Acta Universitatis Sapientiae
Electrical and Mechanical Engineering, 6 (2014) 59–67

Analyzing F0 Discontinuity for Speech Prosody Enhancement

György SZASZÁK, Miklós Gábriel TULICS, Ákos Máté TÜNDIK

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Hungary
e-mail: {szaszak; tulics}@tmit.bme.hu
e-mail: akos.tundik@nokia.com

Manuscript received June 28, 2015; revised September 11, 2015.

Abstract: This research is interested in assessing the pros and contras of using an overall continuous versus a disrupted, not overall defined F0 estimate and compare formal and informal speech styles in this regard. During the evaluation we keep in mind the elaboration of an algorithm capable of handling both speaking styles. The approaches are evaluated in an automatic phonological phrasing task, using a formal and informal speech corpus. A phonological phrasing component is a prosodic unit that has its own stress and intonation contour, which may continue in the next unit. For the different speaking styles we use two speech databases: BABEL for formal speaking style and The Hungarian Spoken Language Database for informal speaking style, both in Hungarian language. Three alternatives of F0 post-processing are compared, ranging from a natural F0 contour disrupted at unvoiced places, over a partial interpolation to an overall continuous contour estimation defined for all unvoiced speech segments. Whereas in formal speech, the more continuous the F0 contour is the better detection rates are observed for phonological phrases, for informal speech a partial interpolation of F0, preserving some fragmentation, yields better results. These results also show that discontinuity of F0 can be an important cue in human perception of informal speech, and also means that the idea of trying to de-spontanize spontaneous speech, in order to be able to treat them in the conventional way, seems to be doubtful.

Keywords: speech prosody, event detection, machine learning, speech signal processing.

1. Introduction

Speech prosody is an important building block of spoken language, carrying information related to several modalities, i.e. prosody provides cues for broad segmentation of the speech stream, carries stress/emphasis, has a discourse

function, reflects speaker attitude and emotions etc. An important speech technology application exploiting prosody is automatic prosodic event detection, i.e. stress detection, or automatic phrasing (segmentation for prosodic units) of the speech flow. Prosodic event detection can be a basic pre-processing step in content or discourse analysis, speech-to-speech translation, that is, in automatic speech understanding applications in general. The three basic acoustically measurable features building up prosody are F0 (fundamental frequency), energy and duration, and they can sometimes be complemented with other, less frequently used features such as jitter, shimmer, harmonics-to-noise ratio (HNR), sub-band energies, etc. In the realisation of prosodic constituents, like stress or intonation patterns, the three basic features interact. The contribution of different features in accomplishing given linguistic functions is language dependent [10], for example, duration is an important cue of stress in American English or German [15], whereas in Hungarian, F0 is believed to be the dominant cue of stress with duration playing almost no role in it [3].

Extracting energy is a basic task, which can usually be carried out without complications. Extraction of duration patterns, on the other hand, may pose problems, especially if the underlying segmental structure (phone segmentation) of the speech signal is unknown. The biggest challenge in this field remains the accurate extraction of F0 [4]. Although several reliable algorithms are known, the F0 estimate is often corrupted by doubling/halving errors [12]. Moreover, the F0 contour is not continuous (it is undefined for unvoiced speech segments), but the human perception of F0 is capable of keeping track of the pitch as if it were continuous. Recent research has found that using a continuous F0 estimate can be advantageous in several applications [5]. Indeed, in speech technology applications, F0 is often interpolated to overcome problems caused by discontinuity [11]. An alternative to this approach has recently been proposed based on probabilistic features [6].

The current paper is interested in exploring these advantages and eventual disadvantages in an automatic prosodic event detection task. A special preference of the evaluation is to keep in mind the elaboration of an algorithm capable of handling both read (formal) and spontaneous (informal) speaking styles. Several studies have proposed phrasing approaches for read and slightly spontaneous speech [1, 2]. In [3] an automatic prosodic phrasing system was implemented for read speech which was able to gain a reliable phrasing down to the phonological phrase level, and also to separate intonational phrase level from the underlying phonological phrase level. The accuracies were ranging between 70-80%. In this approach, prototypes of phonological phrases were clustered, modelled by Hidden Markov Models/Gaussian Mixture Models (HMM/GMM) based on acoustic-prosodic features. In [7] an unsupervised

learning approach for clustering prosodic entities in Hungarian spontaneous speech was described, clustering of such characteristic prototypes led only to partial success.

In this paper the authors compare read and spontaneous speech processing in a prosodic event detection related task, and aim at exploring the advantages and disadvantages of using a continuous vs. disrupted F0 data stream. Three ways of F0 estimations are studied: a continuous, overall interpolated F0 estimation is compared to cases where F0 is fragmented or interpolated only partially. These approaches are evaluated in automatic phonological phrasing system, both for a formal and informal speech corpora. We also believe that these experiments may lead us to a better understanding of spontaneous speech.

This paper is organized as follows: first, the used speech databases are presented, followed by a brief overview of the embedding prosodic event detection system. Experiments are described thereafter, followed by the presentation of the results. Finally, conclusions are drawn.

2. Material and method

This section describes speech databases and basic processing tools used for the experiments later.

2.1 Speech Databases

We use two databases with different speaking styles for the experiments; a formal and an informal one. Both databases got prosodic annotation, representing the top layers of the prosodic hierarchy [9], as follows:

- Intonational Phrases (IP): The IP is the prosodic unit positioned at the top of the prosodic hierarchy, just below the utterance level. The IP is interpreted as a segment of speech that has its own complete prosodic contour, where the first word is accentual (in Hungarian). IP is often found between two pauses.
- Phonological Phrases (PP): The IPs can be divided into PPs: A PP is a prosodic unit that has its own stress and intonation contour, but this contour can be continued in next PP. Syntactically, PP is often related to clitic groups, at least in formal speaking style.

2.1.1 BABEL

The Hungarian BABEL is a read speech corpus (formal speaking style), recorded in a low-noise environment [13]. 60 native Hungarian speakers (30-30

male/female) of varying age and professional background read the utterances. A subset of the database is composed of paragraphs of 5-6 sentences. From this subset, we randomly selected 300 sentences from 22 speakers, and labelled them manually for phonological phrases (2067 in total), according to the 7 types described in [3]. The PP annotation is such that it reveals the IP boundaries unambiguously as well.

2.1.2. Hungarian Spoken Language Database

The Hungarian Spoken Language Database (BEA) [8] is the first Hungarian spontaneous speech database that involves several hundred speakers and a rich speech material in semi-informal and informal speaking styles. The speech material contains different kinds of spontaneous narratives and discourses about personal life, everyday topics, but also includes some sentence repetitions. Informal utterances from 8 speakers (4 female and 4 male) were selected from the database. This subcorpus was manually annotated by two experts for IPs (398 in total) and PPs (751 in total). Again, all phrase boundaries are aligned with word boundaries (this requirement is relatively easy to fulfil in fixed stress Hungarian; however, if stress were unbound, it would not be impossible either).

2.2. Automatic Segmentation for Phonological Phrases

This section describes the automatic PP segmentation algorithm used in the experiments. In this paper, we chose PP segmentation to analyse the effects of varying the post-processing for F0. Before moving on to the experiments, we provide here a brief overview of the system.

As already mentioned, PPs constitute a prosodic unit, characterized by an own stress and some preceding/following intonation contour. The intonation contour might be incomplete (following in the next PP or truncated in informal speech), however, it is specific, hence PPs can be classified/modelled separately, in a data-driven machine learning approach. The distinction between PPs consists of two components: the strength of stress the PP carries and the PPs intonation contour. In this way 7 different types of PPs are distinguished. Acoustic modelling is done with HMM/GMM models, and PP segmentation is carried out as a Viterbi alignment of PPs for the utterances requiring segmentation. The overall approach is documented in detail in [3], featuring as well the 7 types chosen for modelling.

As acoustic-prosodic features, fundamental frequency (F0) and wide-band energy (E) are used. Syllable duration is not used for Hungarian as it was not found to be a distinctive cue in this task [3]. F0 extraction and post-processing alternatives for F0 are described in the respective section later. For energy

computation a standard integrating approach is applied with a window span of 150 ms. Frame rate is 10 ms. First and second order deltas are appended to both F0 and E streams. The evaluation of the PP segmentation is carried out in 10-fold cross-validation. First a PP alignment is generated with models trained on utterances different from the one under segmentation. The generated PP alignment is then compared to the reference obtained by manual labelling. Detection is regarded to be correct if the boundary is detected within the TOL=100-250 ms vicinity of the reference. Here, TOL defines a tolerance interval, ranging between 100 and 250 ms. These values are chosen so to be in the order of a length of a syllable, on average, and much less than average PP length. Once all utterances have the automatic PP segmentation ready, the following performance indicators are evaluated for the PP boundaries:

- recall

$$\left(RCL = \frac{TP}{TP + FN} \right) \tag{1}$$

- precision

$$\left(PRC = \frac{TP}{TP + FP} \right) \tag{2}$$

- the average time deviation (ATD) between the detected and the reference PP boundary. TP refers to the number of true positive, FN to the number of false negative and FP refers to the number of false positive retrieved PP boundaries.

3. Overall and Partial Interpolation for F0

F0 was extracted using the Snack toolkit [14]. F0 data have been subjected to error correction resulting from octal halving/doubling. Keeping the energy unchanged, further post-processing of F0 varies as follows:

- apply only and octal correction, then use the F0 contour as produced by a conventional pitch tracker (Snack V2.2.10 in our case);
- use a continuous contour, interpolated at all unvoiced parts;
- use a partially F0 interpolated contour. This means that the interpolation is omitted if the length of the unvoiced interval is higher than 250 ms, or if the starting F0 value is significantly higher than the value before the unvoiced segment. The criteria used was:

$$F0_{former} \times 1.1 < F0_{current} \tag{3}$$

Speech containing longer periods of silence than 250 ms can hardly be considered as fluent speech. In such situations speakers often may not reset their pitch, so IP (and hence the co-occurring PP) boundaries can be marked by silence. In such cases silence is an acoustic marker by itself and we prevent the interpolation, which may mask the PP boundary. If a medium strength pitch reset occurs it might be smoothed by the F0 interpolation. In the vicinity of long plosives for example, microprosodic disturbances can give false phrase boundary detection. The factor of 1.1 is set to avoid these false detections. These were the motivations to constrain the disruption of the F0 contour in the partial interpolation scenario.

Table 1: Precision (PRC) and recall (RCL) in operating points defined by PRC = RCL and ATD for the 3 evaluated scenarios for formal and informal speech styles TOL = 100 ms).

Style	F0 interpolation	PRC = RCL [%]	ATD [ms]
Formal style	None	62.9	93.0
	Partial	68.2	80.1
	Total	81.2	55.9
Formal style	None	57.7	52.9
	Partial	69.7	44.1
	Total	66.3	44.8

4. Results

Results are shown in Table 1, involving all three tested scenarios for formal and informal speaking styles separately. PP detection rates have different characteristic depending on the speaking style.

Regarding precision (PRC) and recall (RCL), they highly depend on a parameter influencing PP insertion likelihoods during the PP segmentation done with Viterbi alignment (see the PRC-RCL curve in *Fig. 1*). Therefore, segmentation results are shown for operating points where precision and recall are equal (PRC=RCL).

The settings of the tolerance interval TOL also influence results (see *Fig. 2*). We may consider that TOL=100ms corresponds roughly to the length of a half syllable on average and hence TOL=200ms is in the order of the length of a syllable.

As expected, formal and informal speech styles show different characteristics. For read speech, PP detection results are better if the F0 contour is overall interpolated, whereas for the informal speaking style, the partially interpolated F0 contour gives the best results. In case of partial interpolation,

the interpolation was applied only if the length of the unvoiced segment was below the limit of 250 ms, and pitch reset was not suspected.

The results suggest that in the perception of spontaneous speech characterized by informal speaking style, the discontinuity of F0 plays an important role. This also means that the idea of trying to de-spontanize spontaneous speech to that of read speech, in order to treat it with conventional algorithms or tools developed for read speech, seems to be doubtful.

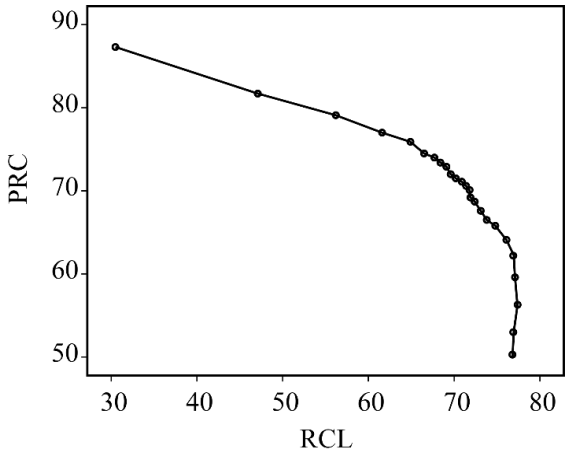

Figure 1: Precision [%] and recall [%] as influenced by the PP insertion likelihood in the automatic PP segmentation. Read speech, total F0 interpolation, TOL = 100ms.

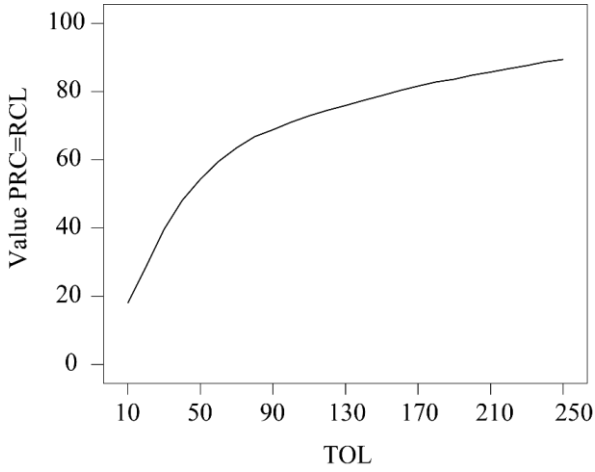

Figure 2: Precision and recall in operating points defined by PRC = RCL [%] depending on TOL [ms]. Read speech, total F0 interpolation.

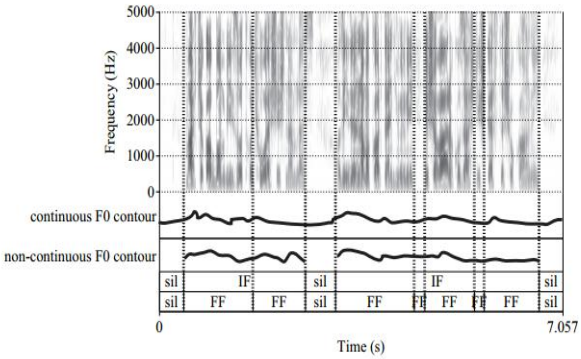

Figure 3: An example with continuous and partially interpolated F0 contours with IP and PP labelling.

5. Conclusions

In this research the authors examined the effect of F0 discontinuity in a phonological phrase segmentation task and drawn a comparison between overall interpolated (continuous) and partially interpolated (fragmented) F0 estimates. Results indicate that the overall interpolated F0 estimate is useful only in read or formal speaking styles. Overall interpolated F0 contour showed better results compared to the partially interpolated one in the phonological phrase segmentation task. The difference between the two methods' precisions was relative 19.1%. The best performing system yielded 81.2% recall and precision in the operating point where type I and type II errors are equal. In terms of precision, this result is fairly comparable to the reported accuracy of other phrase detection tasks by excellent recall rates. Furthermore, a partially interpolated F0 estimate which leaves longer unvoiced periods or pitch reset untouched outperformed the total interpolation one by relative 5.1% for informal speaking style in the same phonological phrase segmentation task. The best precision and recall in the operating point characterized by equal errors was 69.7, which is acceptable for spontaneous speech in international comparison. It follows that piecewise interpolation of F0 seems to be better for spontaneous speech in speech technology applications. Results also suggest considerations regarding human speech perception, but for those we need further investigations with targeted experiments.

Acknowledgements

The authors would like to thank the support of the Hungarian Scientific Research Fund (OTKA) under contract ID PD 112598, titled "Automatic Phonological Phrase and Prosodic Event Detection for the Extraction of Syntactic and Semantic/Pragmatic Information from Speech."

References

[1] Veilleux, N. M., Ostendorf, M., "Prosody/parse scoring and its application in atis," in *Proceedings of the workshop on Human Language Technology*, 1993, pp. 335–340.

[2] Gallwitz, F., Niemann, H., Nöth, E., Warnke, W., "Integrated recognition of words and prosodic phrase boundaries," *Speech Communication* 36, 1-2 (2002) 81–95.

[3] Szaszák, G., Beke, A., "Exploiting Prosody for Automatic Syntactic Phrase Boundary Detection in Speech," *Journal of Language Modeling* 1, 0 (2012) 143–172.

[4] Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., Renals, S., "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, 17, 6 (2009) 365–375.

[5] Yu, K., Young, S., "Continuous F0 modelling for HMM based statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, 5, 19 (2011), pp. 1071–1079.

[6] Garner, P. N., Cernak, M., Motlicek, P., "A Simple Continuous Pitch Estimation Algorithm," *IEEE Signal Processing Letters* 20, 1 (2013) 102–105.

[7] Beke, A., Szaszák, Gy., "Unsupervised clustering of prosodic patterns in spontaneous speech, Text, Speech and Dialogue," *Lecture Notes in Computer Science*, 7499 (2012) 648–655.

[8] Neuberger, T., Gyarmathy, D. Gráczi, T. E., Horváth, V., Gósy, M., Beke, A., "Development of a large spontaneous speech database of agglutinative Hungarian language, Text, Speech and Dialogue," *Lecture Notes in Computer Science* 8655 (2014) 424–431.

[9] Selkirk, E., "The Syntax-Phonology Interface," in the International Encyclopedia of the Social and Behavioral Sciences, Pergamon Press, Oxford, 2001, pp. 15407–15412.

[10] Hirst, D., and Di Cristo, A., "Intonation Systems: A Survey of Twenty Languages," Cambridge University Press, New York 1989, p. 256.

[11] Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S., "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, 2014, pp. 2494–2498.

[12] Murray, K., "A study of automatic pitch tracker doubling/halving Errors," in *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001, 16 pp. 1–4.

[13] Roach, P. S., Amfield, S., Bany, W., Baltova, J., Boldea, M., Fourcin, A., Goner, W., Gubrynowicz, R., Hallum, E., Lamel, L., Marasek, K., Marchal, A., Meiste, E., Vicsi, K., "BABEL: An Eastern European Multi-language database," in *Proeedings of the International Conference on Speech and Language* 1996, pp. 1033–1036.

[14] Sjölander, K., Beskow, A., "Wavesurfer – an open source speech tool," in *Proceedings of the 6th International Conference of Spoken Language Processing*, 2000, 4, pp. 464–467.

[15] Tamburini, F., Wagner, P., "On Automatic Prominence Detection for German," *Proceedings of Interspeech*, 2007, pp. 1809–1812.