can combine them into a unique matching measure (*Mix*) by averaging them. Please, see all mathematical details in (Wagner et al. 2012).

When comparing the behaviour of the *Mix* similarity measure on the examples in Figure 4, we see that it resembles the linguists' intuition of prominence profile comparison, providing a mid similarity score to the first example B1 and a low and a high score respectively to examples B2 and B3.

REFERENCES

Abete, G., F. Cutugno, B. Ludusan & A. Origlia (2010). Pitch behaviour detection for automatic prominence recognition. In *Proceedings of Speech Prosody 2010*, Chicago, 102001:1-4.

Anastasakos, A., R. Schwartz & H. Shu (1995). Duration modelling in large vocabulary speech recognition. In *Proceedings of ICASSP '95*, 628-631.

Arnold, D., P. Wagner & B. Möbius (2011). Evaluating different rating scales for obtaining judgments of syllable prominence from naive listeners. In *Proceedings of 17th ICPhS 2011*, Hong Kong, 252-255.

Arnold, D., P. Wagner & B. Möbius (2012). Obtaining prominence judgments from naïve listeners. Influence of rating scales, linguistic levels and normalisation. In *Proceedings of Interspeech 2012*, Portland, SS09.04.

Arnold, D., P. Wagner & R. H. Baayen (2013). Using generalized additive models and random forests to model prosodic prominence in German. In *Proceedings of Interspeech 2013*, Lyon, 272-276.

Bagshaw, P. (1994). *Automatic prosodic analysis for computer-aided pronunciation teaching*. Edinburgh: University of Edinburgh Ph.D. dissertation.

Beckman, M. E. (1986). *Stress and non-stress accent*. Dordrecht: Foris.

Bertinetto, P. M. (1981). *Strutture prosodiche dell'italiano*. Firenze: Accademia della Crusca.

Brenier, J. M., D. M. Cer & D. Jurafsky (2005). The detection of emphatic words using acoustic and lexical features. In *Proceeedings of Interspeech 2005*, Lisbon, 3297-3300.

Bolinger, D. (1958). A theory of pitch-accent in English. *Word* 14. 109-149.

Camacho, A. (2007). *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*. Gainesville, FL: University of Florida Ph.D. dissertation.

Cohen, H. & C. Lefebvre (2005). *Handbook of categorization in cognitive science*. Amsterdam: Elsevier.

Couper-Kuhlen, E. (1986). *English prosody*. London: Edward Arnold.

Cutugno, F., E. Leone, B. Ludusan & A. Origlia (2012). Investigating syllable prominence with conditional random fields and latent-dynamic conditional random fields. In *Proceedings of Interspeech 2012*, Portland, SS09.06.

Fant, G., A. Kruckenberg & J. Liljencrants (2000). Acoustic-phonetic analysis of prominence in Swedish. In A. Botinis (ed.), *Intonation*, 55-86, Dordrecht: Kluwer.

Fitzpatrik, J. (2000). On intonational typology. In P. Siemund (ed.), *Methodological issues in language typology. Sprachtypologie und Universalienforshung* 53. 88-96.

Garofolo, J. S., L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett & N. L. Dahlgren (1993), *The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM*, NIST order number PB91-100354.

Goldman, J-P., A. Auchlin, S. Roekhaut, A. C. Simon & M. Avanzi (2010). Prominence perception and accent detection in French. A corpus-based account. In *Proceedings of Speech Prosody 2010*, Chicago, 100575: 1-4.

Goldman, J.-P., M. Avanzi, A. Auchlin & A. C. Simon (2012). A continuous prominence score based on acoustic features. In *Proceedings of Interspeech 2012*, Portland, SS09.09.

Goudbeek, M., R. Smits, D. Swingley & A. Cutler (2005). Acquiring auditory and phonetic categories. In H. Cohen & C. Lefebvre (eds), *Handbook of categorization in cognitive science*, 497-514. Amsterdam: Elsevier.

Grover, C., B. Heuft & B. van Coile (1997). The reliability of labeling word prominence and prosodic boundary strength. In *Proceedings of the ESCA Workshop on Intonation*, Athens, 165-168.

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics* 31. 39-62.

Heuft, B., T. Portele, P. Wagner, C. Widera & M. Wolters (2000). Perceptual prominence. In W. Sendlmeier (ed.), *Speech and signals. Aspects of speech synthesis and automatic speech recognition*, 97-116, Frankfurt am Main: Hektor.

Holt, L. L. & A. J. Lotto (2010). Speech perception as categorization. *Attention, Perception & Psychophysics* 72(5). 1218-1227.

Jenkin, K. L. & M. S. Scordilis (1996). Development and comparison of three syllable stress classifiers. In *Proceedings of ICSLP '96*, Philadelphia, 733-736.

Jensen, C. (2004). *Stress and accent*. Copenhagen: University of Copenhagen Ph.D. dissertation.

Jensen, C. & J. Tøndering (2005). Choosing a scale for measuring perceived prominence. In *Proceedings of Interspeech 2005*, Lisbon, 2385-2388.

Jun, S. (2005). Prosodic typology. In S. Jun (ed.), *Prosodic typology: The phonology of intonation and phrasing*, 430-458. Oxford: Oxford University Press.

Kohler, K. (2003). Neglected categories in the modelling of prosody – pitch timing and non-pitch accents. In *Proceedings of ICPhS 2003*, Barcelona, 2925-2928.

Kohler, K. (2005). Form and function of non-pitch accents. In *Prosodic Patterns of German Spontaneous Speech* 35a, AIPUK, 97-123.

Kohler, K. (2006). What is emphasis and how is it coded? In *Proceedings of Speech Prosody 2006*, Dresden, 748-751.

Kopecek, I. (1999). Speech recognition and syllable segments. In *Proceedings of Workshop on Text, Speech and Dialogue – TSD'99*, LNAI 1692, 203-208.

Ladd, D. R., J. Verhoeven & K. Jacobs (1994). Influence of adjacent pitch accents on each other perceived prominence: Two contradictory effects. *Journal of Phonetics* 22. 87-99.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Li, K., S. Zhang, M. Li, W.-K. Lo & H. Meng (2011). Prominence model for prosodic features in automatic lexical stress and pitch accent detection. In *Proceedings of Interspeech 2011*, Florence, 2009-2013.

Low, E. L., E. Grabe & F. Nolan (2001). Quantitative characterisation of speech rhythm: Syllable-timing in Singapore English. *Language and Speech* 43. 377-401.

Ludusan, B., A. Origlia & F. Cutugno (2011). On the use of the rhythmogram for automatic syllabic prominence detection. In *Proceedings of Interspeech 2011*, Florence, 2413-2417.

Mehrabani, M., T. Mishra & A. Conkie (2013). Unsupervised prominence prediction for speech synthesis. In *Proceedings of Interspeech 2013*, Lyon, 1559-1563.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America* 58. 880-883.

Mertens, P. (1991). Local prominence of acoustic and psychoacoustic functions and perceived stress in French. In *Proceedings of ICPhS'91*, Aix-en-Provence, 218-221.

Noetzel, A. (1991). Robust syllable segmentation of continuous speech using neural networks. In *Proceedings of IEEE Electro International Conference Record*, New York, 580-585.

Origlia, A., G. Abete & F. Cutugno (2013). A dynamic tonal perception model for optimal pitch stylization. *Computer Speech & Language* 27. 190-208.

Origlia, A., F. Cutugno & V. Galatà (2014). Continuous emotion recognition with phonetic syllables. *Speech Communication* 57. 155-159.

Pamies Bertran, A. (1999). Prosodic typology: On the Dichotomy between stress-timed and syllable-timed languages. *Language Design* 2. 103-130.

Pfitzinger, H., S. Burger & S. Hid (1996). Syllable detection in read and spontaneous speech. In *Proceedings of ICSLP'96*, Philadelphia, 1261-1264.

Ramus, F., M. Nespor & J. Mehler (1999). Correlates of linguistic rhythm in the speech signal. *Cognition* 73. 265-292.

Russo, M. (ed.) (2010), Prosodic universals: Comparative studies in rhythmic modeling and rhythm typology. Roma: Aracne.

Silipo, R. & S. Greenberg (1999). Automatic transcription of prosodic stress for spontaneous English discourse. In *Proceedings of ICPhS '99*, San Francisco, 2351-2354.

Sluijter, A. & V. van Heuven (1996). Acoustic correlates of linguistic stress and accent in Dutch and American English. *In Proceedings of ICSLP '96*, Philadephia, 630-633.

Sluijter, A., V. van Heuven & J. Pacilly (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America* 101. 503-513.

Spencer, A. (1996). *Phonology*. Oxford: Blackwell.

Streefkerk, B. (1996). Prominent accent and pitch movements. *Inst. of Phon. Sciences Proceedings, University of Amsterdam* 20. 111-119.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Paliwal (eds.), *Speech coding and synthesis*, 495-518. New York: Elsevier.

Taylor, P. A. (1992). *A phonetic model of English intonation*. Edimburgh: University of Edimburgh Ph.D. dissertation.

Taylor, P. A. (2000). Analysis and synthesis of intonation using the Tilt Model. *Journal of the Acoustical Society of America* 107. 1697-1714.

Tamburini, F. (2003). Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system. In *Proceedings of Eurospeech 2003*, Geneva, 129-132.

Tamburini, F. (2005a). *Fenomeni prosodici e prominenza: un approccio acustico*, Bologna: Bononia University Press.

Tamburini, F. (2005b). Automatic prominence identification and prosodic typology. In *Proceedings of InterSpeech 2005*, Lisbon, 1813-1816.

Tamburini, F. (2006). Reliable prominence identification in English spontaneous speech. In *Proceedings of Speech Prosody 2006*, Dresden, PS1-9-19.

Tamburini, F. (2009), Prominenza frasale e tipologia prosodica: un approccio acustico. In G. Ferrari (ed.), *Linguistica e modelli tecnologici di ricerca. Atti del XL Congresso internazionale di studi della Società di Linguistica Italiana*, *Vercelli, 21-23 settembre 2006*, 437-455. Roma: Bulzoni.

Tamburini, F. (2013). Una valutazione oggettiva dei metodi più diffusi per l'estrazione automatica della frequenza fondamentale. In *Atti dell IX Convegno Nazionale dell'Associazione Italiana di Scienze della Voce (AISV2013)* 427-434. Roma: Bulzoni.

Tamburini, F. & C. Caini (2005). An automatic system for detecting prosodic prominence in American English continuous speech. *International Journal of Speech Technology* 8. 33-44.

Tamburini, F. & P. Wagner (2007). On automatic prominence detection for German. In *Proceedings of InterSpeech 2007*, Antwerp, 1809-1812.

Tamburini, F., P. M. Bertinetto & C. Bertini (2014). Prosodic prominence detection in Italian continuous speech using probabilistic graphical models. *In Proceedings Speech Prosody 2014,* Dublin, 285-289.

Taylor, P. A. (2000). Analysis and synthesis of intonation using the Tilt Model. *Journal of the Acoustical Society of America* 107. 1697-1714.

Terken, J. (1991). Fundamental frequency and perceived prominence. *Journal of the Acoustical Society of America* 89. 1768-1776.

Turk, A. E. & J. R. Sawusch (1996). The processing of duration and intensity cues to prominence. *Journal of the Acoustical Society of America* 99. 3782-3790.

Vainio, M., A. Suni & D. Aalto (2013). Continuous wavelet transform for analysis of speech prosody. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody conference (TRASP'13)*, Aix-en-Provence, France, 78-81.

van Kuijk, D. & L. Boves (1999). Acoustic characteristic of lexical stress in continuous telephone speech. *Speech Communication* 27. 95-111.

Wagner, P. (2002), *Vorhersage und Wahrnehmung deutscher Betonungsmuster*. Bonn: University of Bonn Ph.D. dissertation.

Wagner, P. (2005). Great expectations – Introspective vs. perceptual prominence ratings and their acoustic correlates. In *Proceedings of Interspeech 2005*, Lisbon, 2381-2384.

Wagner, P., F. Tamburini & A. Windmann (2012). Objective, subjective and linguistic roads to perceptual prominence. How are they compared and why? In *Proceedings of InterSpeech 2012*, Portland, SS09.02.

Warner, N. & T. Arai (2001). Japanese mora-timing: A review. *Phonetica* 58. 1-25.

Waterson, N. (1987). *Prosodic phonology: The theory and its application to language acquisition and speech processing*. Great Britain: Grevatt and Grevatt.

Wightman, C. W. & M. Ostendorf (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing* 2. 469-481.

Windmann, A., P. Wagner, F. Tamburini, D. Arnold & C. Oertel (2010). Automatic prominence annotation for German speech syntesis corpus: Towards prominence-based prosody generator for unit selection syntesis. In *Proceedings of 7th ISCA Speech Syntesis Workshop*, Kyoto, 377-382.

Wu, S., M. L. Shire, S. Greenberg & N. Morgan (1997). Integrating syllable boundary information into speech recognition. In *Proceedings of ICASSP'97*, Munich, 987-990.

Zahorian, S. A. & H. Hu (2008). A spectral/temporal method for robust fundamental frequency tracking. *Journal of the Acoustical Society of America* 123. 4559-4571.

Fabio Tamburini
Department of Classic Philology and Italian Studies
Alma Mater Studiorum – University of Bologna
Via Zamboni 32, 40126 Bologna
Italy
e-mail: fabio.tamburini@unibo.it