The prosodic structure in French: Analysis with deep learning networks

Philippe Martin
LLF, UFRL, Université de Paris, France
https://doi.org/10.36505/ExLing-2021/0040/000513

Abstract

Experiments using supervised deep learning algorithms were executed on a large set of sentences selected from the SIWIS (Yamagishi et al., 2016) corpus in order to investigate properties of the prosodic structure in French. Based on a phonological model assuming dependency relations between accent phrases (Martin, 2018), the training corpus was annotated with a modified ToBI notation system, encoding prosodic events located on stressed syllables (vowels), i.e. pitch accents, and syntactic groups final syllables (vowels), i.e. boundary tones, which in French occur on the same accent phrases final syllable.

Keywords: Deep learning, French, ToBI, prosodic structure.

Introduction

Various deep learning models are well suited to classify, process and generate time series such as speech prosodic events, and may appear at first very attractive to obtain pertinent phonetic and phonological information in this field. However, the large number of parameters (typically over 1,000,000) obtained from deep learning training make extraction of meaningful phonetic and phonological information almost impossible. Therefore, instead of using raw acoustic data for training (e.g., WaveNet), feature engineering derived from specific knowledge in the domain could be used not only to speed up the learning iteration process, but also to evaluate the pertinence of the descriptive features extracted from raw acoustic data from the performance of the selected deep learning model.

Data

In order to obtain some prosodically meaningful results, sentences from a training corpus were annotated with as melodic contours as well as with a modified ToBI notation. These annotations pertain to prosodic events located on stressed syllables vowels, i.e. pitch accents, and syntactic groups final syllables vowels, i.e. boundary tones. Pitch accents and boundary tones in French occur on the same accent phrases (i.e. stress group) final vowel. Merged prosodic events are classified as rising or falling, above or below the glissando threshold, i.e. perceived as a melodic change or as a static tone. Sentence final and falling prosodic events before pauses constitute special categories.

The glissando value gives an indication pertaining to the actual perception of melodic changes. It is evaluated by the formula (st2-st1) / (t2-t1), with st1 and st2 being the stressed vowel starting and ending F0 frequency value in semitones at times t1 and t2. The glissando threshold, which approximately determines the limit for the perception of a change in pitch, lies in the 0,16 / $(t2-t1)^2$ and 0.32 / $(t2-t1)^2$ range (Rossi, 1971).

Rising and falling melodic changes above the glissando threshold are labelled respectively as H*H- and L*L- in the proposed modified F-TOBI system, and Cris and Cfal using contour notations. Rising or falling melodic changes below the glissando threshold are labelled H*/L* or Cneu. Sentence conclusive terminal prosodic events are annotated as L*L% or Cdec (declarative case) and H*H% or Cint (interrogative case). The falling contour before pause is labelled L*# or Cfap, Sentence post final declarative and interrogative prosodic events in a rheme-theme configuration use the H*/L* or C0n for the declarative case, and H*/H% for the interrogative configuration. These categories have been shown to adequately indicate the dependency relations between stress groups, which in turn define the sentence prosodic structure.

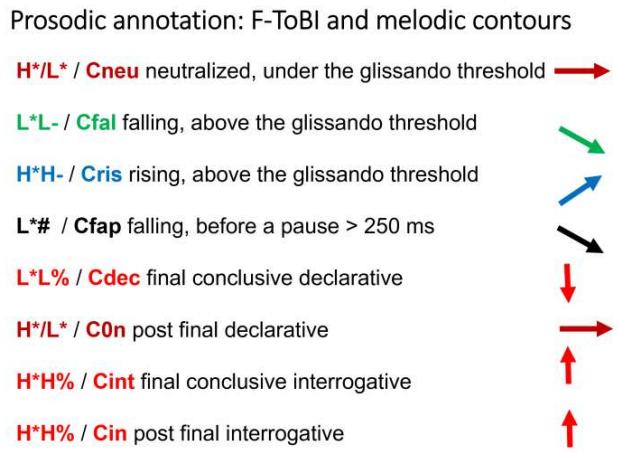

Figure 1. Table of prosodic events annotation symbols in both the modified F-ToBI and contour systems

Training and testing

The French SIWIS corpus (Yamagishi et al., 2016) contains 4477 relatively short sentences with various syntactic structures, read by 31 different speakers. Automatically selected stressed syllables candidates were validated perceptually.

The processing steps implemented in C++ and Python (using Keras and TensorFlow libraries) are as follows:

- Resampling of the 4477 recordings from 44100 Hz to 22050 Hz to better accommodate the automatic segmentation algorithm.
- Automatic segmentation into words and phones with API annotation, using a forced alignment algorithm comparing corpus sentences with TTS generated speech.
- Generation of the fundamental frequency curve for each sentence (spectral based algorithm).
- Data encoding with fundamental frequency values at the beginning and end of annotated stressed vowels, as well as narrow band spectrograms limited to 1300 Hz. These data implicitly include vowel and word duration, syllabic rate, intensity variation, etc.

Experimental results

For indirectly evaluate the efficiency of the selected descriptive features, two deep learning training and testing experiments were conducted, using two types of input data related to stressed vowels:

1. Fundamental frequency curve F0, using the spectral comb method
2. Partial narrow band Fourier spectrum, limited to1300 Hz.

3300 samples constitute the training set, encoded as images of 111 by 25 pixels. The training model has two hidden layers, with respectively 200 and 150 nodes. The Rectified Linear (i.e. x = max(0,x)) activation function was used for both. The output had 8 classes of prosodic events, as defined above (the two post-finals were excluded). The training algorithm used the Adam optimization algorithm, and operates on batches of 32 samples and an embedding of 100.

The training results pertaining to the two types of input data are:

For the fundamental frequency values obtained by the spectral comb method:

```
 1/46 [.............................] - ETA: 0s - loss: 0.4056 - accuracy: 0.85[redacted]
46/46 [==============================] - 0s 650us/step - loss: 0.4760 - accuracy: 0.8070
<tensorflow.python.keras.callbacks.History object at 0x000001DE67350AC0>
```

For the narrow band Fourier spectrum, limited to the0-1300 Hz range:

```
 1/26 [>.............................] - ETA: 0s - loss: 0.0107 - accuracy: 1.0000[redacted]
14/26 [===============>..............] - ETA: 0s - loss: 0.0244 - accuracy: 0.9[redacted]
26/26 [==============================] - 0s 4ms/step - loss: 0.0287 - accuracy: 0.9927
<tensorflow.python.keras.callbacks.History object at 0x000001DE5DB823A0>
```

The final accuracy, respectively 0.8070 and 0.9927, reflects the efficiency of feature selection to encode the data. A low accuracy indicates that some prosodic events belonging to different classes could not be separated, i.e. by using F0 alone, so that narrow band spectrograms appear more appropriate than direct fundamental frequency values for training. This could be explained by occasional errors made in F0 detection, whereas there is no implied calculation of F0 by using a narrow band spectrogram.

Two hundred prosodic events not part of the training set was processed for testing. It reveals that most errors, i.e. wrong classification of prosodic events, were linked to the glissando threshold used in annotation of the training set. For instance, events categorized as H*H- (or Cris) could be occasionally identified as H*/L* (or Cneu) and conversely. Indeed, the glissando threshold is a parametric approximation leading to erroneous classification when its value is close to the threshold between perceived melodic change and static tome perception of melodic events.

Conclusions

Machine learning is a powerful tool to obtain an efficient discrimination of classes of a very large number of objects. However, this classification power does not deliver any comprehension pertaining to the differences between objects belonging to different classes, whereas our intuitive knowledge allows us to establish theses classes and to classify any new object in one of these classes.

We may therefore consider using deep learning processes indirectly to get some insight pertaining to linguistic knowledge, as " machine knowledge" remains hard to interpret.

References

Martin, Ph. 2018. Intonation, structure prosodique et ondes cérébrales, London: ISTE, 322 p.

Rossi, M. 1971. Le seuil de glissando ou seuil de perception des variations tonales pour la parole. Phonetica. n° 23, 1-33.

Yamagishi, J. et al. 2016. The SIWIS French Speech Synthesis Database, University of Edinburgh. School of Informatics. The Centre for Speech Technology Research, https://doi.org/10.7488/ds/1705.