An Automatic Hierarchical Multiple Level Phrase Segmentation Approach for Spontaneous Speech

András Beke[1], *György Szaszák*[2], *Viola Váradi*[3]

[1]Research Institute of Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
[2]Idiap Research Institute, Martigny, Switzerland
[3]Dept. of Phonetics, Eötvös Loránd University, Budapest, Hungary

beke.andras@nytud.mta.hu, gyorgy.szaszak@idiap.ch

Abstract

The present paper investigates automatic prosodic phrasing of spontaneous speech: a two-step segmentation technique is presented, based on unsupervised learning. In the first step, the Intonational Phrases (IP) are detected automatically based on speech energy, spectral centroid and a double-thresholding technique. In the second step, Phonological Phrases (PP) are identified within the IPs. As acoustic features, F0, overall energy and vowel duration are investigated. An adaptive thresholding method is used based on Kullback-Leibler divergence computed in an autocorrelative manner for the feature streams. For Hungarian spontaneous speech, a phrasing accuracy of over 80% can be reached when comparing to a hand-labelled reference phrasing. It is found that in Hungarian sponatenous speech, F0 and energy play an essential role in IP level phrasing, whereas PP level phrasing is most effective using F0 related features alone. Vowel durations are shown not to contribute to prosodic phrasing in Hungarian. Although the evaluation targets the Hungarian language, the applied method is universal and can be easily adapted for other languages.

Index Terms: phrasing, spontaneous speech, hierarchical

1. Introduction

Prosodic phrasing and/or prosodic boundary detection is an important research topic and several phrasing or boundary detection approaches have been developed and analysed for read and slightly spontaneous speech (such as semi-formal speech used in information retrieval systems) [1], [2]. When using a supervised approach, machine learning can be applied on labelled data, which will end in producing a classifier or detector capable of predicting prosodic boundaries based on the associated acoustic-prosodic features. This approach also implies that the entities to be classified or detected (boundary or break and their types) are a priori known and annotated in the training corpus. More recently, unsupervised modelling of prosody [3] or adaptation of seed prosody models in an unsupervised manner has also received attention [4]. Such approaches are of primary interest when dealing with spontaneous speech, as due to its extreme variabilty (disfluencies, atypical or non-canonical realizations in terms of acoustic correlates), prosodic entities worth modelling can be problematic to be identified or clustered even by human experts.

This paper focuses on exploring the prosodic structure of spontaneous speech, work is done on a Hungarian spontaneous speech database. Earlier efforts for prosodic event detection and automatic phrasing in Hungarian read speech showed that based partly on the fixed stress of Hungarian, robust stress detection was possible [5] and a phonological phrase alignment approach was proposed [6], where modelling of F0 and energy contours of phonological phrases was used in a supervised machine learning approach to perform automatic phonological phrase alignment and based on this, a partial, but powerful recovery of the prosodic and, to a lesser extent, of the syntactic structure in read speech. However, not surprisingly, when trying to adapt this automatic approach for spontaneous speech, recall rates fall by approx. 20-30% from around 80%. An effort to try to identify and cluster characteristic prosodic entities or phrase types in Hungarian spontaneous speech by using an unsupervised approach lead only to partial success [7].

However, supposing a hierarchical structure of prosody as described by Selkirk [8], the automatic phrasing implemented for read speech in [6] was able to yield a reliable (accuracies close to 80%) phrasing down to the phonological phrase level, and also to separate intonational phrase level from the underlying phonological phrase level. This approach required clustering of a number of phonological phrase prototypes, which were then modelled by HMM/GMM based on acoustic-prosodic features. As already mentioned, clustering of such characteristic prototypes in sponatenous speech failed. However, our hypothesis is that upon the intonational phrase level, also the phonological phrase level should be identifiable based on prosody in spontaneous speech as well. Moreover, cues for the perception of phonological phrasing are supposed to be simple enough, in order to allow the listener to concentrate also on other modalities (attitudes, emotions, also dialogue management functions) of the complex and rich information transmitted during a spontaneous conversation. Therefore, in this paper an attempt is made to detect intonational and phonological phrases in spontaneous speech with a two-step method, able to separate these two levels according to the prosodic hierarchy.

This paper is organized as follows: First, the spontaneous database is presented, than the intonational and phonological phrase segmentation approaches are described. Phonological phrase detection is implemented and evaluated in the subsequent section, and finally conclusions are drawn.

2. Data and methods

The BEA (BEszélt nyelvi Adatbázis: spoken language database [9]) spontaneous speech database was used in this research. BEA is a multi-purpose database of Hungarian spontaneous speech. 8 spontaneous narratives were selected (4 male and 4 female) from the database. The subcorpus was manually annotated by two different phoneticians, although the annotation will be exclusively used as reference for evaluation. The

10.21437/SpeechProsody.2014-168

annotation contained three levels: Intonational Phrases (IP), Phonological Phrases (PP) and also involved a word level transcription. The intonational phrase can be thought of being a part of speech forming a unity in terms of stress and intonation contour, and is found often between two pauses. The IPs can be further divided into phonological phrases based on intonation and stress pattern. A PP is a unity characterized by its own stress and intonation contour, but this latter can be unterminated (continued in next PP). The corpus contained 398 IPs and 751 PPs in total from the 8 speakers.

2.1. Intonational Phrase segmentation

In the spontaneous speech of the BEA database, turns can usually be further segmented into separate utterance units. However, there is no consensus on how to define an utterance unit [10]. The manner in which speakers segment their speech into intonational phrases undoubtedly plays a major role in its definition. Intonational phrase endings can be signalled through variations in the pitch contour, segmental lengthening and pauses. Cruttenden [11] in his theory uses external criteria for identifying intonation groups defined by *potential* boundaries. One of them is a potential pause following an intonation group, however, pauses are not obligatory boundary markers and may occur within a group. According to Shriberg and his colleagues [12] important cues to boundaries between semantic units, such as sentences or topics, are breaks in prosodic continuity, including pauses. In their system for sentence segmentation, the pause model used by the recognizer was trained as an individual phone. The result showed that this pause model was among the ones with the highest influence in sentence segmentation task. Pauses are without a doubt the most expressive instruments for marking of strong boundaries [13]. For Hungarian language, Gósy's research results showed [14] that pauses were among the most important features for utterance segmentation in human perception. In our database, in most cases the IP boundaries are bounded to pauses. A number of filled pauses and unfinished words were perceived as separate IPs as well by the annotators. These findings suggest using two features: a feature which plays an important role in speech detection such as speech energy or speech centroid; and a second feature which is sensitive to speech intonation such as fundamental frequency. In IP segmentation, the task is not only to find silent regions in continuous speech, but to detect the strong intonation changes in the acoustic signal. There are many solutions to detect silence and speech in the audio signal. In this research we choose a very simple and fast algorithm to segment pauses, created by Giannakopoulos [15] and implemented in MATLAB. In this system, a fundamental frequency estimation algorithm is also implemented.

This algorithm first extracts three features: signal energy, spectral centroid and fundamental frequency. These features are extracted for every frame using a 50 ms long window. The signal energy is higher in case of a speech segment than a silent segment. The spectral centroid is a spectral position and this expresses which frequency region contains the most part of intensity in the spectrum. A higher value of the spectral centroid usually corresponds to a speech segment in the audio signal. F0 can refer to the speech signal, and provides a good representation of intonation movements. On both features, 5 point median filtering is applied to smooth the signal. After the feature extraction, a threshold is calculated to each feature stream. In order to determine the threshold, first the smoothed histogram of the feature is used. The following step is to find local maxima of the histogram. In this step we deviate from the original method. To find the two most frequent values in the histogram we use k-means unsupervised learning algorithm. K-means clustering [16] is one of the simplest and oldest unsupervised learning algorithms. Given a set of data (consisting of n different, d-dimensional observations) and the desired number of clusters (k), this algorithm clusters iteratively the data around the so called centroids. For bootstrapping, k data can be randomly chosen as centroids, then each observation is clustered to the nearest centroid. The nearest centroid is computed using some distance measure, such as the Euclidean distance or sum of squares, for example. Centroids are iteratively updated to the mean of the belonging observations, until a specified level of convergence is reached. The main drawback of the algorithm is that the number of clusters (k) has to be determined prior to the clustering itself. In this process we used two clusters: pause and speech.

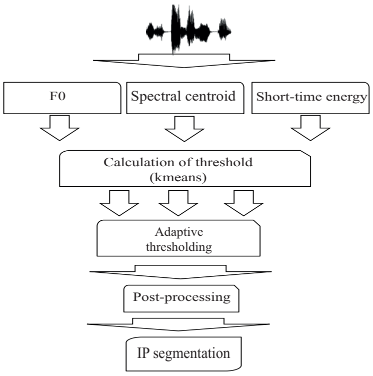

Figure 1: Block diagram of IPs detection

The clustering problem can be formulated as minimizing:

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x_j \in C_i} D(x_j, \mu_i) \tag{1}$$

for clusters $C_1, C_2, ...C_k$, given n observations ($k < n$), using a $D(.)$ distance function to evaluate the distance between centroid means (μ_i) and observations (x_j). Since this problem is NP hard, usually heuristic approximation is used to solve the problem. If the two cluster centers are available, we calculate the threshold using the following equation:

$$T = \frac{W * M_1 + M_2}{W + 1} \tag{2}$$

where W is a user-defined parameter (set to $W = 0.5$ in our case). Once the threshold is calculated, we apply it to the feature streams. The last step is the post-processing step. In this step the overlapping segments are merged using a large window (usually of a length of about $250ms$, see Figure 1).

2.2. Phonological phrase segmentation

Segmentation for PPs is a harder task than detecting IPs in the continuous audio signal. As PPs are embedded into IPs, the output of the segmentation for IPs constitutes the input of the segmentation for PPs (Figure 2), leading to a two-level, hierarchical approach.

As PPs can be described as phrases with own (proper) intonational and stress patterns, we used fundamental frequency

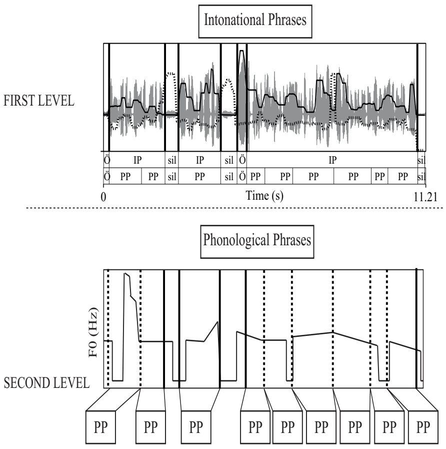

Figure 2: Hierarchical PPs detection

(F0), mid-term speech energy and vowel's duration for their detection. PP boundary detection is carried out as described in the next subsection.

2.2.1. Feature extraction

We use the three basic acoustic correlates of prosody as features: fundamental frequency, energy and duration (tempo).

Fundamental frequency (F_0) is extracted by ESPS method using a 25 ms long window, by a frame rate of 10 ms. The obtained F_0 contour is first filtered with an anti-octave jump tool. This is followed by a smoothing with a 5 point mean filter. F_0 is linearly interpolated in log domain. The interpolation is omitted for voiceless sections longer than 150 ms and also for F_0-rises higher than 110% after an unvoiced part.

Energy is extracted with a 150 ms window by 10 ms frame rate and then a further 5 point mean filtering is applied. As duration features, vowel lengths are used. In order to make the feature extraction automatic, an HMM-GMM based broad phoneme classifier is used. Broad phoneme classes cover vowels, nasal and approximant consonants, plosives, affricates and fricatives. The phoneme classifier uses standard MFCC features as input (with first and second order deltas) and produces a phoneme class level alignment at its output. Resulting vowel length is normalized per speaker, made continuous by a 10 ms frame rate and smoothed in order to obtain a vowel duration contour, called tempo feature. The reason for using a broad phoneme classifier instead of an ASR is twofold: we would like to keep the system generalizable to untranscribed, highly spontaneous speech with no proper language model coverage and hence use a phoneme-class loop grammar, and the phoneme classifier is more accurate in the required pure phoneme-class loop recognition task.

Although the phoneme classifier itself works by some incertaincy, confusions between classes are not crucial if they are systematic, whereby they can even reflect important prosodic information (such as in case of utterance final positions, where vowels are often confused due to low energy and irregularity). For each feature we calculate the first and second order deltas as differential and acceleration coefficients. The dimensionality of the feature vector is hence 9. After the feature extraction we normalise features to 1.

2.2.2. Segmentation using symmetrical Kullback-Leibler distance

The Kullback-Leibler (KL) distance is one of the most commonly used algorithms to measure the dissimilarity between two distributions [17]. The KL distance has been used for various tasks like speaker diarisation, speaker recognition, speech recognition, voice activity detection, etc. The KL distance can be used in speech segmentation and music segmentation as well. In this study, we apply the KL distance to detect phonological phrase boundaries in spontaneous speech. Matthew et al. [18] showed that the symmetric Kullback–Leibler distance is an effective distance metric to facilitate the detection of long-term statistical differences in speech signals. The mathematical background is as follows: let us assume X and Y are two random distributions, and KL is the dissimilarity between these two distributions. The distance $KL(X;Y)$ between X and Y can be calculated as:

$$KL(X;Y) = E_X(log(\frac{P_X}{P_Y}), \tag{3}$$

where E_X stands for the expected value of the probability density function of X. If distributions are modelled by Gaussians, the above equation becomes:

$$\begin{aligned} KL(X;Y) = \frac{1}{2}tr[(\Sigma_X - \Sigma_Y)(\Sigma_Y^{-}1 - \Sigma_X - 1)] + \\ + \frac{1}{2}tr[(\Sigma_Y^{-}1 - \Sigma_X^{-}1)(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T], \end{aligned} \tag{4}$$

where Σ refers to covariance matrices and μ to mean vectors of the respective distributions.

As this function is asymmetric we can symmetrise it with the following method:

$$KL2(X;Y) = KL(X;Y) + KL(Y;X) \tag{5}$$

As said before, if both distributions are considered to be Gaussian, a closed form solution exists for the $KL2$ symmetric KL distance.

In this work, the $KL2$ distance was calculated between two consecutive parts of the signal, corresponding to 4 frames (40 ms) length each. The window step was 1 frame (10 ms). The following task is to find the peaks in $KL2$ value curve. In $KL2$ value a high distance value indicates a possible acoustic change, whereas a low value indicates that the two compared regions of the signal are acoustically similar. From the point of view of the peak detection, it is very important to choose the right window length. For this reason, we used various window sizes from 25 frame (250 ms) up to 70 frame (700 ms) on the $KL2$ curve.

The second problem is the threshold estimation. To detect the changing point, we use two adapted thresholds (thr_A and thr_B). The first is computed as the mean of a window around the given point, multiplied by a constant:

$$thr_A = \alpha \frac{1}{2N_1} \sum(F). \tag{6}$$

where F is the feature vector, N_1 is the length of windows, and α is a constant.

However, in order to be detected as PP boundary, the given value must also be greater than thr_B, which is calculated by:

$$thr_B = \sigma_F + \beta \frac{1}{2N_1} \sum(F) \tag{7}$$

where σ_F is the standard deviation over the windowed area, β is the size of the window. The first threshold ensures that the given value is greater than the surrounding area, calculated over a small window. The second threshold is calculated over a larger window, and ensures that the change takes into account the general trend of the data changes. The window sizes are currently set to 3 and 4 seconds respectively. Use of these thresholds enables us to reduce false positive rate, and to return only the highest value at any possible PP boundary.

3. Evaluation

The techniques described in this section allow us to measure the performance of any segmentation algorithm. For evaluation method we use Brandts GLR method [19]. This method proposes three common measures which show the performance of the automatic segmentation. The first is the insertion (Ins), which means that there an extra boundary (event) in automatic segmentation to the reference segmentation. Omission (Oms) value means that there are left boundaries (missed events) in the automatic segmentation compared to the reference segmentation. The accuracy (Acc) is calculated using the number of correctly matched boundaries ($Corr$) – if the distance between the automatic label and manual label is within a pre-set tolerance –, the insertion value and the omission value:

$$Acc = \frac{Corr - (Ins + Oms)}{All} \qquad (8)$$

The accuracy can measure the performance of the segmentation algorithm. As reference, the PP hand labelling is used. The result depends on the tolerance value, therefore we tried various tolerance values between 25 ms and 100 ms.

4. Results

First, the automatic IP boundary segmentation algorithm is evaluated. Based on speech energy, speech centroid and F0, accuracy of IP boundary detection was 83.1% in spontaneous speech. Leaving one of the features out considerably lowered performance. Most of the errors are caused by the IPs starting or ending by filled pauses. The second aim was to test the automatic PP boundary segmentation algorithm. In our research we tried several features an their combination to detect the PP boundary as well. We compared results obtained from five combinations of the three features. In our first experiment, we focused on the impact of the window size (used for similarity measure calculation in KL2) on PP boundary detection. Window length for the KL2 similarity measure ranges between 100 ms and 400 ms. The result shows the accuracy of the segmentation depending on window length for KL2 and the type of features as well. The best result is obtained by using F0 alone and a 400 ms long window for KL2 (Table 1).

Table 1: *Accuracy of PP segmentation depending on window length.*

Windows length (ms)	100	200	300	400
F0	68.71	75.83	76.33	**80.18**
Tempo	55.33	56.67	56.91	57.38
F0+energy	68.45	74.75	74.25	79.04
F0+tempo	68.79	73.15	72.33	78.22
F0+energy+tempo	68.86	72.60	71.84	77.05

We especially focused on speech tempo in PP boundary detection, as it was found unhelpful for read Hungarian [5]. When we use only the speech tempo for IP boundary detection the accuracy is very poor. However, the accuracy improves when the speech tempo is combined with F0, but it is still lower than using only F0. The accuracy also decreases if speech tempo is combined with F0 and energy. The result shows that the accuracy increases if speech tempo is discarded. Speech tempo does not correlate well with F0 (R=-0.06) or energy (R=-0.04). Based on these results, its contribution to the perception of PPs is likely to be negligible in Hungarian.

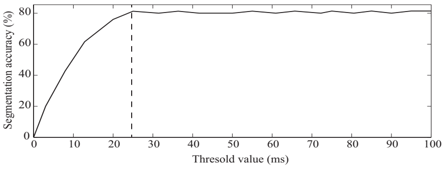

Figure 3: Accuracy of segmentation depending on threshold value

Next, the performance of the segmentation (accuracy) is shown, depending on the tolerance value. We tested our segmentation algorithm using various tolerance values (Figure 3). The result showed that if the tolerance is 25 ms the segmentation accuracy is 80.2%. By further augmenting the tolerance, there is no additional gain, accuracy saturates. This means that the segmentation algorithm is quite precise in time.

These results also suggest that F0 is the basic acoustic cue in PP boundary perception, whereas energy is also important in the perception of upper level unit boundaries is spontaneous Hungarian. The relatively powerful detectability of such boundaries in spontaneous speech strengthens the hypothesis that this kind of segmentation plays a key role in human perception as well.

5. Conclusions

The aim of this research was to segment spontaneous speech based on an unsupervised learning technique and a sophisticated peak detection approach. Phonological phrase segmentation was implemented in two hierarchical steps: first intonational phrase segmentation was performed using k-means clustering. Thereafter, intonational phrases were further segmented for phonological phrases, based on prosodic event detection exploiting symmetric Kullback–Leibler distance in an autocorrelation like approach. KL-distance features were themselves used as derived features, and peak detection was carried out on these. In our research we tried several features an their combination to detect the PP boundary as well. The results showed that fundamental frequency can be clearly associated with the phonological phrase level, as the best PP segmentation result is yielded by using F0 features alone by allowing only 25 ms time deviation between the detected PP boundary and the reference one. In this case, the accuracy was 80.2%. Results showed that speech tempo had no identifiable role in PP boundary detection. Regarding energy-based features, they seem to be associated to the IP level and did not improve PP detection based on F0 alone. However, some lack of robustness is supposed for the energy cues despite using normalization, due to the high degree of channel variability in the corpora (for example the speech recorded with a single microphone in a two-party conversation). These results are comparable to results seen for read speech [5], especially by the low uncertainty of the PP boundary detection regarding accurate placement in time.

6. References

[1] N. M. Veilleux and M. Ostendorf, “Prosody/parse scoring and its application in atis,” in *Proceedings of the workshop on Human Language Technology*, ser. HLT ’93. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993, pp. 335–340.

[2] F. Gallwitz, H. Niemann, E. Nöth, and W. Warnke, “Integrated recognition of words and prosodic phrase boundaries,” *Speech Communication*, vol. 36, pp. 81–95, 2002.

[3] C. Chiang, S. Chen, H. Yu, and Y. Wang, “Unsupervised joint prosody labeling and modeling for mandarin speech,” *Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1164–1183, 2009.

[4] A. S. and N. S., “Automatic detection of disfluency boundaries in spontaneous speech of children using audiovisual information,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 138–149, 2009.

[5] K. Vicsi and G. Szaszák, “Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features,” *International Journal of Speech Technology*, vol. 8, no. 4, pp. 363–370, 2005.

[6] G. Szaszák and A. Beke, “Exploiting Prosody for Automatic Syntactic Phrase Boundary Detection in Speech,” *Journal of Language Modeling*, vol. 1, pp. 143–172, 2012.

[7] A. Beke and G. Szaszák, “Unsupervised clustering of prosodic patterns in spontaneous speech,” in *Lecture Notes in Computer Science*. Springer, 2012, pp. 648–655.

[8] E. Selkirk, “The syntax-phonology interface,” in *International Encyclopaedia of the Social and Behavioural Sciences*, N. Smelser and P. Baltes, Eds. Oxford: Pergamon, 2001, pp. 15 407–15 412.

[9] M. Gósy, “BEA - A multifunctional Hungarian spoken language database,” *PHONETICIAN 105-106*, 2012, pp. 50–61.

[10] P. A. H. Traum, David R., “Utterance units in spoken dialogue,” in *Dialogue Processing in Spoken Language Systems — ECAI-96 Workshop, Lecture Notes in Artificial Intelligence*, pp. 125–140, 1997.

[11] A. Cruttenden, *Intonation*. Cambridge University Press, 1997.

[12] D. H.-T. G. T. E. Shriberg, A. Stolcke, “Prosody-based automatic segmentation of speech into sentences and topics,” pp. 127–154, 2000.

[13] P. Hansson, *Prosodic phrasing in spontaneous Swedish*. *Travaux de l’institut de linguistique de lund 43*, Lund University, 2003.

[14] M. Gósy, “Vitrual sentencse in spontaneous speech,” in *Speechresearch 2003*. 2003, Budapest, Hungary, pp. 19-43.

[15] T. Giannakopoulos, “Study and application of acoustic information for the detection of harmful content, and fusion with visual information,” Ph.D. dissertation, Dpt of Informatics and Telecommunications,University of Athens, Greece, 2009.

[16] J. B. MacQueen, “Some Methods for classification and Analysis of Multivariate Observations,” in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, pp. 281–297.

[17] C. L. Boite Jean Marc, “Speaker tracking in broadcast audio material in the frame work of the THISL project,” in *Proceedings of the ESCA ETRW workshop Accessing Information in Spoken Audio*, pp. 84–89, 1999.

[18] B. R.-R. S. Matthew Siegler, Uday Jain, “Automatic segmentation, classification, and clustering of broadcast news audio,” in *Proceeding of DARPA Speech Recognition Workshop*,pp. 97–99, 1997.

[19] D. P. S. Jarifi and O. Rosec, “Brandts GLR method and refined HMM segmentation for tts synthesis application,” in *Proceeding of European Signal Processing Conference, EUSIPCO2005*, pp. 23–33, 2005.