DOI: 10.17469/O2104AISV000014

VALENTINA SCHETTINO, ANTONIO ORIGLIA, FRANCESCO CUTUGNO

Dynamic time warping and prosodic prominence

In this study, an investigation on methodological issues linked with prosodic prominence rating is carried out. Our main research goal is the resolution of a particular issue related to a specific rating scale – i.e. the one resulting from the PromDrum method (Samlowski, Wagner, 2016). In this approach, namely, the rater is left free to *drum* as many syllabic units as perceived, thus resulting in rated material in which the number of rated units possibly does not correspond to the number of actually expected syllables. In order to solve this problem, a forced alignment algorithm is developed with the Dynamic Time Warping procedure. In this way, we are able to find the best possible alignment without subjective choices. Moreover, the procedure allows a qualitative evaluation of the rated material.

Keywords: prosodic prominence, rating scales, Dynamic Time Warping.

1. *Prosodic Prominence: An introduction*

In the last years, many scholars have investigated a prosodic phenomenon with an important communicative function (Streefkerk, 2002), known as prosodic prominence. Prominence could be defined as "the property by which linguistic units are perceived as standing out from their environment" (Terken, 1991: 1768). In spoken productions, salient units are produced and perceived in a more detailed way, being fundamental in the understanding of the linguistic message. The way in which this emphasis is achieved, however, is a complex and multi-layered process, that comprehends both acoustic behaviours and linguistic expectancies (Streefkerk, 2002). The disentanglement of the different influences and of their relative importance is particularly complex. Furthermore, every language seems to adopt its own manner of signaling prominence, both through different acoustic correlates and through the linguistic meaning assigned to each; prominence patterns can scarcely be compared in different linguistic and phonetic settings in the same language too: this means that the same prominence pattern could bear two different meanings in two different contexts. Eventually, prominence can be referred in various ways to different prosodic domains: intonation patterns are strictly related to prominence, seen that the shape of the pitch curve is related to both phenomena; stress and prominence mutually influence each other, too. As such, then, prominence is a complex phenomenon, acting on different prosodic levels; its production and perception, indeed, cannot be properly described without observing the various contribution to it from different, although related, fields. Disentangling the different features contributing to this phenomenon on different levels is a big challenge that the academic community is facing (Wagner, Origlia, Avesani, Christodoulides, Cutugno,

D'Imperio, Escudero Mancebo, Gili Fivela, Lacheret, Ludusan, Moniz, Chasaide, Niebuhr, Rousier-Vercruyssen, Simon, Simko, Tesser, Vainio & 2015). One of the most interesting issues related to this topic, however, is the annotation of prosodic corpora regarding prominence. Rating procedures, indeed, are both interesting for cognitive reasons and for methodological interest: how comes that some syllables are perceived as standing out, and which elements are considered most relevant in the perceptual phase? In order to shed some light on these questions, in this paper we deal with a specific rating methodology, explaining its theoretical benefits but also presenting some drawback aspects linked with it. Specifically, we suggest an improvement of the rating scale known as PromDrum method (Samlowski, Wagner, 2016), using Dynamic Time Warping.

In the next sections, the problem will be presented, and our proposal will be depicted. We will present the material used for the investigation and display the results achieved through the suggested implementation.

2. *Prosodic Prominence and rating scales*

One of the first issues to be solved when approaching the annotation of prosodic prominence is the kind of rating scale to use for prominence levels, linked with the more general question of whether prominence should be regarded as a discrete or continuous phenomenon. In the literature, in fact, different kinds of scales have been used: given that the linguistic community still does not agree upon the number of relevant linguistic categories as regards prominence scales, some settle a binary system, in which the distinction between prominent and non-prominent syllables is considered sufficient for the description of the phenomenon; on the other side, however, we find completely opposed approaches, in which prominence categories are discretized in 31 different classes.

In particular, the so-called 31 degrees scale, first introduced by Fant, Kruckenberg (1989), aims at a fine-grained evaluation of perceived prominence levels; this type of scaling can be easily related to a physical perspective and presupposes a gradient functioning of prosodic patterns, both on the level of production and perception. However, it could be difficult for the annotators to discretize between such close categories.

In the other scale type mentioned above, known as binary scale, syllables are expected to be either prominent or non-prominent, no further category is contemplated (Wightman 1993; Streefkerk, Pols, Ten Bosch, 1999); this approach should result in an easier job for the annotator: nevertheless, this simplistic view risks to leave aside important distinctions and consequent interpretation of intonational patterns. Indeed, different degrees of prominence, related to different linguistic levels, are entirely lost in this kind of approach.

Another scale type is a compromise between the other two and is in fact known as intermediate scale: it uses four different levels of perceived prominence and usually distinguishes between non-prominent, almost non-prominent, almost promi-

nent and prominent syllables (Jensen, 2004). However, results are still not satisfying.

Lately, a new methodology has been developed (cf. Samlowski, Wagner, 2016), which exploits the prosody-gesture link and tries to avoid all drawbacks of the other approaches. In this case, prominence perception is related to the beating movement: participants in the experiment are asked to listen to some short sentences and reiterate them by beating on a DrumPad, modulating the intensity of the beat in a directly proportional way. The drumming task permits an easy, intuitive processing of prosodic prominence, allowing drummers to produce a fine-grained annotation without necessarily being experts: discretizing between close categories results in an intuitive task in this case, because drummers are not confronted with the choice between very similar, close categories, but rather reproduce what they hear exploiting the prosody-gesture link; moreover, this procedure proved to be very fast, thus enabling the annotation of very large corpora in a consistent way as regards prosodic prominence. Moreover, perception of prominence patterns remains central all over the annotation task.

All things considered, the PromDrum method is in our opinion the best suited approach for the investigation of prosodic prominence perceptual patterns, and for different reasons: firstly, it does not require a long preparation phase and it allows naïve speakers to take part in the perceptual experiment of rating spoken productions; secondly, it is fast and enables the annotation of very large corpora; lastly, but most importantly, native speakers (and listeners) should be able to evaluate the degree of prominence of the data reflecting the actual functioning of perceptual processes related with prosodic prominence, allowing an examination of the complex dynamics developed along sequences of prominence peaks in relation with non-prominent, contextual units.

Summing up, this methodology seems suitable for many purposes: from our point of view, the major advantage of this approach is that it refers to a definition of prominence that is strongly intuitive and based on acoustics, but that at the same time does not leave aside the mental categorization of the phenomenon: it succeeds, then, in mixing both signal-based and expectation-based factors. For these reasons, we will use this methodology for annotating a corpus of spontaneous speech. However, we will also describe the limitations still inherent in this approach and propose an improvement on that side.

3. *PromDrum method: An issue*

As we have stated, prosodic prominence is a quite complex phenomenon with different elements interplaying at the same time on different levels. As such, a good research angle is on the perceptual side, as prominence is mainly referred to as a perceptual process, with units "*perceived as standing out from their environment*". In analyzing this phenomenon, then, an important methodological choice regards the rating scale used for examination. In the preceding section, we have presented

the most widespread ones and explained why we chose the so-called PromDrum method (Samlowski, Wagner, 2016). In this part of the paper, however, we intend to depict an issue inherent in this approach, in order to propose an improvement on this side towards the end of this contribution.

Indeed, the main problem with this approach is that the drumming procedure – with which annotators rate prominence exploiting the prosody-gesture link through an electric drum pad – leaves raters free to decide how many beats they hear. Traditionally, manually annotated corpora for prosodic prominence have used a strict correspondence of syllables and annotation units: annotators, indeed, were forced to rate the relative salience of each syllable as marked by the researchers (cf. Fant, Kruckenberg, 1989; Wightman, 1993; Streefkerk et al., 1999; Jensen, 2004). The drumming procedure has the advantage of removing this constraint: in fact, drummers are left free to produce the number of beats they consider to be the best representation of what they have heard in the input audio file. This choice, however, implies that, in our data, the drumming associated with a specific file may not contain an amount of beats that is equal to the number of reference syllables[1]. In Samlowski, Wagner (2016), authors examined only drummed sentences in which the number of expected syllables and the number of beats coincide, as their aim consists mainly in the validation of the drumming procedure. In our case, on the contrary, we are interested in exploiting the freedom left to the annotators, because in this way we believe we can further understand the connection between perceptual processes and rating procedures. Nevertheless, we still have to overcome the alignment problem between rating and drumming: in fact, once we have our input audio file and the concerning annotation, we have to be reasonably sure that a given drummed beat can be correctly assigned and aligned with a given spoken chunk. In order to display our proposal to overcome this issue, we have to present the material used for the investigation first. In the next section, we will proceed with that; after that, we will introduce our proposal.

4. *Material*

Prosodic prominence (Terken, 1991) has gained attention over the past decades. Nevertheless, the different elements related to this topic still have to be examined in a detailed way, disentangling notational problems and domain mix-up (Wagner et al., 2015). Due to these investigation issues, large corpora annotated on the prosodic level for prominence are very small in number, especially when dealing with multilingual data and L2 acquisition – with some exceptions[2]. In addition to the lack of comparability of databases, it is also hard to compare annotation methods:

[1] With the concept *reference syllables*, we refer to the syllabic units that can be expected in an utterance on the basis of lexical and/or phonematic constraints.

[2] Cf. Kohler, (1996); Campione, Véronis, (1998); Ostendorf, Price & Shattuck-Hufnagel, (1995); Oostdijk, (2000); Cheng, Greaves & Warren., (2005); Hirst, Bigi, Cho, Ding, Herment & Wang, (2013), among others.

indeed, a variety of methodological approaches as regards annotation schemata and rating scales has resulted in a low level of comparability between different works and a core notational problem (Wagner et al., 2015). Attempting an annotation of prominence can be both a difficult challenge and, at the same time, an important occasion for examining this complicated situation; in order to gain further insights on this investigation, we decided to annotate prominence from a perceptual point of view (Portele, Heuft, 1995; 't Hart, Collier & Cohen, 1990). Moreover, we decide to concentrate this study on the examination of German L1 and Italian L2, mainly because of personal competencies. However, we also find stimulating the idea of investigating perceptual processes in L2 productions: indeed, if the DTW algorithm can be applied to L2 material, too, the implementation scope of this technique would be *a fortiori* wide.

To our knowledge, at present there is only one available corpus structured for prosodic studies taking into consideration German L1 and Italian L2 together (Schettino, 2015). This database consists of 24 German native speakers producing more than nine hours of spoken speech, both in German L1 and Italian L2. Most of the speakers were university or school students, with just one teacher of Italian. Mean age was 25.6 years, with a total amount of nine men and 15 women; different levels of fluency in Italian were examined, and specifically 14 speakers of the level A, eight of the level B and two of the level C (CEFR); the diatopic variation was not taken into consideration. In the following tables, precise quantitative information is reported.

Table 1 - *Quantitative information about the corpus*

	Read speech	Commentaries	Dialogue (German L1)	Dialogue (Italian L2)
Speakers	24	9	24	24
Recording time	1h 36m 40s	50m 04s	2h 52m 18s	3h 49m 18s

Table 2 - *Additional information about the informants' fluency level in Italian L2 (CEFR)*

	A1	A2	B1	B2	C1	C2
Male	4	0	2	2	1	0
Female	7	3	1	3	0	1
Total	11	3	3	5	1	1

The elicited spontaneous productions consisted of two participants (for each session) who were asked to play TicTacToe together. The material is thus characterized for similar syntactic organization, comparable lexicon and congruous duration across files; moreover, this game is a perfect situation for analyzing prominence distribution predictability (Watson, Arnold & Tanenhaus, 2008). Participants played alternatively in German and in Italian, with a randomized sequence of languages. The starting move of the game and the first speaker were randomized, too. At the

beginning of every game, the participants were instructed on the starting move and the language they should have used for that particular game. In total, every pair of participants played eight games in Italian and eight in German, for a total amount of 16 games per couple. All games were recorded in a single session. The files were registered with high quality microphones in an anechoic chamber.

As regards the segmentation procedure, the above described "dialogues" were then segmented in speech turns. For this study, we used a sample of this segmented dialogic turns, both in German and in Italian; in particular, we used nine Italian drummers, each of them evaluating the degree of perceived prominence of 61 different turns, and three German drummers, drumming 51 turns each. Turns were selected trying to locate the files in which intonation and stress patterns differed from the norm: if a stress was put on the "wrong" syllable, or the pitch shape diverged from usual Italian[3] patterns and/or alignment, the file was considered to be a good element of investigation. In total, then, we let twelve annotators drum prominence, listening to about 700 files. In this way, we obtained 700 drummed files in which information about prominence perception and functioning is concealed. Still, we have to overcome the alignment problem with the original audio file. In the next section, we will advance our proposal for solving this issue.

5. *PromDrum and alignment: Our proposal*

As previously mentioned, prosodic prominence has been described as a complex phenomenon, in which bottom-up features and top-down knowledge are intertwined in the perception phase, resulting in a complex dynamic whose different elements and their mutual influences are difficult to be told apart. In this respect, we believe we can further understand the nature of this phenomenon through the examination of the connection between perceptual processes and rating procedures. For this reason, it is important to overcome the alignment problem between ratings and drumming. In order to do that, we develop an objective procedure that is able to both evaluate the quality of the annotations in the first place and to find the optimal alignment of drummed beats and expected syllabic units. We choose to use the Dynamic Time Warping (henceforth DTW, Sakoe, Chiba, 1978), because this algorithm does not assume that the number of units in the sequences to align has to be the same; furthermore, it reports alignment paths that minimize a given distance function. As such, then, it is compatible with both our aims of evaluating the quality of the annotations and to find the best suited alignment with respect to the reference number of syllables.

Concerning the qualitative evaluation of the annotated files, we have to bear in mind that – given the degree of freedom left to the annotators – it is possible that, in some cases, the number of beats does not exactly match the amount of ref-

[3] We do not expect German productions to be mis-produced, as speakers in this corpus are German native speakers.

erence syllables. Little discrepancies between drummed beats and syllabic units are acceptable for our analysis: we do not impose a "right" number of units that have to be recognized, but rather leave freedom of perception to the drummer. In our opinion, indeed, these little differences could be a big help in our investigation, as we regard them as possible expressions of perceptual processes, with which it would be possible to interpret the relationship between signal and perception in a more extensive way. Moreover, differences in the amount of reference syllables and beats do not represent a problem in our approach, because the DTW procedure is able to align beats and signal in a straightforward way. For example, if a drummer drums 10 beats instead of 11, the algorithm would be able to calculate to which units the beats are probably referred, and it is possible to retrace the non-drummed syllable. At this point, it becomes possible to add linguistic interpretation to the missing beat, trying to understand why that particular syllable was not perceived in the rating phase. On the contrary, drumming sequences that greatly differ from the reference ones cannot be used for further analyses and must be discarded: if – for instance – a drummer should have drummed 17 beats, but only 5 beats are found in the file, the alignment cannot be objectively reconstructed.

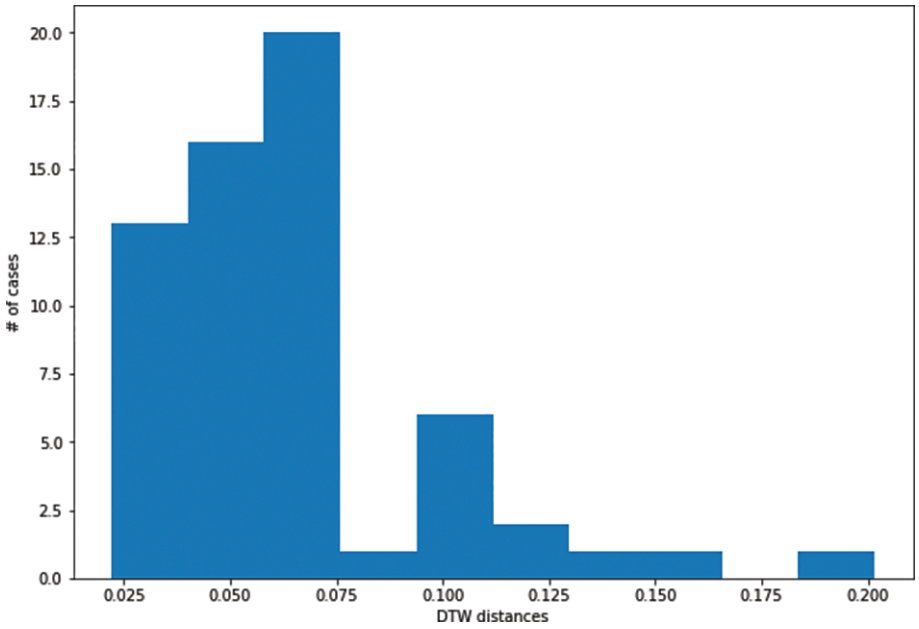
For all these reasons, we reckon that the DTW procedure applied to the PromDrum rating method can successfully improve this technique and help us improving our understanding of the phenomenon known as prosodic prominence. In the next section, specific results corroborating this assumption will be depicted and discussed.

6. *PromDrum method and DTW implementation: Our results*

In this section, results about the application of the DTW algorithm to the PromDrum rating method will be presented.

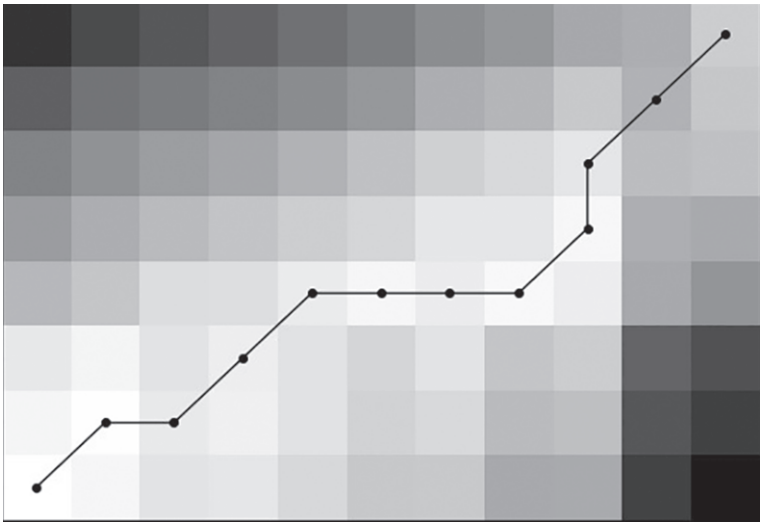
In the first place, the minimum distance provided by the DTW procedure – interpreted as the effort that the algorithm has to do in order to connect beats and reference syllables – gives us indications about the annotation quality: the less effort the system makes, the lower the DTW distances values are, and the surer we can be about the quality of the drummed sentences. In Figure 1 it is possible to observe the minimum distance value plotted along the number of cases: in this specific case, it seems that most of the sentences drummed by this drummer have a minimum distance ≤ 0.075. With this procedure, we can calculate the qualitative threshold for each drummer, leaving aside all the drummed files that diverge too much from the reference, thus being sure that the actually evaluated files have been produced with a good degree of correlation between perceived units and linguistic signal.

Figure 1 - *Minimum distance value plotted along the number of cases. Qualitative evaluation of the drummer*



Along the whole set of our drummers, we calculated that the best threshold is 0.07 in our data: most of the annotators, in fact, have produced the vast majority of acceptable drummed files under this DTW distance value, indicating that it could be considered a good qualitative limit, sufficiently strict but quite fair, too.

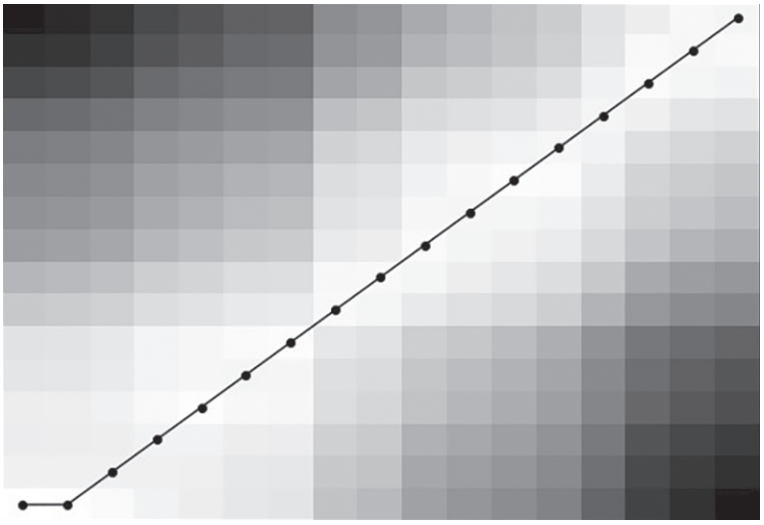
Figure 2 - *Warping path of a non-acceptable drummed file. Too big discrepancies between the reference syllables distribution and the beats one*



The second advantage of the DTW approach is that this procedure indicates what is the best alignment between the two sequences: this is named warping path. In this way, the alignment does not result from a subjective choice: it is the result of an optimization procedure that minimizes the chosen distance function.

In Figures 2 and 3, two different accounts of the warping path are reproduced: in the first case, there are consistent discrepancies between the rating annotations and the number of reference beats: as can be seen, the warping path does not find a one-to-one correspondence and the line appears to be consequently not straight.

Figure 3 - *Warping path of an acceptable drummed file. Tolerable discrepancies between the reference syllables distribution and the beats one*



In the second figure, on the contrary, the relationship between number of reference beats and actual beats is much more uniform: the only drumming hit that prevents a bijective correspondence is the first one. In this specific case, the Italian word *io* "I" – that is expected to contain two syllabic units – is drummed as a single beat: however, for the DTW algorithm it is not much onerous to recognize that two adjacent, very close syllables may be perceived as one unit and drummed accordingly; this is shown in the Figure, too: given that the colour of the cells correlates with the alignment cost – with white signalling a relative smooth and effortless alignment and black cueing almost impossible connections, we can observe that the dots relative to the first two reference syllables, although connected with one single beat, are positioned in white cells, indicating the relative ease with which they are related to the same drumming hit.

As regards the forced alignment of the beats' values and the acoustic cues, we decide to relate drummed beats with the vocalic portion of the syllable. The vowel, indeed, is the portion of speech that carries most of the relevant acoustic features; furthermore, it was proven in many studies that the quality and the distribution of

vocalic phones in an utterance is the only acoustic correlate of rhythmic categorization that seems to be valuable in comparative works (cf. Dauer, 1987; Mehler, Dupoux, Nazzi & Dehaene-Lambertz, 1996; Ramus, Nespor & Mehler, 1999; Ling, Grabe & Nolan, 2000; Grabe, Low, 2002). As the PromDrum method mainly reflects prominence perception on the rhythmic level, i.e. provides a reflection of the rhythmic perception of sequences of "strong" and "weak" units – it seems appropriate to base our examination on vocalic productions.

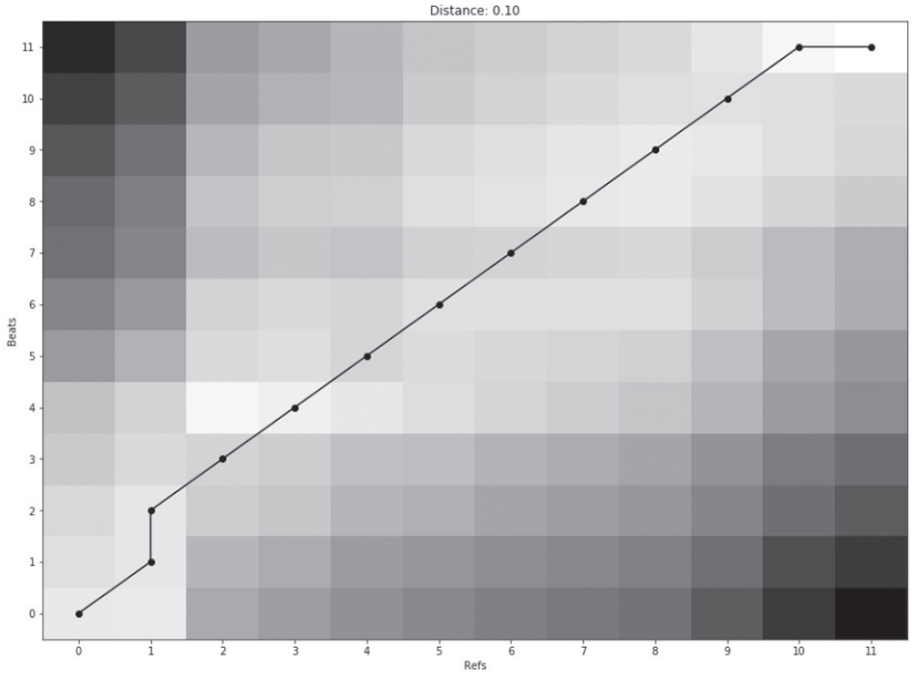
Concerning the temporal alignment of the two units (beat and vowel), we set the onset of the vowel as the point in time to which the beats are connected. The temporal alignment, indeed, may cause problems in our approach: as we mentioned, the DTW procedure reports alignment paths that minimize a given distance function, in our case, the Euclidean distance. In case we have less beaten units than vowels, then, the temporal distance between the beats will be counted as a relevant indication in the DTW algorithm in the alignment process. As a consequence, the temporal alignment of the beat with the vowel is crucial in the calculation of distance and in the following assignment of beats to vowels. However, the alignment can be carried out in a successful way if we make an important consideration: as a beat is realized faster than a vocalic breath emission, aligning the beat with – for example – the intensity peak of the vowel introduces delays due to the different speed with which the intensity curve can reach its peak. Putting the reference on the vowel onset, on the contrary, marks well the moment in which the vowel is perceived. The best way to align beaten units and vowels, then, is using the vowel onset as a fixed reference in the signal.

The obtained successive values relative to the beats sequences are normalized in the time domain: we assume that the first beat and the first vocalic onset are located at time 0, while the last units will be displaced at time 1. In this way, each file has a temporal sequence that can be aligned without too much effort in a considerably successful way. Pauses also play a role, helping in the right alignment process and disambiguating between different alignment paths.

The described procedure allows the optimal alignment without forcing us to make too strong assumptions: unlike the approach found in Samlowski, Wagner (2016), where – in order to validate the methodology – only the drummed file with an equal amount of beats and reference syllables are counted, we do not need to set a number of linguistic units that should be regarded as the right amount. The DTW algorithm, indeed, is able to prevent strict constraints in the empirical phase, avoiding the obligation of enforcing some methodological protocols due to underlying theoretical assumptions. In our case, moreover, the simple fact that reference syllables and beaten units are equal in number does not suffice in assuring a good quality of the alignments: it could well be possible, indeed, that a drummer – in the attempt of reiterating the "correct" number of syllables, concentrates in repeating the exact amount of expected units, without being successful in reproducing the rhythmic contour of the input file. In this case, the Euclidean distances between the beats would not reflect the temporal distribution of vowels in the audio file,

thus resulting in a higher degree of effort in the DTW algorithm and consequently in a badly correlated – perhaps not acceptable rated file. An example of this sort can be observed in Figure 4: although the two sequences have the same number of units in it, the alignment is not straightforward: most of the cells are grey, with the algorithm interpreting the Euclidean distances between the different units as a bad rhythmic reflection of the input file. As a consequence, this drummed sequence is not accepted, because the degree of effort in the alignment procedure exceeds the threshold of acceptability for the given drummer.

Figure 4 - *Warping path of a non-acceptable drummed file in which the number of reference syllables and does number of beats coincide, but their distribution over time do not*



This is not a drawback in our opinion: in this way, in fact, we can be sure that only the drummed files that really reflect rhythmic perceptual processes are accepted.

On the other side, we can find drummed sequences that contain a number of beats that is not at all equivalent to the reference number of syllables, but that still can be aligned in a successful way through the DTW procedure. In Figure 5, an example of this type of drumming is shown: in this case, although we have a number of beats lower than the expected reference syllables, the algorithm is capable of aligning the two sequences with not too much effort: as we can see, the points are mostly distributed in white cells. Evidently, even if the produced drummed beats are less than expected, their sequence succeeds in reflecting the rhythmic contour of the input audio file.

Figure 5 - *Warping path of an acceptable drummed file in which the number of reference syllables and the number of beats do not coincide, but their distribution over time does*

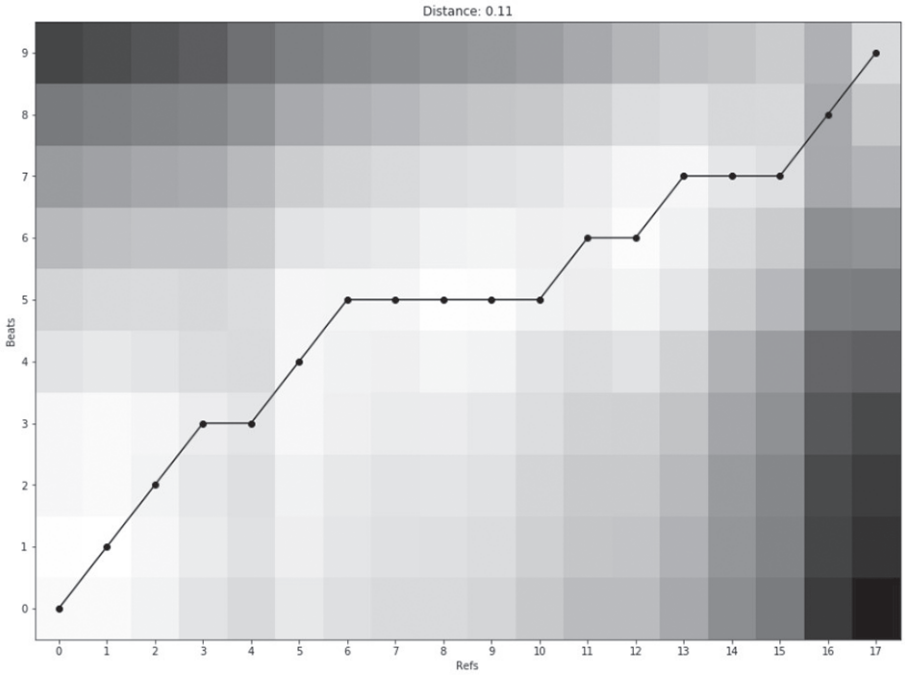
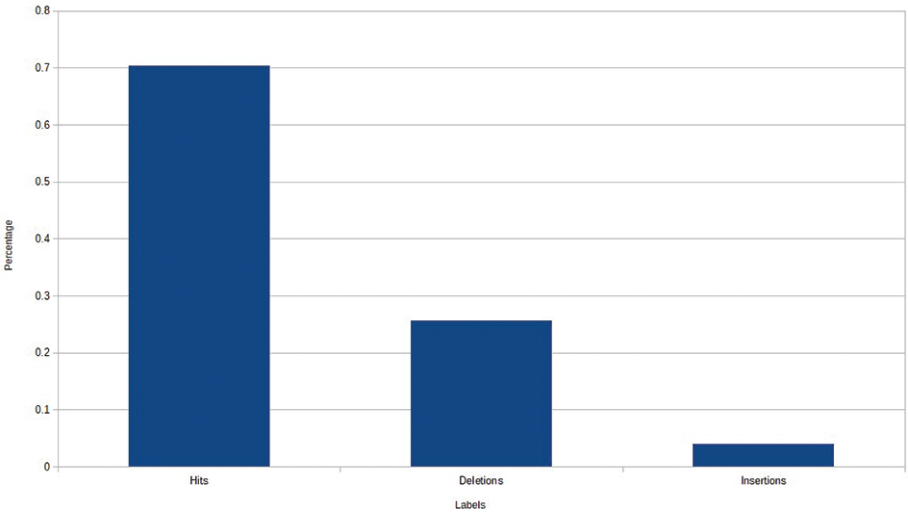


Figure 6 - *% of Hits, Deletions and Insertions in the DTW procedure*

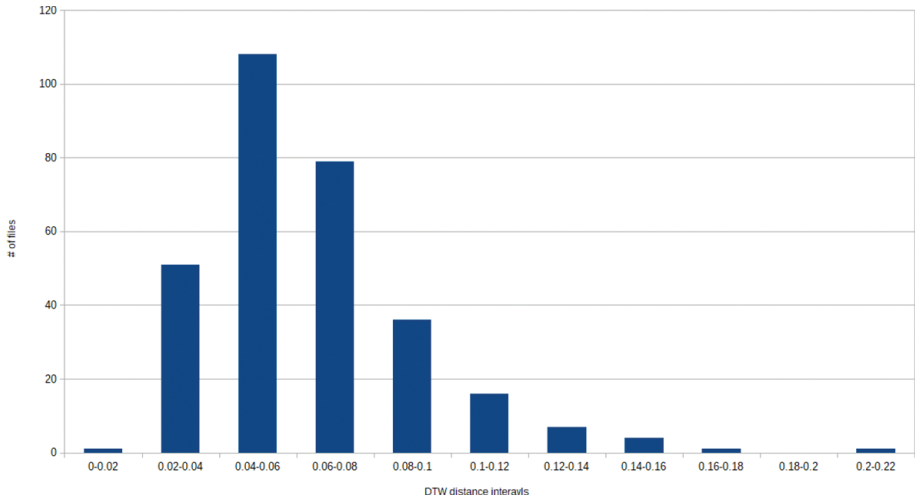


As it can be observed in the last two figures, the methodology developed in this work allows both the insertion of "non-expected" beats and the deletion of "expected" one, with respect to the reference number of syllabic units. In general, we are

expecting a great amount of beats that actually match the reference unit without alignment problems; we are defining these cases as Hits. Deletions – i.e. absences of an expected beat – are also expectable, given the methodological organization of our approach and in particular the freedom left to the drummer to rate the amount of syllables they hear; however, their influence on the analysis is not so big, because the DTW provides a qualitative level of the alignments: as a result, missing beats do not automatically result in a file ineligible for analysis. Insertions, on the contrary, are not expected to take place often: it is not probable that drummers produce more beats than what they listened to, even if it is possible – as shown in Figure 4 – that the warping path assigns two different beats to the same reference syllable.

In Figure 6, the percentage of Hits, Deletions and Insertions for the whole examined corpus is reported. Our expectations are confirmed: Hits represent the 70% of the cases, a value that further legitimates the PromDrum procedure enhanced with our methodological improvements; Deletions instances are about a quarter of the cases, whereas Insertions are really rare (less than 5% of the whole investigated data). As it can be seen, although Hits represents a vast majority of the cases, the number of files in which this correspondence does not take place is statistically significant. As a consequence, developing a procedure that allows the use of most drummed sentences is a worthwhile effort. Therefore, the DTW technique can be considered a fruitful improvement in order to obtain the best possible alignment between the beats sequence and the relative audio file.

Figure 7 - *Relationship between the number of examined files and DTW distance intervals*



A last acknowledgement about the value of this approach relies in the number of acceptable drummed sequences: if with the basic PromDrum method only those annotations could be accepted in which the number of reference syllables and the number of actual beats coincided, with the DTW technique it is possible to accept most of the annotations, disregarding only those qualitatively scarce because not

reflecting the rhythmic nature of the input audio file. As it can be seen in Figure 7, in fact, a histogram about the relationship between the number of examined files and the DTW distance intervals tells us that we are on the right track: indeed, most of the files can be found under the 0.1 distance threshold. A limited amount of files display a distance value higher than 0.12: for those ones, the effort put by the algorithm in finding the best possible alignment is too high, and they must be discarded. But the whole rest can be accepted and used for linguistic analyses and interpretation.

7. *Conclusion and future work*

In Samlowski, Wagner (2016), the prosodic prominence rating procedure admitted the possibility of annotated files having non-equivalent numbers of beats and expected syllables, but solved the issue just discarding the considered files. In this work, we develop a procedure for considering those files, too, as good candidates for perceptual analyses of prosodic prominence. The DTW algorithm suits our aim, and for different reasons: firstly, the best alignment between the two sequences, the so-called warping path, does not result from a subjective choice, but is the manifestation of objective parameters. The second advantage is that it allows us to evaluate the quality of the annotations: the minimum distance provided by the DTW algorithm – interpreted as the effort that the algorithm has to do in order to connect beats and reference syllables – gives indications in this sense. Furthermore, we can calculate the qualitative threshold for each rater, leaving aside all the drummed files that diverge too much from the reference. Little discrepancies between the amount of drummed beats and syllabic units are acceptable for our analysis; indeed, we regard them as possible expressions of perceptual processes. On the contrary, drumming sequences that greatly differ from the reference ones cannot be used for further analyses and must be discarded.

We consider this procedure as a big improvement in the drumming methodology, and in general in the rating of prosodic prominence. With this approach, indeed, it is possible to rate prosodic prominence in a simple and intuitive way, exploiting the existing link between prosody and gestures; in this way, large corpora of spoken speech can be annotated in a rather fast way. Moreover, the great enhancement connected with the DTW procedure consists in the acquittal from background assumptions: we do not need to state how many units have to be rated, nor do we need to set them manually. In this way, we reckon that the perceptual process can be investigated in a more straightforward way, together with the relationship between perceived degree of prominence and acoustic correlates. This concrete understanding can be used in a future phonological description of the phenomenon.

As regards possible improvements of this approach and particular precaution to take when using this methodology, we observed that the best quality of the analysis comes from drummed data registered with a set sound on the DrumPad whose mean pitch is attested between 150 and 250 Hz. Moreover, input audio files with

long pauses in it should be regarded as bad candidate for the examination, because long silences and their relative distance to adjacent vowels are improbable to be reproduced in a satisfying way, and the pertaining files are systematically discarded during the DTW procedure.

Summing up, we think that this procedure can ameliorate the DrumPad method, which in turn could result in a better annotation system, applicable to large corpora of spontaneous speech.

Bibliography

Campione, R., Véronis, J. (1998). A Multilingual Prosodic Database. *Proceedings of the Fifth International Conference on Spoken Language Processing*, 7, 3163-3166.

Cheng, W., Greaves, C. & Warren, M. (2005). The Creation of a Prosodically Transcribed Intercultural Corpus: The Hong Kong Corpus of Spoken English (prosodic). In *International Computer Archive of Modern and Medieval English (ICAME) Journal*, 29, 47-68.

Dauer, R. (1987). Phonetic and Phonological Components of Language Rhythm. *Proceedings of the XIth International Congress of Phonetic Sciences*, 447-450.

Fant, G., Kruckenberg, A. (1989). Preliminaries to the Study of Swedish Prose Reading and Reading Style. In *Speech Transmission Laboratory. Quarterly Progress and Status Reports*, 1-83.

Grabe, E., Low, E.L. (2002). Durational Variability in Speech and the Rhythm Class Hypothesis. In Warner, N., Gussenhoven, C. (Eds.), *Papers in laboratory phonology* 7. Berlin: Mouton de Gruyter, 515-546.

t' Hart, J., Collier, R. & Cohen, A. (1990). *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press.

Hirst, D.J., Bigi, B., Cho, H., Ding, H., Herment, S. & Wang, T. (2013), Building OMProDat: An Open Multilingual Prosodic Database. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, 11-14.

Jensen, C. (2004). *Stress and Accent*. Ph.D. Dissertation, University of Copenhagen.

Kohler, K.J. (1996). Labeled Data Bank of Spoken Standard German: The Kiel Corpus of read/Spontaneous Speech. *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP)*, 3, 1938-1941.

Ling, L.E., Grabe, E. & Nolan, F. (2000). Quantitative Characterizations of Speech Rhythm: Syllable-timing in Singapore English. In *Language and speech*, 43(4), 377-401.

Mehler, J., Dupoux, E., Nazzi, T. & Dehaene-Lambertz, G. (1996). Coping with Linguistic Diversity: The Infant's Viewpoint. In Morgan, J.L. & Demuth, K. (Eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Mahwah: Lawrence Erlbaum Associates, 101-116.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and First Evaluation. www.lrec-conf.org/proceedings/lrec2000/pdf/110.pdf. Accessed 27.10.2018.

Ostendorf, M., Price, P.J. & Shattuck-Hufnagel S. (1995). The Boston University Radio News Corpus. In *Technical Report ECS-95-001*, 1-19.

Portele, T., Heuft, B. (1995). Two Kinds of Stress Perceptions. In Elenius, K., Branderud, P. (Eds.), *Proceedings of the 13th International Conference of Phonetic Sciences*, 1, 126-129.

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of Linguistic Rhythm in the Speech Signal. In *Cognition*, 73(3), 265-292.

Sakoe, H., Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1), 43-49.

Samlowski, B., Wagner, P. (2016). PromDrum-Exploiting the Prosody-gesture Link for Intuitive, Fast and Fine-grained Prominence Annotation. *Proceedings of Speech Prosody 2016*, Paper 212.

Schettino, V. (2015). Prosogit: A Corpus for Prosodic Studies in German L1 and Italian L2. In *AION*, 25(1/2), 237-255.

Streefkerk, B.M. (2002). *Prominence. Acoustic and lexical/syntactic correlates*. Utrecht: LOT.

Streefkerk, B.M., Pols, L.C.W. & Ten Bosch, L.F. (1999). Towards Finding Optimal Features of Perceived Prominence. In Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D. & Bailey, A.C. (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences*, 1-7 August 1999, 1769-1772.

Terken, J. (1991). Fundamental Frequency and Perceived Prominence of Accented Syllables. In *The Journal of the Acoustical Society of America*, 89(4), 1768-1776.

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., Escudero Mancebo, D., Gili Fivela, B., Lacheret, A., Ludusan, B., Moniz, H., Chasaide, A.N., Niebuhr, O., Rousier-Vercruyssen, L., Simon, A.C., Simko, J., Tesser, F. & Vainio, M. (2015). Different Parts of the Same Elephant: A Roadmap to Disentangle and Connect Different Perspectives on Prosodic Prominence. *Proceedings of the 18th International Congress of Phonetic Sciences*, Paper 202.

Watson, D.G., Arnold, J.E. & Tanenhaus, M.K. (2008). Tic Tac Toe: Effects of Predictability and Importance on Acoustic Prominence in Language Production. In *Cognition*, 106 (3), 1548-1557.

Wightman, C. (1993). Perception of Multiple Levels of Prominence in Spontaneous Speech. In *The Journal of the Acoustical Society of America*, 94(3), 1881-1881.