Automatic Prosodic Segmentation

by Elena Khasanova and Siyana Pavlova

Introduction

Prosodic segmentation takes important part in speech recognition, in particular, in discourse comprehension and various information extraction tasks. Across languages, prosody helps to convey structural, semantic, and functional information. Prosodic features normally comprise pausing, changes in pitch range and amplitude, global pitch declination, melody and boundary tone distribution, and speaking rate variation. Multiple studies argue for significant improvement of information extraction methods if based on prosodic cues due to several reasons [Shriberg et al.: 2000]. First, prosodic features are unaffected by word identity. Second, many prosodic features (e.g. duration, intonation) are invariant to changes in channel characteristics. Third, prosodic feature extraction does not require much computational load, so it can be easily added to existing acoustic models and thus improve the performance of the system [ibid.].

The present project focuses on the correlation between prosodic and syntactic structure, in particular, automatic determination of the intonation group boundaries based on acoustic cues, such as silent pauses, duration, F0 movement and phonation type. Christodulides [2018] suggests that these are the most salient cues to prosodic boundary identification. They are also language specific, for instance, in French, since the main accent occurs on the last syllable of the prosodic unit, it signals the prosodic boundary. Mertens and Simon [2013] discuss the qualitative view on prosodic boundaries identification, which suggests boundary types associated with hierarchical prosodic units, such as phonological word or intonation phrase, and functionally defined boundary strengths, such as continuation and final prosodic boundaries. Orosanu and Jouvet [2015] provide 10 acoustic features that take part in the intonation group boundaries identification. Following the latter studies, our project targets continuation and final prosodic groups based on the 5 prosodic cues (listed in the Methodology section) extracted from a speech fragment in French. The aim of the project is to develop a program that automatically segments speech signal into Intonation Groups (organizational unit in prosody, which contains one or several words) according to the rules provided.

Methodology

The goal of our project is, given a text file with a specified format, to parse the file and collect the relevant information, split the provided signal into intonation groups according to a set of rules, and add this information to a .TextGrid file.

Our program uses empirical rules based on prosodic features. These are summarised in Table 1.

Feature	Definition	Condition	Signal	Explanation
VNDurNorm	the normalized duration of the last vowel	>150%	IG boundary	Stressed vowel should be longer than usual
VNF0Delta	the F0 delta (difference between the last and the first syllable) of the last vowel of a word	>= 8	IG boundary	F0 movement should be significant to indicate the IG
VNF0SlopeT2	the F0 slope (VNF0Slope) on the last syllable multiplied by squared VNDurNorm	>= 16000	IG boundary	rate of pitch change, rising or falling; direction and amplitude
VNF0Level	F0 of the last vowel	> 85% or < 15%	IG boundary, continuation pattern if high, final pattern if low	fundamental frequency increases - melody goes up, decreases - goes down; high or low
lastF0Level	F0 of the last vowel normalized by speaker	> 20 or < 20	continuation pattern if high, final pattern if low	normalization is important as it takes into account speaker variation

Table 1. Summary of prosodic features used in the project

The data containing acoustic parameters is presented in a text file. Information about each word, in order of occurrence in the original signal, is provided. Each line of the file contains a number of features (WrdLabel, WrdStart, WrdEnd, VNDurNorm, etc.) for a given word.

The architecture of our program is presented in Fig. 1.

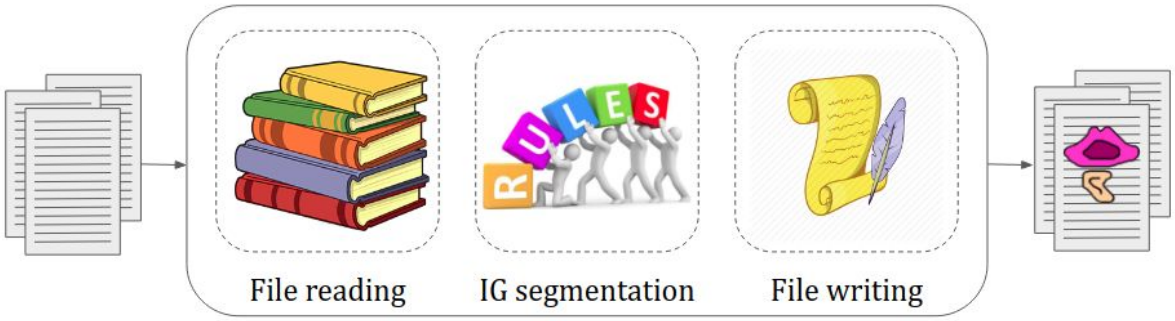

Fig. 1. Program architecture

For the first part, file parsing, we read the file, parse each line using regular expressions and record the features and their values for each word in a dictionary. The output of this part is a list of said dictionaries. While we need only a subset of these features in order to perform IG segmentation, we decided to record all features for each word. In this way, this module becomes reusable.

The second part, IG segmentation, is based on a set of empirical rules. Prosodic units are limited by the word boundaries.
For each word provided, we first apply the rules on VNDurNorm (>150%), VNF0Delta (>=8) and VNF0Slope2 (>=16000). If all of these are above the specified thresholds, we apply the two rules on VNF0Level and lastF0Level, which tell us whether the current word is an IG boundary, and if yes, what type - continuation pattern or final pattern. It should be mentioned that we take into account only the parameters of the last syllable since stress is always on the last syllable and IG boundary is unlikely to occur in the middle of the word.
Furthermore, we apply an additional set of rules to find the pauses between IGs. If the words surrounding a pause do not qualify for an IG boundary, the pause does not demarcate a new IG and is included in the next IG. The output of this part if a list of dictionaries, where each item is an interval. An interval has a beginning, an end and a label (IG_C for continuation pattern, IC_F for final pattern or # for a pause).

The final part, file writing, takes an existing .TextGrid file, copies its contents to a new file along with information about intonation groups, in a format readable by Praat.

Results

The final output of our program is a TextGrid file readable by Praat, where in addition to the phone segmentation and word segmentation provided in PreparedSpeech.TextGrid, we add the segmentation of the input into intonation groups as well as intervals to mark pauses.

Using the methodology described above, and the PreparedSpeech.proDataV3 file, which contained 340 words, we obtained 76 intervals of which 27 are pauses, 38 are continuation patterns and 11 are final patterns. This means that the average phrase length we have obtained is close to 7 and the ratio of continuation patterns to final patterns is roughly 4 to 1.

Conclusion

The quality of our segmentation can be evaluated by comparison with a reference file, linguistic intuition, or perception experiments. The reference file PreparedSpeechREF.TextGrid contains a different type of prosodic segmentation and takes into account more intonation groups and thus can not be completely considered as a standard. Our program is limited only by detection of the final and continuation patterns, while the reference file marks the prosodic type or each word. Linguistic intuition may suggest that different intonation groups are demarcated by the pauses, however, in our program pauses between the intonation groups are added after the detection of IG boundaries, therefore some intonation groups include pauses between the words. However, the type of intonation (continuation or final) seems to be captured accurately based on linguistic intuition of the authors of the project. Perception experiments have not been conducted within the span of our project.
In conclusion, given the size of input and the number of intonation groups as well as linguistic intuition, our segmentation seems sufficient, however, there is room for improvement. Our program is based on acoustic features of the last syllable of the words and word boundaries, no lexical cues are taken into account. Studies [Orosanu and Jouvet: 2015] suggest that the latter significantly improve segmentation. Also, more subtle intonation patterns may be considered to better represent the prosodic system of the language.

Literature

1. Christodulides, G. Acoustic Correlates of Prosodic Boundaries in French: A Review of Corpus Data. Rev.Estud.Ling., Belo Horizonte, v. 26, n. 4, p.1531-1549, 2018.
2. Mertens, P. Simon., A.C. Towards automatic detection of prosodic boundaries in spoken French. Proceedings of the Prosody-Discourse Interface Conference 2013 (IDP-2013).
3. Orosanu, L., Jouvet, D. Combining lexical and prosodic features for automatic detection of sentence modality in French. 3rd International Conference on Statistical Language and Speech Processing, SLSP 2015.
4. Shriberg, E., Stolcke, A., Hakkani-Tur, G., Tur, G. Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. To appear in Speech Communication 32(1-2) Special Issue on Accessing Information in Spoken Audio (September 2000).