Tone-First Mandarin Pronunciation Feedback System

ELEC5305 Acoustics, Speech and Signal Processing

Author: Guanzhen Li

SID: 540281325

GitHub guli0019

GitHub Project Link: https://github.com/guli0019/elec5305-project-540281325

Project Overview

The project creates a pronunciation feedback system which pays attention to the Mandarin tones and provides a sentence-level score and character-level feedback. The system takes a short audio recording of a read speech, a corresponding reference text, and t converts the text to pinyin with tone numbers and tone numbers, derives a stable fundamental-frequency (F0) contour of the audio, and finds the similarity between the shape of the the F0 contour of each syllable and a set of tone templates, and maps similarity to a 0–1 tone-correctness score. The scores are summarized in a sentence and given with visuals that can be interpreted: there is an audio player, the F0 curve as a function of time and a color bar below each character (green/yellow/red) to show what is right. The design focuses on clear signal-processing decisions and also pedagogically practical feedback, and is derived on proven strands of pronunciation scoring, text-speech alignment, and contemporary pitch tracking [1][2][3].

Background and Motivation

In Mandarin, lexical identity of syllable is determined by the tone contour; inaccurately toned syllables can drastically lower the intelligibility in the second language learners. Previous research in computer-aided pronunciation training has demonstrated that likelihood/posterior-based techniques at the segment level are associated with human judgments and forced alignment is a common technique in the localize phones and syllables where accurate boundaries are required. Meanwhile, probabilistic YIN (pYIN) provides strong F0 estimates suitable to contour analysis. These threads drive a tone-based system which measures the fit of an observed contour to the desired tone directly and describes that fit using visualizations that are readily comprehensible [1][2][3].

Proposed Methodology

Text processing and timing

The text used is converted to pinyin including tonic numbers. A light voice activity detector divides speech into speech and non-speech, the length of syllables is estimated in read speech with the help of the prosodic features (energy peaks, inter-syllabic pauses) and time normalization in accordance with which each reference syllable is assigned to a time

interval.

Pitch extraction and preprocessing

pYIN is applied to capture an F0 trajectory at every span of syllable. It is smoothed (low-pass or median filtering) and short unvoiced pauses are interpolated, and the output is length-normalized to a constant grid, to allow meaningful comparisons between syllables.

Template construction and matching

Multi-speaker tone templates (T1-T4) are based on native monosyllables and calculated robust averages (varying bands are permitted) of normalized F0 contours. During the run-time, the syllable contour being observed is compared to all templates by a dynamic time warping (DTW) algorithm or by correlation to piecewise-linear prototype contours; the most similar tone and its score of similarity is retained. Similarity is then mapped to a 01 tone correctness score in the target tone on a development set using a calibrated mapping. Examples of tones in the same native corpus are indexed to allow users to audition a reference when they are producing something themselves.

Aggregation and presentation

Sentence-level scoring is based on a weighted average of per-syllable scores (weights represent duration or anticipated prominence). The user interface superimposes the F0 curve onto the waveform, and displays a character-aligned color bar with tooltips which show numeric scores and the nearest competing tone; this displays a quantitative feedback with an at-a-glance diagnosis, which can be interpreted reliably by instructors and learners. A small demo site (GitHub Pages) will contain 2-3 curated examples, with audio, contours, and color bars to the generated underlying JSON data to be reproducible.

Methodological context

Although the scoring of the system is tone-centric, it is also informed by the current techniques: phone-level pronunciation scoring on the basis of likelihood/posteriors, which is extensively used in CALL; text-speech alignment that can be trained to estimate boundaries with high accuracy; and probabilistic F0 tracking that can be trained and works well to extract their contours. Those offer some theoretical basis and directions on how to work towards a further improved system and leave the current system in the context of interpretation and tone feedback [3].

Expected Outcomes

Deliverables will include: a running prototype that takes input audio and text and displays the correct tone per-character, a score on a sentence level, and visual feedback; an interactive plot and a curated collection of examples; and an evaluation report. Primary measures consist of tone classification accuracy on monosyllabic/short-word tests and per-syllable precision/recall/F1 of incorrectness-of-tone flags on a small, de-identified set with instructor annotations. In cases where possible, a small group of sentence-level human ratings will be gathered in order to report Pearson correlation with sentence score. F0 tracking will be tested against error analysis, which will catalogue typical instances of neutral tone and tone-sandhi cases as well as low-energy segments, and will guide setting the thresholds and visualization modifications.

Timeline (Weeks 6–13)

Week	Task
6	Finalize scope; set up repo & Pages
7	Data and preprocessing pipeline
8	Syllable timing + contour normalization
9	Template matching and per-syllable scoring
10	Sentence-level scoring and UI
11	Evaluation design and labeling
12	Experiments and analysis
13	Packaging and submission

References.

[1] Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, *30*(2-3), 95-108.

[2] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017, August). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech* (Vol. 2017, pp. 498-502).

[3] Mauch, M., & Dixon, S. (2014, May). pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 659-663). IEEE.