Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

Acoustic Cues for Automatic Determination of Phrasing

Agnieszka Wagner

Department of Phonetics, Adam Mickiewicz University in Poznań, Poland
wagner@amu.edu.pl

Abstract

This paper proposes a framework of automatic determination of phrasing using acoustic features derived from the speech signal. The feature vectors were defined in a series of analyses investigating the acoustic-phonetic realization of minor and major phrase boundaries and different boundary types. The resulting representation was used to train statistical classifiers to automatically determine phrase boundary position and type. The output of the classifiers can be used to provide speech corpora with phrasing information to enhance the performance of TTS or ASR systems, or to generate a comprehensive feedback in prosody tutoring systems. Apart from providing an efficient means for automatic phrase boundary detection, the study presented in this paper sheds also light on the role of timing and F0 cues in signaling phrase boundaries.

Index Terms: phrasing, boundary tones, prosody labeling

1. Introduction

Next to prominence, *phrasing* belongs to the most important linguistic functions of intonation. It organizes an utterance into a hierarchical prosodic structure. Intonation phrases which occur almost at the top of this hierarchy (below *utterance*) include one obligatory (nuclear) pitch accent and are characterized by semantic and syntactic coherence. They constitute the domain of recurring intonation patterns and can be considered as units of information [1]. Although the correspondence between syntactic and semantic units on the one hand, and intonation phrases on the other is not straightforward, it is generally agreed that to some extent intonation phrases correspond to clauses.

1.1. Functional aspects of phrasing

Apart from a binary distinction between boundary presence vs. absence intonation phrase boundaries are classified with respect to *strength* (minor vs. major phrase boundary) and *type* (rising – signaling continuation or interrogation as in yes/no questions, falling – characteristic of statements, but also wh-questions). This kind of phrasing information can be used to resolve ambiguous parses and to disambiguate the meaning that can be assigned to a given phrase by the hearer e.g. "what?" can be interpreted as a request for repetition of what the interlocutor has just said when realized with a rising boundary (– I have it. – What? – I have it.) or as a request for further information when realized with a falling pitch at the phrase boundary (– I have it. – What? – A dress.). This kind of knowledge can enhance the performance of the language model component of ASR or dialogue systems, e.g. [2].

1.2. Acoustic cues

Perception and production studies have shown that phrase boundaries are signaled mainly by timing cues – duration of syllables and vowels increases significantly in the vicinity of phrase boundaries [3], [4], [5].

There is no agreement as regards correlation between boundary presence and strength on the one hand, and duration of the silent pause on the other. The significance of the silent pause in signaling phrase boundaries was shown in [6] where the accuracy of the automatic phrase boundary detection using only pause duration achieved 95.8%. The results of a perception study [4] prove that pause duration is an important cue to boundary strength, but in the prediction of upcoming boundaries listeners use it only in the absence of a distinct pre-boundary lengthening (cf. [5]). The analyses reported in [3] revealed that less than 50% of phrase boundaries are followed by a silent interval, which suggests that pause can not be systematically used as a cue to boundary presence.

Some studies point out the role of F0 cues [5], [7], [8] and voice quality [9] in signaling phrase boundaries. The results presented in these studies show that location and strength of upcoming phrase boundaries can be reliably judged without an access to lexical or syntactic information which indicates that acoustic cues are of primary importance in this task.

1.3. Automatic determination of phrasing

Most of the state-of-the-art approaches to automatic determination of phrasing rely on large vectors consisting of one or more types of features: acoustic, lexical and/or syntactic. Statistical modeling methods applied to automatic boundary detection and classification include neural networks, classification trees, maximum entropy models or HMMs. The complexity of the models depends on speaking style, the number of categories that need to be recognized and the number of tasks performed by the model.

As regards the best state-of-the-art models they achieve an overall accuracy well above 90% [6], [10], [11]. The performance of the best models is thus comparable to inter-transcriber agreement for phrase boundary position identification [3], [12]. The same studies report on lower consistency (i.e. below 90%) in manual boundary type annotation. In automatic phrase boundary type classification accuracy rates are also about 5-10% lower than in detection of boundary position [13], [14].

2. Methodology

2.1. Speech material

The speech corpus used in the current study was built for the Polish module of BOSS (Bonn Open Source Synthesis) - a unit selection speech synthesis system. The corpus contains recordings of a professional male speaker (approx. 4 hours) reading phonetically rich and balanced sentences, fragments of fiction (including dialogues and examples of expressive speech) and reportage. From this corpus a subset consisting of 1052 utterances (15566 syllables including 1880 phrase-final syllables) representative of the whole speech material included in the corpus was selected.

The speech material was annotated at the segmental and suprasegmental level. Transcription and segmentation at the phone, syllable and word level was obtained automatically. At the word level verification of the automatically inserted stress markers was carried out using a large pronunciation lexicon.

An inventory of five pitch accents and five phrase boundary types was defined. Unlike in the ToBI [12] intonation labeling framework, the description of boundary phenomena used in the current study is based on and provides two types of information: 1) the strength of the phrase break – where ***2*** marks minor phrase boundary and ***5*** major phrase boundary, and 2) the type of the distinctive pitch movement occurring at the phrase end – where dot indicates falling and a question mark rising boundary.

2.2. Phrase boundaries types

The inventory of phrase boundary types consists of three falling and two rising boundaries. One rising and one falling boundary type are associated with minor intonation phrase boundaries, and one rising and two falling boundaries – with major intonation phrase boundaries.

2,? and ***5,?*** are rising boundaries realized by a rising pitch movement from a lower target level on the penultimate or ultimate syllable in the phrase to a higher F0 target associated with the phrase boundary. The two boundaries differ in the amplitude of the rise at the phrase edge (which is greater in case of ***5,?***) and scaling of the F0 targets (the rise starts higher and ends lower in speaker's range in case of ***2,?***). ***5,?*** is associated with major phrase boundaries, whereas ***2,?*** – with minor phrase boundaries.

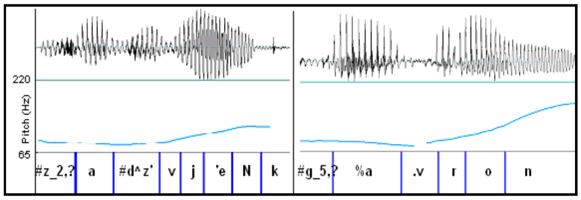

Figure 1: *Rising boundaries:* ***2,?*** *(left) and* ***5,?*** *(right).*

2,. 5,. and ***5,!*** are falling boundaries realized by a falling pitch movement from a higher target level on the penultimate or ultimate syllable in the phrase to a lower F0 target associated with the phrase edge. The three boundary types differ in the amplitude of the fall (which is the greatest in case of ***5,!*** and the smallest in case of **5,.**) and scaling of the F0 targets (in case of **5,.** they are positioned significantly lower in speaker's range in comparison to **2,.** and ***5,!***). **5,.** and ***5,!*** are associated with major phrase boundaries, whereas **2,.** – with minor phrase boundaries. In the current study ***5,!*** boundaries were not taken into account, because they were underrepresented in the speech corpus.

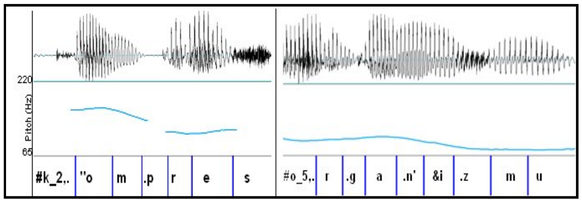

Figure 2: *Falling boundaries:* **2,.** *(left) and* **5,.** *(right).*

2.3. Feature extraction

For the analysis of the acoustic-phonetic realization of phrase boundaries for each syllable and its vocalic nucleus a number of features describing variation in F0 and duration was automatically extracted with a Praat script. The features included: 1) F0 value at syllable/vowel start/end and in the middle of the syllable/vowel, 2) maximum, minimum, mean F0 and standard deviation from the mean F0, 3) amplitude, steepness and duration of the rising and falling pitch movement, 4) Tilt model parameters [15], 5) slope parameter describing the amount of pitch variation, 6) absolute position (in ms) of F0 maximum and minimum, 7) syllable and vowel start, end time and duration. From these parameters further features were derived and normalized with respect to the overall F0 level over the length of the phrase (in case of F0 parameters) or expected duration determined for a given vowel or syllable type (in case of duration parameters, see [16]). For each syllable and its vocalic nucleus features of the two previous and two next syllables/vowels were provided as well. The resulting inventory consists of acoustic features which are commonly used in the analysis of intonation.

3. Production study

3.1. Feature selection

In a series of ANOVA and discriminant function analyses the effect of a major/minor phrase boundary presence/absence on variation in the acoustic features extracted from utterance's F0 and timing cues was investigated. The objective of the analyses was to identify features that can be regarded as the best acoustic cues signaling intonation phrase boundaries and distinguishing among boundaries of a different type. The analyses were performed at the word level i.e. in the investigation of cues to boundary position only word-final syllables (6844) were taken into account and the acoustic-phonetic realization of different phrase boundaries types was based on the subset of pre-boundary syllables (1880).

3.2. Time domain

ANOVA results showed that in the time domain phrase boundaries are signaled most of all by an increased duration of the pre-boundary syllable (+b=1.34 vs. –b=0.9, mean values, F=504.01), increased duration of the vowel of the word-penultimate syllable (+b=1.23 vs. –b=0.9, F=399.26) and to lesser extent – duration of the vowel of the pre-boundary syllable (+b=1.47 vs. –b=0.9, F=23.64). The effect of phrase boundary presence on variation in the duration features is statistically significant (p<0.01).

As regards distinction among boundaries of a different type (2,? 2,. 5,? 5,.) timing cues play no significant role.

3.3. F0 domain

3.3.1. Cues to boundary presence

The results of ANOVA analyses prove the significance of F0 features in signaling intonation phrase boundaries. The most important F0 features are:

- **tilt** (F=92.61) – feature describing the shape of the pitch contour on the vowel of the word-final syllable: value -1 indicates falling pitch movement, 1 indicates rising movement and 0 indicates the same amount of rise and fall [15]
- **F0mean** (F=81.07) – overall F0 level on the vowel
- **slope** (F=64.03) – feature expressing the amount of pitch variation on the vowel of the word-penultimate syllable
- **c1** (F=25.31) rising amplitude on the vowel

All these features are significantly affected by the presence of a phrase boundary (p<0.01). The figure below shows the effect of phrase boundary presence (+b) and absence (-b) on the shape of the pitch contour (*tilt,* left) and overall F0 level (*F0mean*, right) on vowels in the word-final syllables.

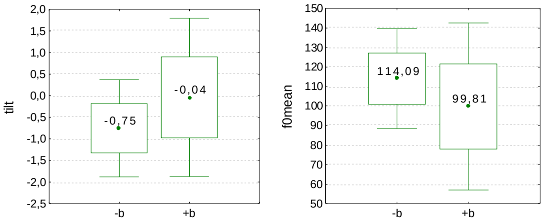

Figure 3: *Mean tilt and F0mean of vowels in pre-boundary (+b) and word-final syllables(-b).*

It was observed that vowels in syllables of a phrase-final position (+b) are characterized by significantly less falling pitch (indicated by higher average *tilt* value) than vowels in syllables which do not precede a boundary. This effect can be attributed to rising boundaries which signal continuation or interrogative mode. Vocalic nuclei of phrase-final syllables have significantly higher rising amplitude (c1 +b=6,13) compared to other vowels (-b=0,38, mean values). The former are also characterized by significantly lower F0mean, which can be attributed to falling boundaries. The role of F0 features in signaling phrase boundaries is also confirmed by significantly greater amount of pitch variation (slope) on vowels in +b than –b class (129,4 Hz/s vs. 88,12 Hz/s).

Similar effects to those reported here were found in [9]. It was shown there that variation in F0 (expressed in terms of parameters such as rhyme-final F0 level, F0 drop and slope) plays an important role in signaling prosodic boundaries. Moreover, in the absence of cues such as final lengthening and silent pause duration listeners rely on F0 features in prediction of boundary position and strength.

3.3.2. Cues to boundary type

In the ANOVA and discriminant function analyses it was found that phrase boundary type has the greatest effect on the following F0 features:

- **F0end** (F=914.15): F0 level at the end of the word-final syllable
- **F0mean** (F=566.68): overall F0 level on the nucleus of the word-penultimate syllable – this feature can be useful for the distinction among different boundary types, because it is correlated with other significant features, namely syllable and nucleus relative duration and syllable distance to the next pause (in ms)
- **direction** (F=341.77): describes the direction of the distinctive pitch movement at the edge of the phrase; it is calculated as a difference between overall F0 level on the vowel of the word-penultimate and word-final syllable

The effect of phrase boundary type on the variation in the F0 features listed above is statistically significant (p<0.01). Fig. 4 illustrates the distribution of F0end and F0mean values in the four classes of phrase boundaries. It can be seen that rising boundaries are characterized by higher F0end than falling boundaries and that minor phrase boundaries (2,. and 2,?) can be effectively distinguished from major phrase boundaries (5,. and 5,?) by higher F0mean.

As regards the *direction* feature it distinguishes not only rising from falling boundaries, but also weaker from stronger boundaries (2 vs. 5).

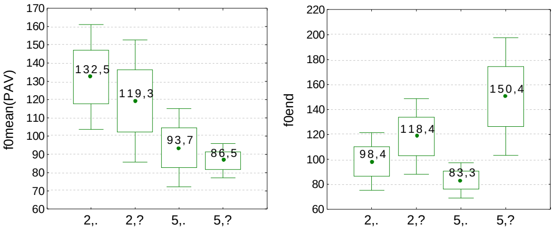

Figure 4: *Variation in the average values of F0mean and F0end depending on phrase boundary type.*

3.4. Other features

Instead of using silent pause duration to distinguish between minor and major intonation phrase boundaries we propose to use a feature describing the *distance* of the word-final syllable *to the next silent pause* measured in syllables.

As expected, it was found that minor phrase boundaries are significantly less frequently signaled by pauses than major boundaries – similar effects are reported in [3]. In our speech corpus the average distance of syllables of the final position is 9-10 syllables in minor intonation phrases and only 1 syllable in major intonation phrases. These differences are statistically significant (p<0.01). ANOVA (F=341.77) also indicates that *distance to the next pause* can be a very efficient feature in distinguishing between boundaries of a different strength.

4. Automatic determination of phrasing

4.1. Features

The framework of automatic phrasing proposed in this study can be summarized as follows:

- Statistical modeling techniques applied to automatic determination of phrasing include neural networks (multilayer perceptrones – MLP, radial basis function – RBF networks), discriminant function analysis (DFA) and decisions trees (DT). The models are designed semi-automatically using Statistica 6.0 software.
- Phrase boundary detection and classification are performed separately (i.e., by different models).
- The models rely on small feature vectors which provide a compact acoustic-phonetic description of the realization of boundary phenomena. Contrary to [11] where syntactic features are used as well, our feature vectors consist mostly of *acoustic* features (10 altogether) which can be easily derived from utterances' F0 and timing cues. The only lexical feature used in the current study is syllable's distance to the next pause (sec. 3.4).
- Determination of phrasing is performed at the word level.
- Contrary to [6], [10] in the task of detection of boundary position we account for both minor and major phrase boundaries.

4.2. Detection of phrase boundary position

All models (MLP, RBF, DFA and DT) designed in the current study perform much better than a chance-level detector which assigns the most frequent label (here –b) to all syllables. DFA yielded the highest average accuracy – 82.05% (in the cross-validation test). However, the best overall performance had the RBF network – apart from yielding high average accuracy (80.42%) it enabled correct identification of boundary position in 81.55% cases and of non-phrase-final syllables in 79.29% cases. The performance of all models is summarized in Table 1; the numbers in brackets (column *class*) show chance-level accuracies.

The results of sensitivity analysis carried out on the inputs to the best-performing network (RBF) are similar to those of the ANOVA analyses as regards the contribution of various F0 and duration features to the distinction between phrase-final vs. non-final syllables and vowels. The results prove that for the detection of phrase boundary position features reflecting variation in pitch (*tilt, slope*) and overall pitch level (*F0mean*) are as important as features describing variation in duration.

Table 1: *Detection of boundary position (test sample).*

class	MLP (4:14:1)	RBF (7:23:1)	DT (5 splits, 6 terminal nodes)	DFA
-b (72.59)	81.99	79.29	84.33	90.05
+b (27.14)	76.26	81.55	78.6	74.04
Average%	79.13	80.42	81.47	82.05

4.3. Classification of boundary type

In general, performance of models designed for classification of phrase boundary types is comparable to that of the boundary detection models. The average accuracy yielded by the former varies between 80.98% (DFA) and 87.61% (RBF network, 54 neurons in the hidden layer). The average recognition accuracy achieved with the classification tree (9 splits, 10 terminal nodes) and MLP network (20 neurons in the hidden layer) is above 84%. All the models perform much better than a chance-level boundary type classifier.

Weak boundaries (2,. and 2,?) were more difficult to recognize than strong boundaries (5,. and 5,?). The average recognition accuracy of the former did not exceed 85%, whereas the latter were recognized with at least 92% accuracy (results computed for the test sample). In the table below the best classification results (RBF network) are summarized.

Table 2: *Boundary type classification (test sample).*

class	2,. (20.42)	2,? (33.42)	5,. (37.93)	5,? (8.22)
accuracy %	70.13	84.92	98.6	96.7

5. Discussion and outlook

The main focus of the paper was to propose an efficient framework of automatic determination of phrasing based on a compact acoustic-phonetic representation consisting of two feature vectors. This representation can be easily derived from timing cues and utterance's F0. Seven F0 and duration features are used as an input to statistical models performing automatic detection of phrase boundary position and only four features (describing F0 variation and distance to the next pause) are used to automatically classify phrase boundary types (2,. 2,? 5,. 5,?). The results of sensitivity analyses (computed for the neural networks) and importance ranking of predictor variables (computed for the classification trees) confirmed the significance of the selected features to the detection and classification of phrase boundaries. Our results are in line with those of other perception and production studies [5], [8], [9] which indicate that both duration and F0 variation are important cues to phrasing, although boundary presence has greater effect on duration than on F0.

The performance of our phrase boundary detection models (average accuracy between 79 and 82%) compares favorably with [17] where the distinction between phrase-final vs. non-final syllables was accurate in 71% on average, and is similar to the performance of the models described in a recent study [5]. The fact that our models are less efficient than the best models presented in the literature (above 90% average accuracy) may be due to the fact that we take into account both minor and major phrase boundaries. As regards boundary type classification our models achieve average accuracy similar to that achieved by the best models [6], [10], [14].

Apart from high efficiency (high average accuracy in boundary detection and classification) the advantage of the framework proposed in this paper is the use of a compact acoustic-phonetic representation consisting of features which can be easily extracted from utterance's F0 and timing cues.

In the future it is planned to generalize this framework to other speakers and speaking styles, and to investigate the contribution of the selected acoustic features to the perception of phrase boundaries and recognition of boundary types by human listeners.

6. References

[1] Hirst, D. and di Cristo, A.: Intonation Systems: A Survey of. Twenty Languages. Cambridge University Press, Cambridge 1998

[2] Gallwitz, F., Niemann, H., Nöth, E. and Warnke, V.: Integrated recognition of words and prosodic phrase boundaries. Speech Comm. 36(1): 81-95 (2002)

[3] Yoon, T.J., Cole, J. and Hasegawa-Johnson, M.: On the edge: Acoustic cues to layered prosodic domains. Proc. 16th Int. Cong. Phon. Sci., Saarbruecken 2007, pp. 1017-1020

[4] Hsin-Yi, L. and Fon, J.: Perception of Temporal Cues at Discourse Boundaries. Proc. INTERSPEECH 2009, Brighton 2009

[5] Aguilar, L., Bonafonte, A., Campillo, F. and Escudero, D.: Determining Intonational Boundaries from the Acoustic Signal. Proc. INTERSPEECH 2009, Brighton 2009

[6] Bulyko, I. and Ostendorf, M.: Joint prosody prediction and unit selection for concatenative speech synthesis. Proc. ICASSP, Salt Lake City 2001, pp.781-784

[7] Carlson, R., Hirschberg, J. and Swerts, M.: Cues to upcoming Swedish prosodic boundaries: subjective judgment studies and acoustic correlates. Speech Comm. 46(3/4): 326-333 (2005)

[8] Kim, H., Yoon, T-J., Cole, J. and Hasegawa-Johnson, M.: Acoustic differentiation of L- and L-L% in Switchboard and Radio news speech. Proc. Speech Prosody, Dresden 2008, paper 214

[9] Carlson, R. and Swerts, M.: Perceptually based prediction of upcoming prosodic breaks in spontaneous Swedish speech materials. Proc. 15th Int. Congr. Phonet. Sci., Barcelona 2003, pp. 507-510

[10] Wightman, C.W., Syrdal, A., Stemmer, G., Conkie, A. and Beutnagel, M.: Perceptually Based Automatic Intonation labeling and Prosodically Enriched Unit Selection Improve Concatenative Text-To-Speech Synthesis. Proc. ICSLP, Beijing 2000, pp. 71-74

[11] Ananthakrishnan, S. and Narayanan, S.: Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence. IEEE Transactions on Audio, Speech, and Language Processing 16(1): 216-228 (2008)

[12] Pitrelli, J.F., Beckman, M.E. and Hirschberg, J.: Evaluation of prosody transcription labeling reliability in the ToBI framework. Proc. ICSLP, Yokohama 1994, pp.123-126

[13] Sridhar, R., Bangalore, V.K. and Narayanan, S.S.: Exploiting Acoustic and Syntactic Features for Automatic Intonation labeling in a Maximum Entropy Framework. IEEE transactions on audio, speech and language processing, 16(4): 797-811 (2007)

[14] Schweitzer, A. and Möbius, B.: Experiments on automatic prosodic labeling. Proc. INTERSPEECH 2009, Brighton 2009

[15] Taylor, P.: Analysis and synthesis of intonation using the tilt model. J. Acoust. Soc. Am 107(3): 1697-1714 (2000)

[16] Rapp, S.: Automatic labeling of German prosody. Proceedings of ICSLP, Sydney 1998, pp. 1267-1270

[17] Wightman, C.W. and Ostendorf, M.: Automatic Labeling of Prosodic Patterns. IEEE Transactions on Speech and Audio Processing, 4(2): 469-481 (1994)