SPEECH SYNTHESIS AND PRONUNCIATION TEACHING

Waris Quamer
Texas A&M University
quamer.waris@tamu.edu

Anurag Das
Texas A&M University
anuragdiisc.ac.in@tamu.edu

Ricardo Gutierrez-Osuna
Texas A&M University
rgutier@tamu.edu

ABSTRACT

This entry reviews advances in speech-to-text (i.e., speech recognition) and text-to-speech (i.e., synthesis) technologies and how these advances may be used to develop two distinct approaches for pronunciation feedback: *explicit* feedback that uses speech *recognition* techniques to help L2 learners detect, identify and correct pronunciation errors in their speech, and *implicit* feedback that uses speech *synthesis* techniques to generate synthetic voices that L2 learners can use as personalized models. We will provide a brief history of speech-to-text recognition and text-to-speech synthesis through the lens of computer assisted pronunciation training, and present two state-of-the-art models based on modern deep-learning techniques.

KEYWORDS

Computer assisted pronunciation training (CAPT), automatic speech recognition (ASR), mispronunciation detection (MPD), text-to-speech (TTS) synthesis, self-imitation training

1 FEEDBACK IN PRONUNCIATION TRAINING

Conventional wisdom about second language (L2) learning states that simple immersion in the L2-speaking environment will lead to pronunciation improvements. And this appears to be true for the first year in the new environment. However, research also shows that, without instruction that focuses on pronunciation (e.g., vs. lexical/grammatical development), further gains in pronunciation beyond that initial period are negligible (Derwing & Munro, 2013). Further, pronunciation training requires individualized attention and intense practice, which do not lend themselves well to the format of modern language classrooms. For these reasons, computer-assisted pronunciation training (CAPT) has received significant attention over the past two decades. CAPT allows L2 learners to follow personalized lessons, work at their own pace, practice as often as they like, and avoid the anxiety and potential embarrassment of practicing pronunciation in a social setting (Felps et al., 2009).

Critical to the success of any pronunciation training program (computer- or human-mediated) is providing effective feedback to learners. In CAPT, feedback is arguably the biggest challenge. Early CAPT systems relied on visualizations (e.g., speech waveforms, spectrograms) that are both difficult to interpret for non-specialists and potentially misleading: two utterances can have different acoustic representations (e.g., due to pitch and vocal-tract length differences across speakers) despite both having been pronounced correctly. Automatic speech recognition (ASR) addresses many of these limitations and has long been promoted for use in CAPT systems. However, prior to the deep-learning revolution of the past decade, ASR had limited accuracy, in no small part due to the inherent variability of L2 speech. Errors in ASR feedback can be so disruptive to L2 learners that early critics suggested CAPT should rely on implicit rather than explicit feedback (Felps et al., 2009).

In what follows, we discuss how advances in ASR techniques over the past decade can be used to provide *explicit* feedback to the learner by detecting, identifying and correcting pronunciation errors in their speech,

and how text-to-speech (TTS) techniques can also be used to provide *implicit* feedback by generating synthetic voices that L2 learners can use as personalized models for self-imitation training.

2 EXPLICIT FEEDBACK VIA SPEECH RECOGNITION

The primary form of feedback in CAPT, to which we refer as *explicit* (or corrective), is based on mispronunciation detection (MPD). In this setting, the learner is given a sentence to read, and the algorithm highlights severe deviations from the sentence's canonical pronunciation. MPD is closely related to ASR, in that both seek to produce a text transcription of an utterance, but there are major differences between the two. An ASR system (such as those on Siri or Alexa) treats mispronunciations as noise; that is, the goal of ASR is to transcribe the intended sequence of sounds. In contrast, an MPD system treats mispronunciations as the signal to be extracted; in other words, MPD seeks to identify errors, not overlook them. State-of-the-art ASR systems are trained on massively large speech corpora containing millions of utterances from millions of speakers, regardless of their dialect or accent. As an example, the publicly available LibriSpeech ASR corpus contains approximately 1,000 hours of speech from over 1,000 speakers. In contrast, training an MPD system requires a dedicated speech corpus that is phonetically annotated to capture the actual sounds that were produced, a process that is manually intensive and perceptually demanding. As an example, the largest publicly available corpus for MPD (L2-ARCTIC) only contains 24 hours of speech, one hour for each of 24 non-native speakers of English whose first languages (L1s) were Arabic, Hindi, Korean, Mandarin, Spanish and Vietnamese.

Not surprisingly, then, a significant number of studies on L2 pronunciation training rely on off-the-shelf ASR systems. This is largely possible because the practice words/sentences are generally known in advance. Thus, mispronunciations can be detected by analyzing the output probabilities of an ASR model trained on native speech. The classical example is the Goodness of Pronunciation (GOP) measure, which is computed as the ratio of the posterior probability of the target phoneme (i.e., in the practice word/sentence) relative to that of the most likely phoneme (i.e., according to the ASR model). Then, a mispronunciation is detected by comparing the GOP measure against a threshold that has been optimized through cross-validation. However, the accuracy of such methods relies heavily on the accuracy of the alignment between speech audio and text transcription (i.e., forced alignment), which can be challenging if the L2 learner's production differs significantly from the target pronunciation. Moreover, these methods only deal with substitution errors, but not insertion errors. To avoid these issues, newer approaches consider alternative pronunciation sequences that are likely to be uttered by the learner, known as extended recognition networks. While this approach allows the system developer to take advantage of common error patterns and phonotactic constraints, building these extended recognition networks is time consuming and requires linguistic expertise. Further, these systems are unable to detect mispronunciations that are not included in the set of hand-crafted rules. Techniques have also been developed to discover pronunciation error patterns in a data-driven fashion, though this requires access to fully transcribed non-native corpora, which are generally limited.

As more accurate MPD models continue to be developed, it is unclear whether corrective feedback that is 100% accurate is necessary in order for L2 learners to benefit from it and improve their pronunciation. A recent study by Silpachai et al. (in press) sheds light on this issue. In the study, Chinese learners of English were trained to produce nine sound contrasts in English while receiving corrective feedback from a web-based MPD system. Unknown to the students, the MPD system was being partially operated by one of the experimenters following a Wizard-of-Oz paradigm. L2 learners' recordings were provided to the experimenter in real time, who then selected the appropriate corrective feedback. Finally, the experimenter's recommendations were changed randomly so that the system was accurate 33%, 66%, or 100% of the time (the latter being the unmodified feedback from the experimenter). When rated for accuracy by native speakers of English, L2 learners' productions were similar for the 100% and 66% accuracy conditions, and both were more accurate than for the 33% accurate condition. In other words, MPD feedback was beneficial even when it was less accurate than the gold standard (human feedback).

3 IMPLICIT FEEDBACK THROUGH SPEECH SYNTHESIS

Several studies during the last three decades have suggested that it would be beneficial for L2 students to be able to listen to native voices similar (if not identical) to their own voices. In a landmark study, Probst et al. (2002) sought to identify important dimensions of similarity that a model voice should have for it to be a good match for a given L2 learner. Their rationale was that matching an L2 learner to a model voice with similar characteristics, would "free" the learner from having to attend to these variables, thus reducing the complexity of the pronunciation learning task. In a pronunciation training experiment, L2 learners who imitated a well-matched speaker improved more than those who imitated a poor match, suggesting the existence of a user-dependent "golden speaker." Among the three characteristics they considered (gender, pitch and speed of articulation), the latter was the most influential factor. It is worth noting that a similar concept has emerged in the talker-adaptation literature: the existence of a "a "golden talker" whose inherent range of systematic variability allows generalization to a novel talker" (Baese-Berk et al., 2013).

Building on this prior research, Felps et al. (2009) argued that the ideal golden speaker for each L2 learner would be their own voice, resynthesized to have a native accent. A handful of studies over those three decades had shown that when L2 learners imitate their own voices (self-imitation) with native-like prosody they were able to improve their pronunciation, and these findings have been corroborated across multiple L1-L2 language pairs (Japanese learners of English, L2 learners of English, German and Italian with various L1 backgrounds). A recent study by Pellegrino (2024) summarized this prior research and examined the extent to which self-imitation leads to pronunciation improvements. In her study, Japanese learners of Italian practiced three different speech acts (commands, requests and grants), first without guidance or instructions (pre-test), then with prosodically corrected versions of their own voices (post-test). Compared to their pre-test utterances, L2 learners' utterances at post-test showed convergence to the prosody of L1 utterances, measured in terms of duration, F0 mean and F0 max per syllable. More interestingly, convergence to the L1 prosody was stronger when the initial acoustic distance between the L1 model and the L2 speaker was larger, corroborating results. However, the study did not include a control condition (i.e., L2 learners who only practiced with the L1 model), so these results are inconclusive.

To our knowledge, Ding et al. (2019) were the first to examine self-imitation of both prosodic and segmental characteristics. In the study, Korean learners of English practiced with a "golden speaker" version of their own voice, resynthesized using statistical machine learning techniques running on a web application (Golden Speaker Builder). L2 learners' productions at posttest were rated (by independent listeners) as being more comprehensible and fluent than those at pretest. Further, L2 learners reported that practicing with their "golden speaker" voices helped them perceive differences in intonation and stress in their unmodified speech. As in Pellegrino (2024), the study lacked a control group, so it is possible that practicing with a different model voice would have been equally or more effective.

4 CASE STUDIES

The field of speech technology experienced remarkable progress from 2010 to 2020 and does not appear to be slowing down. ASR systems have accuracies that exceed that of humans (less than 5%). State-of-the-art synthesized speech is perceptually indistinguishable from natural speech, though that is not anymore always a good thing (e.g., President Biden's deepfake robocalls in 2024 are not a desirable outcome of this technology). At the same time, such advances make it possible to develop pronunciation training tools that in 2010 would appear in the "future work" section of grant applications. This section explores two state-of-the-art systems that have been specifically developed for pronunciation training. The first system can generate "accent conversions" of L2 speech with independent control of segmental and prosodic characteristics. The second system combines text-to-speech recognition with mispronunciation detection to improve detection accuracy.

4.1 Accent conversion

The problem of generating "golden speaker" voices for L2 learners is commonly referred to as "accent conversion," in reference to the closely related problem of "voice conversion." In voice conversion, one seeks to transform utterances from a source speaker to sound as if a (known) target speaker had produced them. The conversion aims to match the organic properties of the target speaker (vocal tract configurations, glottal characteristics, age, gender) as well as the speaker's pronunciation patterns, including their dialect or accent; the only information to be preserved from the source utterance is its orthographic transcription. In contrast, accent conversion seeks to disentangle organic properties from pronunciation. Its goal is to produce speech with the orthographic content *and* pronunciation characteristics of a (native) source utterance with the organic characteristics of the (non-native) target speaker. Thus, accent conversion is more challenging than voice conversion. Not only does it require a more fine-grained disentanglement of voice characteristics, but it must do so without a ground truth since recordings of the (non-native) speaker producing speech with the desired (native) target accent do not exist.

Early approaches for accent-conversion required parallel recordings of the native and non-native speakers producing the same sentences. Before building a model, parallel recordings would first have to be aligned, frame by frame. The resulting lookup table of short audio frames for both speakers would then be used to build a multivariate regression mapping via machine learning algorithms. To generate an accent conversion, a source utterance from the native speaker would be split into frames and each frame mapped into the corresponding frame for the target speaker while minimizing acoustic discontinuities at the output. The critical step of these models was the initial alignment step; otherwise, the machine-learning model would preserve the accent of the non-native speaker.

Modern systems based on deep-learning techniques do not require parallel recordings, thus avoiding the challenge of aligning source and target utterances. Instead, these systems are trained to reconstruct the input speech at the output (self-supervision) under carefully designed constraints built into the architecture. For example, the system illustrated in Figure 1 can generate speech with the segmental characteristics of a first utterance (U1), the voice identity of a second utterance (U2) and the prosody characteristics of a third utterance (U3). Utterance U1 is passed through a speech recognizer, which generates a speaker-independent representation of the utterance's contents in the form of a phonetic posteriorgram (a vector that represents the probability each frame belongs to one of roughly 5,000 triphones). This phonetic posteriorgram is passed through a module that removes speaking-rate information. This ensures that the speech synthesizer downstream learns to reconstruct the utterance based on the prosody contained in utterance U2, and the voice quality of utterance U3. The speaker recognizer and speaker encoder are pretrained on a large speech corpus. Thus, only the speech synthesizer and the prosody encoder need to be trained. This type of deep-learning architecture represents the segmental content, prosodic content and voice quality of utterances as high-dimensional numerical vectors (embeddings) that can be combined in a manner that resembles algebraic operations.

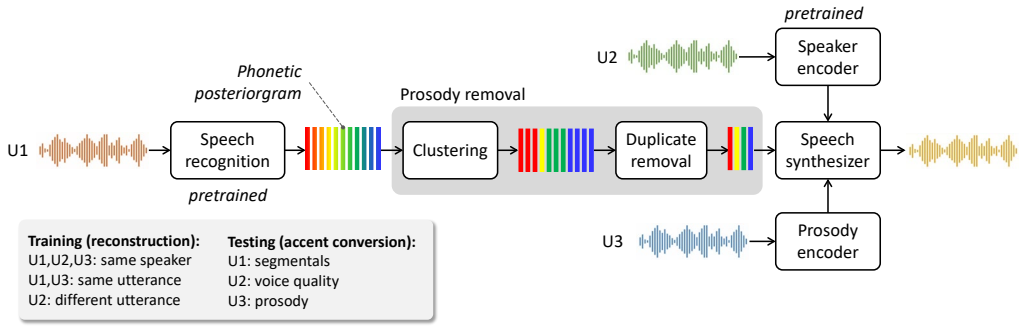

Figure 1. Deep learning architectures provide a high level of abstraction by representing the content of an utterance (segments, prosody, voice quality) as numerical vectors and combining them in an algebraic fashion.

4.2 Joint mispronunciation detection and speech synthesis

Research in machine learning shows that solving multiple tasks simultaneously can improve learning efficiency and generalization performance, so long as the tasks are related. For example, models trained to perform speech recognition and speaker recognition simultaneously outperform models trained to perform each task separately. Similar findings have been reported in other related areas such as natural language processing and text-to-speech (TTS) synthesis. The goal of TTS is to generate speech waveforms from text inputs. Early systems were called formant synthesizers because they generated speech by combining different frequencies (formants) to mimic the sound of human speech. Stephen Hawking's original robotic voice was based on formant synthesis. The next generation of TTS systems worked by concatenating sounds from a database of short speech segments. The database could be relatively small as in the case of diphone synthesizers, or large speech corpora (i.e., hours of speech) as in the case of unit-selection synthesis. As with speech recognition, the field of TTS made enormous strides between 2010 and 2020. Unit-selection systems, which were state-of-the art around 2010, were slowly replaced by so-called statistical parametric speech synthesizers based on hidden Markov models. Over the past 10 years, though, the field has come to adopt deep learning techniques.

Considering that both MPD and TTS operate at the level of individual sound symbols (i.e., orthographic, phonetic), a recent study by Das et al. (2024) examined whether combining the two problems in a multi-task learning fashion would boost MPD performance. The basic architecture of the system is illustrated in **Figure 2**. An utterance from an L2 learner is passed through a speech recognizer to generate a phonetic posteriorgram, and the phonetic posteriorgram is passed to a sequence-to-sequence (seq2seq) model along with the text prompt given to the L2 learner. The seq2seq model combines both sources of information to produce a phonetic sequence. The phonetic sequence is then compared against the canonical transcription of the text prompt, and any mismatches (additions, deletions, substitutions) are flagged as potential mispronunciations. This can be treated as a baseline MPD system. To improve performance, Das et al. (2024) connected the output of the MPD system to a text-to-speech synthesizer that was trained to reconstruct the audio waveform at the input from the phonetic transcription and an embedding that captured the voice characteristics of the speaker. Effectively, this provided the MPD system with two error signals to minimize: (1) differences between the MPD output and the ground truth phonetic transcription, and (2) differences between the original speech and its reconstruction. The combined system (MPD+TTS) significantly outperformed the baseline system (MPD) in correctness and accuracy, both on seen and on unseen utterances, indicating that the TTS reconstruction task forced the MPD system to learn a better decoder function.

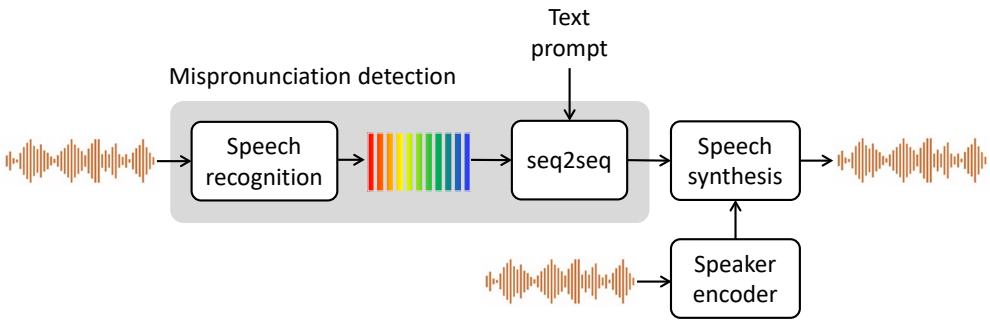

Figure 2. Deep-learning architecture for joint mispronunciation detection and text to speech synthesis

5 CONCLUSIONS

In the previous section, we presented two modern systems for pronunciation training that take advantage of major advances in speech technology. As shown in the two figures, these systems blur the lines between the complementary problems of speech recognition (speech-to-text), speech synthesis (text-to-speech) and

speaker recognition (biometric authentication). They also allow us to disentangle the various contributors to the speech signal, such as speaker characteristics, regional/non-native accents, segmental and prosodic information, and later recombine them to generate a variety of speech stimuli for pronunciation training – see Suggested Reading.

6 CROSS-REFERENCES

See Also

wbeal20673

wbeal20641

wbeal20660

wbeal20672

wbeal20669

REFERENCES

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *J Acoust Soc Am*, *133*(3), El174-180. https://doi.org/10.1121/1.4789864

Das, A., Quamer, W., & Gutierrez-Osuna, R. (2024). Improving mispronunciation detection using speech reconstruction. *IEEE Transactions on Audio, Speech and Language Processing*, *under review*.

Derwing, T. M., & Munro, M. J. (2013). The Development of L2 Oral Language Skills in Two L1 Groups: A 7-Year Study. *Language Learning*, *63*(2), 163-185. https://doi.org/10.1111/lang.12000

Ding, S., Liberatore, C., Sonsaat, S., Lučić, I., Silpachai, A., Zhao, G., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2019). Golden speaker builder – An interactive tool for pronunciation training. *Speech communication*, *115*, 51-66. https://doi.org/https://doi.org/10.1016/j.specom.2019.10.005

Felps, D., Bortfeld, H., & Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. *Speech communication*, *51*(10), 920-932.

Pellegrino, E. (2024). After Self-Imitation Prosodic Training L2 Learners Converge Prosodically to the Native Speakers. *Languages*, *9*(1), 33. https://www.mdpi.com/2226-471X/9/1/33

Probst, K., Ke, Y., & Eskenazi, M. (2002). Enhancing foreign language tutors – In search of the golden speaker. *Speech communication*, *37*(3), 161-173. https://doi.org/https://doi.org/10.1016/S0167-6393(01)00009-7

Silpachai, A., Neiriz, R., Novotny, M., Gutierrez-Osuna, R., Levis, J. M., & Chukharev-Hudilainen, E. (in press). Corrective feedback accuracy and pronunciation improvement: Feedback that is "good enough". *Language Learning & Technology*.

SUGGESTED READING

Agarwal, C., & Chakraborty, P. (2019). A review of tools and techniques for computer aided pronunciation training (CAPT) in English. *Education and Information Technologies*, *24*, 3731-3743.

Korzekwa, D., Lorenzo-Trueba, J., Drugman, T., & Kostek, B. (2022). Computer-assisted pronunciation training—Speech synthesis is almost all you need. *Speech communication*, *142*, 22-33.

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

CONTRIBUTOR BIOS

Anurag Das received his BTech degree in Electronics and Communication Engineering from the National Institute of Technology, Silchar (India) in 2016. He is currently pursuing his PhD degree in Computer Science from Texas A&M University College Station, TX, USA. His research interests include mispronunciation detection, developing machine learning solutions for diabetes management. His notable publications include Understanding the Effect of Voice Quality and Accent on Talker Similarity (Interspeech, 2020).

Waris Quamer received his BTech degree in Computer Science and Engineering from Indian Institute of Technology, (ISM) Dhanbad in 2019. He received his Master of Science degree in Computer Science from Texas A&M University College Station, TX. He is currently pursuing his PhD degree in Computer Science from Texas A&M University College Station, TX, USA. His research interests include speech synthesis for accent and voice conversion. His notable publications include Zero Shot Accent Conversion without native reference (Interspeech, 2022), and Decoupling segmental and prosodic cues of non-native speech through vector quantization (Interspeech, 2023).

Ricardo Gutierrez-Osuna received his B.S. in Industrial Engineering from the Polytechnic University of Madrid (Spain) in 1992, and M.S. and Ph.D. degrees in Computer Engineering from North Carolina State University in 1995 and 1998, respectively. He is a professor in the Department of Computer Science and Engineering at Texas A&M University. His research interests include intelligent sensors for digital health applications, speech and language processing, and machine learning. For the past 15+ years he has conducted research in speech modification and resynthesis of non-native speech.