Deep learning and intonation in Text to Speech systems

Philippe Martin
LLF, UFRL, Université de Paris, France
https://doi.org/10.36505/ExLing-2019/10/0035/000397

Abstract

Although Recurrent Neural Networks deliver excellent results in Speech-to-Text and Text-to-Speech (TTS) applications, the generation of satisfactory synthetic sentence prosody remains one of the main causes of the quality differences between human and synthetic speech. These differences do not involve only emotions or attitudes, but also the prosodic structure which determines the way the listener processes the speech flow. This paper explores the theoretical and technical reasons for these difficulties and proposes a better feature engineering approach for deep learning based on an alternate model of sentence intonation, applied to French.

Key words: Deep learning, artificial intelligence, text-to-speech, prosodic structure

TTS by unit selection

Most recent text-to-speech systems are based on the concatenation of segments of various sizes of actual human speech recordings (normally produced by one speaker) stored in a database. As complete required sentences are rarely found in the database, typical operations proceed by assembling selected segments and adjusting the syllabic prosodic parameters duration, fundamental frequency (pitch), and intensity. Theses parameters have to be predicted from the available information such as class of syllables and type of syntactic groups of words (syntagms), among others.

As dominant phonological models of sentence intonation (e.g. the Autosegmental-Metrical model), may or do appear confusing for speech engineers, many approaches rely on statistical prediction of segment duration and fundamental frequency patterns. These methods generally do not integrate any linguistic process such as the prosodic structure, but they incorporate constrained prosodic characteristics of human prosody linked to emotions, attitude and socio-geographic attributes.

Deep learning: RNN, LSTM, BLSTM, DBLSTM

The advent of Recurrent Neural Network (RNN) and their variants such as Long Short-Term Memory (LSTM), Bidirectional BLSTM or Double bidirectional LSTM (DBLSTM) turn to be methods of choice used to generate appropriate sequences of prosodic parameters for TTS systems. They can be viewed as generalizations of statistical approaches implemented earlier in TTS

Systems. A deep learning implementation such as WaveNet (van den Oort et al. 2016) delivers synthesized speech of excellent quality very close to human speech. However, it generally fails to generate an appropriate prosodic structure from a given text, except for very short ones, where an acceptable pitch pattern is likely to be in the database.

Indeed, although the LSTM and BLSTM processes may capture time dependencies of some prosodic parameters, their practical implementation prevents them from apprehending long term dependencies characteristic of pitch movements from a raw wave signal. For instance, with a step size of 5 ms, 10 seconds of training speech would handle a memory of 2000 epochs, possibly involving a very large number of hidden layers. The number of connections to apply a gradient descent algorithm to determine connections between hidden layers could quickly become intractable. To cope with this problem, one has to turn to feature engineering, i.e. selection of features of the speech signal considered pertinent by prosodic experts.

Feature engineering

One of the most important functions of the prosodic structure is to allow the listener to proceed to the decoding and the understanding of the message conveyed by the speaker. This may be difficult to operate by the listener in real-time from the text only. Indeed, the average short-term auditory memory is limited to some 2 to 3 seconds for speech. Therefore, it is imperative to process incoming speech quickly and efficiently. For this purpose, melodic contours located on stressed syllables (actually on stressed vowels) facilitate the process by segmenting the incoming speech flow and grouping the resulting chunks (the stress groups) hierarchically before the actual identification of the text. Melodic contours function as dependency markers to indicate to the listener how to assemble the successive stress groups in the sequence (which defines the prosodic structure).

The prosodic structures displayed on the synthesized and original example illustrated below were obtained automatically by the WinPitch software (2019) from the sequence of annotated melodic contours, following a dependency model. In this model, based on a generalization of Delattre (1966) *continuation mineure* and *continuation majeure*, but enacted with the contrast of melodic slope principle (Martin, 1975), the minor continuation contour C2 ↘ indicates a dependency towards a major continuation contour C1↗ located further in the sentence ("to the right"), which in turn indicates a dependency towards a termination conclusive contour C0↓, also "to the right". A neutralized contour Cn→ marks a dependency towards either C2↘, C1↗ or C0↓ located further in the sentence. Besides, a neutralized termination contour C0n← indicates a dependency relation this time "to the left", i.e. towards a terminal conclusive contour C0 that precedes it. Details of the mechanisms defining the prosodic structure from a sequence of melodic contours can be found in Martin (2018).

Contrary to the original proposal by Delattre, melodic contours are not global, but are instead aligned on the vowel of (non-emphatic) stressed syllables. They are acoustically defined by their glissando values, above or below a threshold of change of pitch perception. Above this threshold, which depends on the rate of frequency change, the contour is perceived as a melodic change, below as a static tone (Rossi, 1971). C1↗ and C2↘ are respectively rising and falling contours above the glissando threshold, whereas Cn→ and C0n← are below the threshold. The terminal conclusive contour C0↓, falling in the declarative modality case and Ci↑ rising in the interrogative case, are acoustically defined as reaching respectively the lowest and the highest level of the pitch in the sentence.

An example

A read speech sentence from the SIWIS corpus (FR_A1_08_000 Yamagishi et al., 2017) has been selected as an example:
L'agriculture marocaine bénéficie d'un traitement privilégié pour ses exportations vers l'Europe "Moroccan agriculture benefits from privileged treatment for its exports to Europe".

API segmentation, melodic contours assignment and prosodic structure generation are automatically generated by WinPitch. The original realization read by a female speaker is shown Fig 1, whereas Fig. 2 displays the realization of a TTS system (Claude voice in Microsoft Windows).

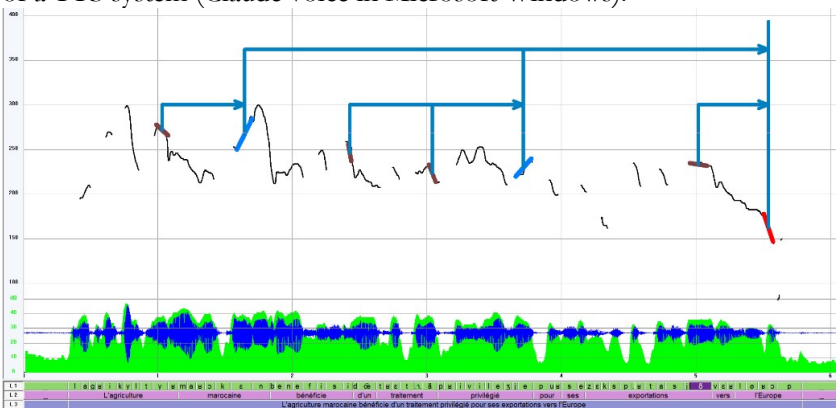

Figure 1. Original version of the sentence *L'agriculture marocaine bénéficie d'un traitement privilégié pour ses exportations vers l'Europe* read by a female speaker. The order of processing by the listener, as indicated by the prosodic structure, is [*L'agriculture Cn→ marocaine C1↗*], [*bénéficie Cn→ d'un traitement Cn→ privilégié C1↗*] and [*pour ses exportations Cn→ vers l'Europe C0↓*].

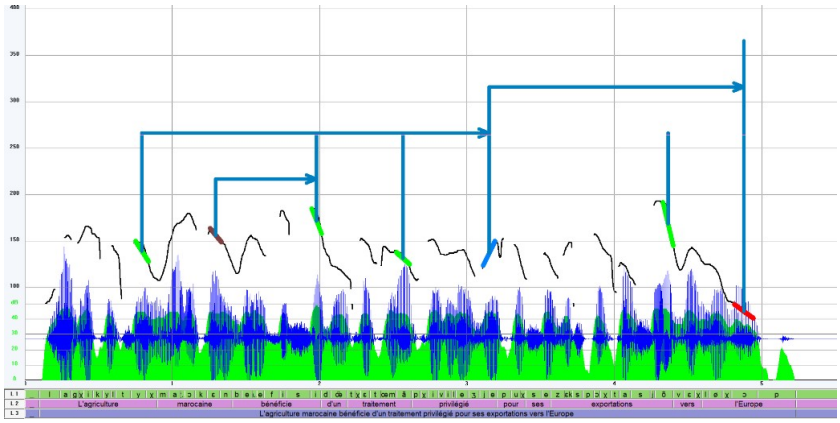

Figure 2. TTS version of the sentence *L'agriculture marocaine bénéficie d'un traitement privilégié pour ses exportations vers l'Europe* (Microsoft voice Claude). For this TTS realization, the order of processing by the listener, as indicated by the prosodic structure, is [*L'agriculture C2*↘ [*marocaine Cn*→ *bénéficie C2*↘] *d'un traitement C2*↘ *privilégié C1*↗] and [*pour ses exportations C2*↘ *vers l'Europe C0*↓]. The falling penultimate melodic contour C2↘ is a-grammatical, and should have been realized as a neutralized contour Cn→, below the glissando threshold.

References

Delattre, P. 1966. Les dix intonations de base du français, French Review 40, 1-14.

Martin, Ph. 1975. Analyse phonologique de la phrase française, Linguistics, 146, 35-68.

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 23, 1-33.

van den Oord, A. et al. 2016. A Generative Model for Raw Audio, https://arxiv.org/abs/1609.03499

Yamagishi, J., Honnet, P-E., Garner, Ph., Lazaridis, A. 2017. The SIWIS French Speech Synthesis Database, 2016 [dataset]. University of Edinburgh. School of Informatics. The Centre for Speech Technology Research. https://doi.org/10.7488/ds/1705.

WinPitch. 2019. www.winpitch.com