Consistency of Prosodic Annotation of Spontaneous Speech for Technology Needs

Jolanta Bachan, Agnieszka Wagner, Katarzyna Klessa & Grażyna Demenko

Faculty of Modern Languages and Literatures, Institute of Linguistics, Department of Phonetics
Adam Mickiewicz University in Poznań, Poland
jolabachan@gmail.com, wagner@amu.edu.pl, klessa@amu.edu.pl, lin@amu.edu.pl

Abstract

Speech prosody is the carrier of information essential for speech processing and understanding. In this paper, we present a prosodic annotation specification for spontaneous speech, created for the needs of a Polish ASR system (Demenko et al., 2011; Demenko et al. 2012) with a view to improve the system's performance. We also report on an interlabeller consistency experiment carried out in order to evaluate the proposed annotation specification. Results of annotation of a 70-minute speech corpus by 6 trained annotators are inspected with regard to labelling consistency with a special focus on boundary labels insertion and prominence annotation. The results of annotation agreement analysis are comparable with the results presented in the literature and imply simplifying the annotation specification.

Keywords: prosodic annotation, annotation consistency, ASR

1. Introduction

Three factors shape the prosodic structure of an utterance: (1) the distribution and strength of syllable prominence which is connected with the accent/stress and located in those places of the utterance where new or key information for the discourse occurs, (2) the distribution and strength of prosodic phrase boundaries connected mainly with the syntactic structure of the utterance, and (3) intonation (i.e. pitch contours of various kinds).

The main interest of the present study is to investigate selected aspects related to the first two of the factors mentioned above. Prominence and phrasing might be realised by speakers using a range of acoustic cues such as segmental durations, pitch, intensity, spectral characteristics, and voice quality which the listeners effectively use for the representation of the prosodic structure. However, the perception of prosodic structure, and above all, the prominence, is affected by the linguistic and non-linguistic factors which need to be taken into account while analysing the results of perception-based prosodic annotation. Multiple approaches to prominence and the diversity of its possible correlates lead to emergence of a wide variety of definitions and ways of understanding of this term (Wagner et al., 2015).

The use of prosodic information in speech technology systems has been a subject of many studies in the recent years (Gibbon, 1992; Batliner and Möbius, 2005; Maier, 2009; Demenko et al., 2011; Windmann et al., 2011). It is already known that the effective modelling of prosody is reflected in the naturalness of speech generated by speech synthesis systems, and in automatic speech recognition using prosodic information can have a significant impact on the correctness of recognition. In both cases, it is necessary to create speech corpora containing annotations providing consistent and reliable information on the course of the prosodic structure of speech, and the creation of tools for annotation mining, processing and further use.

In this paper, a prosodic annotation specification based on speech perception of Polish spontaneous speech is introduced. The specification was designed with a view to improve performance of an automatic speech recognition system for Polish (Demenko et al., 2011; Demenko et al. 2012). The current study is conducted on the only corpus of Polish spontaneous speech containing such detailed prosodic annotations for highly diversified authentic speech material.

2. Critical assessment of existing prosodic annotation systems

The usefulness of the prosodic transcription system can be assessed by examining the consistency of annotations made in accordance with its specifications on the same material for at least several listeners, where high consistency proves the validity of such a system. The results of annotation agreement in ToBI (*Tones and Break Indices*) (Silverman et al., 1992; Beckman and Hirschberg, 1994; Beckman and Ayers, 1997; Beckman et al., 2004) are divergent because of the type of verbal material (read vs. spontaneous speech) and transcription (full vs. simplified). As shown in the work of Yoon et al. (2004), the agreement of annotations as regards the position of melodic accents (prominence) in spontaneous speech was 86%, while for prosodic phrase boundaries it reached 89%. A similar study of the RaP system (*Rhythm and Pitch*) (Breen et al., 2012) showed that for a binary decision on prominence, the annotation agreement stands at 89%, and in the case of the boundaries of phrases at 90% (Breen et al., 2012). In both systems, the transcription of consistency or agreement significantly decreases with increasing accuracy of annotation (Breen et al., 2012; Grice et al., 1996; Jun et al., 2000; Pitrelli et al., 1994; Syrdal and McGory, 2000; Yoon et al., 2000). Restricting the prosodic annotation to the layer called *rhythm* in the RaP system, and thus denoting only the strength/level of prominence and strength/level of phrase boundaries and leaving out the tonal description, i.e. the types of accents (e.g. H + L *) and phrasal and boundary tones (e.g. LL %), seems justified. In the context of speech recognition such information is less important, and at the same time it reduces significantly the

consistency of annotations and slows its course. Overall, the results indicate the need to reduce the number of possible labels and to simplify a very detailed prosodic annotation to achieve a more reliable and a more general annotation specification, but still taking into account the most relevant information from the point of view of the prosodic structure of utterances. The above postulate was the basis for the creation of specifications of prosodic annotation of a spontaneous speech corpus for the present ASR system.

3. Specification of prosodic annotation of spontaneous speech

The present specification of prosodic annotation was created for the need of using prosodic information in a Polish ASR system to improve the correctness of automatic speech recognition. In the description of prosodic phenomena, the following factors are taken into consideration: two levels of prominence of syllables - *strong* and *weak-medium* prominence enhancement, and three levels of prosodic phrase boundary strength - the *weak*, *medium* and *strong* boundaries. In addition, elements of discourse with high impact on the prosodic structure of speech are taken into account.

The annotation of prominence and phrase boundaries are guided by both meaning, i.e. the syntactic, semantic and discourse cues, and the acoustic features of speech. In order to reconcile the two criteria, (1) labels marking weak phrase boundaries were introduced (boundary type /) in places where syntactically and semantically such a boundary occurs, but the acoustic cues are very subtle, and (2) labels indicating ungrammatical phrase boundaries (type \) which are clearly marked by prosody, but appear in "unexpected" locations from the point of view of the semantic, syntactic and/or discourse structure of an utterance. The proposed specification of prosodic annotation is summarised in Table 1.

4. Prosodic annotation: analyses and results

4.1. Speech material and annotators

The analysed material consisted of 70 minutes of authentic recordings of spontaneous speech: 36 minutes of the material was transcribed in advance (it contained text and annotation of noises, e.g., fillers, speaker noises like breathing, intrusive and stationary noises coming from outside, labels indicating the excerpt as a trash), another 34 minutes of the material only contained recordings with no annotation or transcribed text provided in advance. The material was composed of 5 categories of recordings:

- P – parliamentary speeches,
- I – police inspections,
- DT – telephone conversations,
- DC – informal dialogues in a car,
- C – conference lectures.

All the material was annotated independently by 6 people. All annotators were trained according to the same procedure and worked in the same quiet laboratory room, the speech signals were presented from PCs via headphones.

Label	Usage
2	weak prominence
3	strong prominence
1	difficult to tell whether prominent or not
/	weak phrase boundary
//	phrase boundary of medium strength
///	strong phrase boundary
%	hard to asses strength of boundary (/// or //)
$	grammatical indication of phrase boundary but prosodic cues are not unequivocal
\	ungrammatical boundary (speaker inserts a pause/breath/filler within the utterance)
!	word or phrase realised with emphasis
{*text*}	inclusion
/..	apposition
/@	backchannel
/~	incomplete utterance cut off at the end
~/	incomplete utterance cut off at the front

Table 1. Summary of specification of perceptual annotation of prosody

4.2. Interlabeller consistency analysis & annotation mining

A sample annotation of a short excerpt of the recordings is presented in Table 2. The graph displayed in Figure 1 presents the number of the differences found in this excerpt measured using Levenshtein distance (Gibbon, 2014).

For annotation agreement assessment, an average agreement metric across all coders and items from NLTK was used, *avg_Ao* (Lippincott, 2014): *"The agreement coefficient calculates the amount that annotators agreed on label assignments beyond what is expected by chance"*. Also *pi* (Scott, 1955) (here, *multi-pi*, equivalent to *K* from Siegel and Castellan (1988)) and *s* (Bennett et al., 1954) were used, but their results are not presented here because they were very similar to *avg_Ao* values. The annotation material was processed earlier in order to align the annotations and create triples[1]:

<annotator_ID, number_of_item, label>

	Text with the annotation
A1	jest to po3dział / taki na2sz umo3wny / mo2żna powie3dzieć //, [spk=b]
A2	jest to po2dział taki na1sz umo3wny /, mo1żna powie3dzieć //, [spk=b]
A3	jest to po3dział ta1ki na1sz \ umo3wny /, można powiedzieć /, [spk=b]
A4	jest to podzia3ł taki na2sz \ umo3wny można powie2dzieć /,
A5	jest to po2dział taki nasz umo3wny //, mo2żna powie3dzieć //,
A6	jest to po2gląd ta2ki nasz umo3wny / mo2żna powie3dzieć ///. [spk=b]

Table 2. A sample result of annotation of the same phrase by 6 people

[1] The authors wish to thank Dr. Ewa Kuśmierek (Poznań Supercomputing and Networking Center) for developing annotation alignment software and preparing data for agreement analysis.

Additionally, for this part of the analysis, all the prominence labels were removed, and the phrase boundary labels were normalised to one label "&" followed by the punctuation mark.

Annotation agreement for the whole material, i.e. 70 min. 52 sec. of speech altogether (55 files, the mean length of a file was 77 sec., 11226 items for each annotator) equals: *avg_Ao* – 0.809, where 1.000 means complete agreement. Annotation agreement of the material where text was already available was of 0.086 point better, and the annotation performed on the material without any text inserted earlier was the most coherent for parliamentary speeches (Table 3). The biggest differences were found in the annotation material of conference recordings.

In order to further inspect the interlabeller variability in the annotated material, we analyse the labellers' agreement in prominence ratings, boundary markers and other types of labels included in the specification of prosodic annotation. Figure 2 shows the quantitative summary of prominence labels inserted in the annotations by all labellers. Figure 3 and Figure 4 show statistics of phrase boundary labels, without punctuation marks for the boundaries /, //, /// and %.

The results of statistical analyses of the frequency of occurrence of prominence and phrase boundary labels show substantial differences in the usage of the labels by particular annotators, cf. e.g., the difference in the number of labels inserted by the annotators A4 and A6 (Figure 4), which indicates a significant subjectivity in the perception of phrase boundaries and prominence of syllables. Additionally, it was observed that some labels, e.g. {} and % were very rarely used, therefore their removal from the specification guidelines should be considered. However, the results presented here also show a high interlabeller agreement with regard to the general shape of the structure of the phrasing of speech (the position of a boundary, without its strength) and they are comparable with the results presented in the literature (see Section 2).

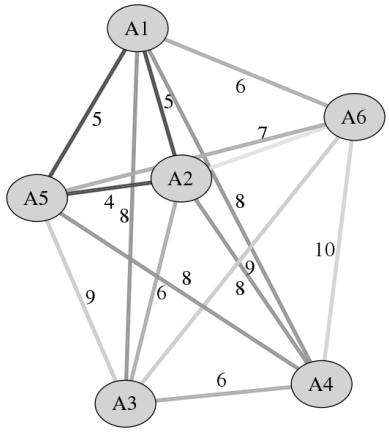

Fig. 1: Graph showing the number of differences among annotations presented in Table 2 measured using Levenshtein distance (Gibbon, 2014)

Rec. Cat.	Time	File/ Avg_time	Items(*6)	*avg_Ao*
Material with text text				
P	15m 40s	12/78 s	2585	0.858
I	5m 10s	4/77 s	525	0.873
DT	13m 51s	13/64 s	2311	0.829
DC	1m 28s	2/44 s	219	0.860
All	36 8s	31/70 s	5640	0.857
Material without text text				
P	15m 13s	9/101 s	2253	0.861
I	5m 27s	9/81 s	618	0.710
DT	12m 6s	2/59 s	2357	0.711
C	1m 58s	4/82 s	358	0.698
All	34m 43s	24/87 s	5586	0.771

Table 3. Results of annotation agreement for each of the recording categories, on the material with and without text. The number of items is provided per annotator (the total number of items used for the analyses was thus 6 times higher than given in the table).

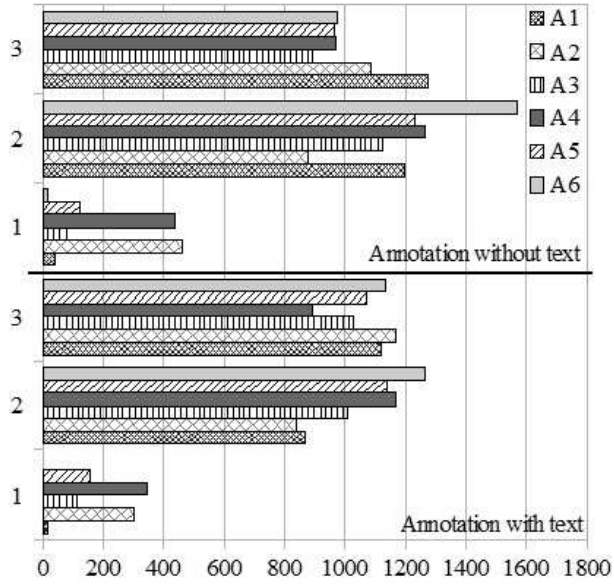

Fig. 2: The results of prominence annotation

In order to preliminarily inspect the potential acoustic-phonetic correlates of perceived prominence based on the present speech material, we looked, among others, at segmental durations. As reported in the literature (e.g., Wagner et al., 2015), segmental duration can be an important cue for the perception of prominence. The average duration of the realizations of the vowel /a/ with different prominence labels (1, 2, 3, as defined in Table 1) extracted from annotations of the annotator A3 is presented in Figure 5. In Polish, /a/ can occur as part of other words as well as a self-standing conjunction *a* (Eng. 'and', 'versus'); the conjunctions were excluded from this part of analyses. As it can be seen, the durations appear to be the longest for the realizations labelled as strong prominences and shorter for the weaker ones. This indicates that at least for this particular labeller, segmental duration could be confirmed as one of the correlates of the strength of perceived prominence in

Polish spontaneous speech. The nature of the relationships and the variability of timing patterns as an extended interlabeller comparison is a subject of a separate on-going investigation.

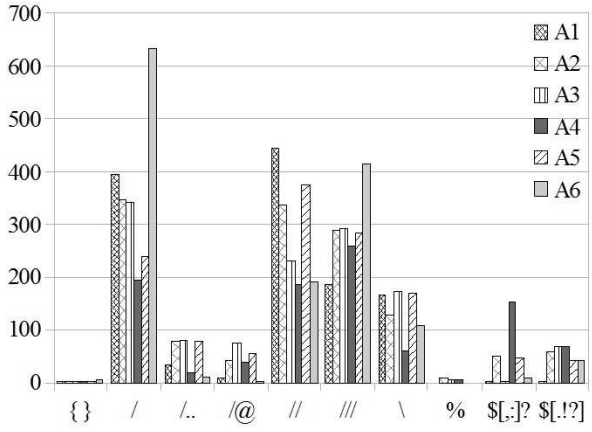

Fig. 3: Summary of the results of annotation of phrases in the material with text

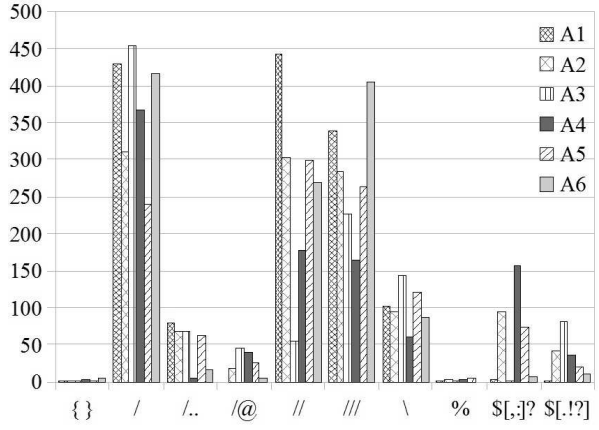

Fig. 4: Summary of the results of annotation of phrases in the material without text

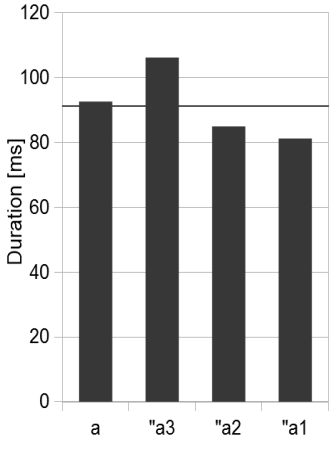

Fig. 5: Mean durations of the realizations of /a/ vowel depending on prominence labels (horizontal line denotes the overall mean).

5. Discussion

As shown by a preliminary analysis of the consistency of annotation, the proposed specifications require some modifications (e.g. merging some of the labels and the removal of others), and the overall results of annotation agreement of phrase boundaries do not differ significantly from those presented in the literature (see Section 2). The results imply the need to simplify the specifications (in terms of the number of levels of prominence and phrase boundaries).

However, at this stage of analysis, the proposed annotation specification seems to be generally well suited for application in the training corpus for an ASR system, because it is fully based on human perception (no visual inspection needed), and does not require perceptual assessment of fundamental frequency levels – a task which is very difficult for the annotators who are not musically talented (opinion based on authors' experience in this kind of annotation). The annotation labels are also inserted in the text/transcription layer of annotation which speeds up the costly annotation process and does not require multilayer annotation. It should be admitted however, that for certain purposes the analysis of data obtained from multilayer annotation would be more effective and easier, due to better separation of labels (less advanced text parsing and data processing required).

6. Conclusion

The proposed specification of prosodic annotation of spontaneous speech turned out to be quite complex and the results of the annotation consistency study showed the need of its simplification. The overall result of annotation agreement, 0.809, was reasonable considering the type of annotated material (spontaneous speech, dialogues in diversified recording environment), and the quality of recordings which was very low in some cases. However, the rare or uncommon use of some labels imply their removal from the annotation specification and integration of their functions with other labels.

The frequency of occurrence of the three prominence labels for particular recording categories tends to be relatively similar in the annotations of the 6 labellers, although obviously certain individual differences are still present. An interesting future work would be to investigate the specifity of these individual differences and similarities in the perceived levels of prominence and their acoustic-phonetic correlates. The pilot study of vowel segmental durations reported above confirms the expected positive correlation between the level of prominence and duration of realisations of /a/.

Generally, each prosodic transcription must be statistically verifiable with respect to a set of objectives corresponding to its assumed application. The objectives enabling implementations in speech technology systems on one hand impose certain level of detail, formalisations and constraints and, on the other, require simplicity. Informal preliminary tests showed also that the enrichment of prosodic annotation through perceptual analysis, based not only on the strength of boundaries and accent prominence, but also on features such as ungrammatical structure and discourse markers, is a good basis for automation of modelling of Polish accent, especially for the detection of phrase boundaries.

Acknowledgments

The present study was supported by Polish National Centre for Research and Development, project no.: **DOBR/0008/R/ID1/2013/03,** *"A system of automatic recognition of Polish speech to text in the environment of creating and circulation of documents within the law enforcement agencies and the judiciary",* http://speechlabs.pl/.

References

Batliner, A. and Möbius, B. (2005). Prosodic Models, Automatic Speech Understanding, and Speech Synthesis: Towards the Common Ground? In: *The Integration of Phonetic Knowledge in Speech Technology*, Springer Dordrecht, 2005, 21-44.

Beckman, M.E. and Hirschberg, J. (1994). *The ToBI annotation conventions*. Ohio State University.

Beckman, M.E. and Ayers, G. (1997). Guidelines for ToBI labeling (Version 3.0). Manuscript and accompanying speech materials. The Ohio State University. Retrieved from: http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf. Access date: September 19, 2015.

Beckman, M.E., Hirschberg, J.B. and Shattuck-Hufnagel, S. (2004). The original ToBI system and the evolution of the ToBI framework. In: *Prosodic models and transcription: Towards prosodic typology*, 9-54.

Bennett, E.M., Alpert, R., and Goldstein, A.C. (1954). Communications through limited response questioning. In: *Public Opinion Quarterly* 18, 303–308.

Breen, M., Dilley, L. C., Kraemer, J. and Gibson, E. (2012). Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). In: *Corpus Linguistics and Linguistic Theory* 8(2), 277-312.

Demenko, G., Szymański, M., Cecko, R., Lange, M., Klessa, K. and Owsianny, M. (2011). Development of Large Vocabulary Continuous Speech Recognition using phonetically structured speech corpus. In: *Proceedings of ICPhS 2011*, Hong Kong.

Demenko, G., Szymański, M., Cecko, R., Kuśmierek, E., Lange, M., Wegner, K., Klessa, K. and Owsianny, M. (2012) Development of Large Vocabulary Continuous Speech Recognition for Polish. In: *Acta Physica Polonica A*, 121, 86-91.

Gibbon, D. (1992). Prosody, time types and linguistic design factors in spoken language system architectures. In: G. Görz, (ed.), *KONVENS '92*. Berlin, Springer, 90-99.

Gibbon, D. (2014). Visualisation of distances in language quality spaces: DistGraph, a teaching tool for language typology data mining. Bielefeld: U Bielefeld. Retrieved from: http://wwwhomes.uni-bielefeld.de/gibbon/DistGraph/. Access date: September 19, 2015.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J. and Batliner, A. (1996). Consistency in transcription and labelling of German intonation with GToBI. In: *Proceedings of ICSLP*, 1716–1719.

Jun, S. A., Lee, S. H., Kim, K. and Lee, Y. J. (2000). Labeler agreement in transcribing korean intonation with K-toBI. In: *Proceedings of Interspeech 2000*, 211-214.

Lippincott, T. Natural Language Toolkit: Agreement Metrics. Copyright (C) 2001-2015 NLTK Project. Retrieved from: http://www.nltk.org/_modules/nltk/metrics/agreement.html#AnnotationTask.avg_Ao. Access date: September 19, 2015.

Maier, A., Hönig, F., Zeissler, V., Batliner, A., Körner, E., Yamanaka, N., Ackermann, P.D. and Nöth, E. (2009). A language-independent feature set for the automatic evaluation of prosody. In: *Proceedings of Interspeech* 2009), Brighton, England, 6.-10.9.2009, pp. 600-603.

Pitrelli, J. F., Beckman, M. E. and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the tobi framework. In: *Proceedings of ICSLP*, 123–126.

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. In: *Public Opinion Quarterly*, 19(3), 321-325.

Siegel, S. and N.J. Castellan, Jr. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw Hill, Boston, MA.

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., Price, P. and Hirschberg, J. (1992). ToBI: A standard scheme for labeling prosody. In: *Proceedings of the Second International Conference on Spoken Language Processing*, 867-879.

Syrdal, A. K. and McGory, J. T. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In: *Proceedings of Interspeech 2000*, 235-238.

Wagner, P., Origlia, A., Avesani, C., et al. (2015). Different parts of the same elephant: a roadmap to disentangle and connect different perspectives on prosodic prominence. In: *Proceedings of the ICPhS*, Glasgow, UK.

Windmann, A., Jauk, I., Tamburini, F. and Wagner, P. (2011). Prominence-Based Prosody Prediction for Unit Selection Speech Synthesis. In: *Proceedings of Interspeech 2011.*

Yoon, T., Chavarria, S., Cole, J. and Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In: *Proceedings of Interspeech 2004*, 2729–2732.