Interspeech 2018
2-6 September 2018, Hyderabad

Visualizing Punctuation Restoration in Speech Transcripts with Prosograph

Alp Öktem[1], *Mireia Farrús*[1], *Antonio Bonafonte*[2]

[1]Universitat Pompeu Fabra, Spain
[2]Universitat Politècnica de Catalunya, Spain
[1]{alp.oktem, mireia.farrus}@upf.edu
[2]antonio.bonafonte@upc.edu

Abstract

We have developed a neural architecture that tests the effect of lexical, morphosyntactic and prosodic features in restoring punctuation in speech transcriptions. Having outperformed a baseline model in terms of precision and recall, we further extend our performance tests by attaching it in a speech recognition pipeline. The visual and interactive testing environment that we prepared helps us observe how our models generalizes in unseen data and also plan our next steps for improvement.
Index Terms: prosody, punctuation, automatic speech recognition

1. Introduction

It is well-known that punctuation plays an important role in written language. Among its uses, we find the segmentation of discourse into comprehensible units through sentence structure, marking sentence modality (statement, interrogative etc.), the resolution of ambiguity, or the increase of readability through grammatical rules.

In applications that use *speech-to-text* (conversion of speech into textual form), correct punctuation placement remains as an issue outside of the automatic speech recognition (ASR) problem if: (1) humans will read the transcription, or (2) the output punctuation plays an important role in subsequent processing steps (e.g. machine translation, parsing). Also, we note that punctuating spoken language is different than merely punctuating its transcription. Prosody realization in speech, in many cases, influences position and type of the punctuation marks in the speech transcript. For example, pauses between words often signal a sentence boundary or a comma usage, rising intonation is a probable indicator of a question, or intensity and pitch reset is often observed at paragraph or sentence boundaries.

Based on these observations, Öktem et. al [1] developed and studied a neural network based architecture for building punctuation models on speech transcripts using lexical and prosodic features. Various models that combine word-aligned, prosodic and morphosyntactic features are tested. It is shown that certain prosodic features help to improve accuracy of punctuation generation in transcribed speech and also each punctuation mark can show different behaviors with respect to the features used.

In this paper, we present a software interface that we developed to demonstrate and test the lexical-prosodic punctuation models presented in [1]. By using a interface based on *Prosograph* [2], we are able to visualize prosodic movements in speech recordings and emulate interactively how our models would perform on real-world data. Input can be given to the interface either by recording with microphone or as a prerecorded file which is then sent to a state-of-the-art ASR system for transcription. The transcriptions, punctuated with our models, are displayed together with their graphical prosodic visualizations. The interface makes it possible to configure which of the lexical and prosodic features will be used for both punctuating and visualizing on screen.

2. Overview

The pipeline of the testing interface can be summarized as follows: (1) Obtaining a recording from either microphone or a waveform audio file, (2) Transcription using speech-to-text software, (3) Prosodic and syntactic feature extraction, (4) Punctuation restoration, (5) Visualization of punctuated versions of transcript together with acoustic measurements.

2.1. Voice input

A punctuation restoration module can be useful mainly in two uses-cases: (1) Transcription of simultaneous speech (e.g. for real-time subtitling) or (2) Transcription of prerecorded speech. For this, we prepared the testing interface so that it is able to record speech from microphone or take speech audio files as input.

2.2. Feature extraction

The trained lexical-prosodic punctuation models work with word-aligned acoustic features. The features are obtained by taking raw pitch/intensity measurements using Praat [3] software and then converting them to semitones according to the speaker's mean values. This is to ensure models learn on prosodic variances instead of absolute values.

2.3. Visualization

Same features used by the punctuation modules are used by Prosograph for visualize the acoustic characteristics of the input speech. By default, it displays the words in spoken segments and also silenced intervals in a transcript. Additionally, it provides controls for selecting which word-level prosodic features to display. These features include: mean f0, mean intensity, range f0, range intensity, f0 contour and intensity contour.

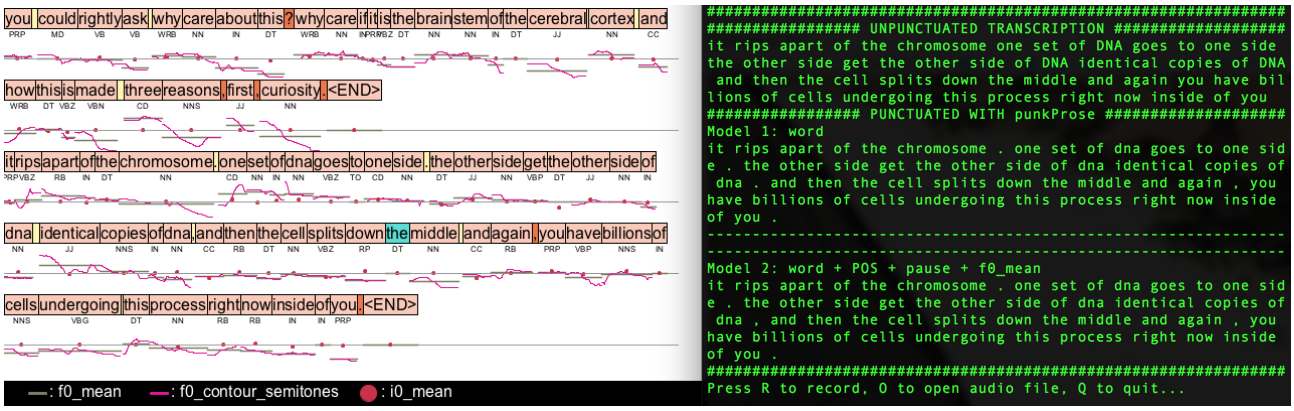
3. Software Framework

We have employed the following software components for our demo software:

Google Cloud Speech-to-Text for transcribing speech input from microphone. This service proved to be the best working one among other alternatives.

Montreal Forced Aligner [4] for obtaining word timings in the audio with respect to the transcription. This is essential for aligning the prosodic features with words.

Figure 1: *The two window interactive test environment. Recordings are presented through the command line interface (right) and visualized directly on Prosograph (left).*



Praat [3] for extracting raw prosodic features during each word interval. Intensity is extracted in *dB* and fundamental frequency in *Hz*.

Proscript Python package [5] for data storage and manipulation. This library makes it possible to extract and normalize acoustic-prosodic features and also store them.

punkProse [6] for punctuating the prosodically annotated speech transcripts. The *punctuator* script takes unpunctuated input in *Proscript* format and generates punctuation using models trained on TED talks. For more information on this work see [1].

Prosograph [7] for visualization and audio playback. This software makes it possible to visualize speech transcripts together with their prosodic features in a simple and clear interface [2]. Also, represented speech recordings can be played through its interface thanks to recent developments on the software.

3.1. Usage

The usage of the demo software is facilitated through two windows as seen in Figure 1: (1) Command-line interface for recording speech or opening an already recorded audio sample and (2) Prosograph for visualizing and playback of the speech input. The command-line interface carries out the process of getting the speech input, transcribing it and then outputting the punctuate transcript into appropriate files for Prosograph to read. After each recording, Prosograph needs to be refreshed to visualize the most recent set of recordings.

Also through Prosograph, we are able to select an interval of words and listen to the corresponding portion of the recording. This is to facilitate a careful study of the prosody of the recordings.

Punctuation generation results of different models can be viewed from both panels. Command-line interface prints the punctuated results of the most recent input. Visualizations can be switched to punctuated versions by pressing numerical keys of the keyboard where each key loads results of different models.

Each new recording is listed at the end of the Prosograph display to facilitate comparison. For example, as a test case, two utterances of the same sentence can be realized with different pausing or tone variations to see how punctuation varies.

4. Discussions and Conclusions

We have developed a visual and interactive testing environment to see how our lexical-prosodic punctuation models perform with an ASR system. Through this, we have the possibility to directly see the effects of prosodic realizations to punctuation placement in speech transcripts and evaluate our models. By demonstrating our work to the scientific community, we hope to build ideas to develop our methodology and find possible applications for it.

5. Acknowledgements

The first author has received Maria de Maeztu Reproducibility Award from Department of Information and Communication Technologies of Universitat Pompeu Fabra in 2018 through presentation of this work. The second author is funded by the Spanish Ministry of Economy, Industry and Competitiveness through the *Ramón y Cajal* program.

6. References

[1] A. Öktem, M. Farrús, and L. Wanner, "Attentional parallel RNNs for generating punctuation in transcribed speech," in *5th International Conference on Statistical Language and Speech Processing (SLSP)*, Le Mans, France, 2017.

[2] A. Öktem, M. Farrús, and L. Wanner, "Prosograph: A tool for prosody visualisation of large speech corpora," in *Proc. Interspeech*, 2017, pp. 809–10.

[3] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer software], retrieved from http://www.praat.org/," 2017.

[4] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Proc. Interspeech*, 2017, pp. 498–502.

[5] A. Öktem, "Proscript: Python library for prosodic annotation of speech segments," 2018. [Online]. Available: https://github.com/alpoktem/proscript

[6] ——, "punkProse: Code for building speech punctuation generation models using lexical, syntactic and prosodic features," 2018. [Online]. Available: https://github.com/alpoktem/punkProse

[7] ——, "Prosograph: A visualizer for prosodically annotated speech corpora written in processing," 2018. [Online]. Available: https://github.com/alpoktem/Prosograph