STATISTICAL PHONE DURATION MODELING TO FILTER FOR INTACT UTTERANCES IN A COMPUTER-ASSISTED PRONUNCIATION TRAINING SYSTEM

Wai-Kit Lo, Alissa M. Harrison and Helen Meng

The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
{wklo, alissa, hmmeng}@se.cuhk.edu.hk

ABSTRACT

We study the use of a statistical phone duration model for separating intact utterances from corrupted ones in a computer-assisted pronunciation training system. Our system performs forced alignment between the input utterance and the canonical transcription of the prompted text. Intact utterances contain spoken content that correspond to the text prompt. For these utterances, our system performs detailed phonetic analysis of the alignment and generates corrective feedback to highlight the occurrence of phonetic errors. Corrupted utterances result from disfluencies, truncated recordings, or spoken content that does not correspond to the text prompt. For these cases, the appropriate feedback is to invite the user to record again. We develop a filtering mechanism for intact input utterances by means of phone duration modeling. The likelihood-ratio-test involving the phone-specific duration probability and an antimodel probability gave the best EER of 17.16%, which is a 20% relative improvement over the baseline approach that incorporates phone-posterior probabilities.

Index Terms— computer-aided pronunciation training, phone duration modeling, user interface

1. INTRODUCTION

Computer-Assisted Pronunciation Training (CAPT) [1, 2] uses automatic speech recognition (ASR) technology to help improve the learner's pronunciation. Pronunciation exercises and objective feedback are critical for language learning. The major benefit of CAPT is that language learners can practice speaking in a private, self-paced and possibly round-the-clock environment.

We have developed a research prototype for a CAPT system. The system presents a pre-designed sentence to the learner and prompts for an input utterance. It then performs a forced alignment between the input utterance and extended phonetic transcriptions of the text prompt. The canonical phonetic transcription is obtained by dictionary lookup. From the canonical version, our system automatically predicts possible mispronunciations using phonological rules or a data-driven approach [3, 10] and generates extended phonetic transcriptions that are also made available in forced alignment. Thereafter, the system performs detailed analysis for the phonetic alignment to perform mispronunciation detection and diagnoses. This enables generation of corrective feedback messages that inform the user about detailed phonetic errors.

In this usage context of a self-directed learning tool, most users (i.e. learners) are cooperative. However, anecdotal observations based on new users show that there are several common factors that may cause corruptions to the input utterances. For instance, there may be disfluencies (such as false starts, repairs, repetitions). Users may stop reading before completing the prompt text, due to distractions, side conversations, etc. The recording may also be truncated at its end-points, possibly due to the user pressing the <stop> button too early. These corrupted utterances should be handled differently by the system, as compared with an intact utterance whose spoken content corresponds well with the text prompt. More specifically, our system generates corrective feedback for an intact input utterance to inform the user of discovered phonetic errors. However, appropriate feedback for a corrupted input should prompt the user to record again. Hence, there is strong motivation to develop a filtering mechanism that separates the two types of utterances, as illustrated in Figure 1.

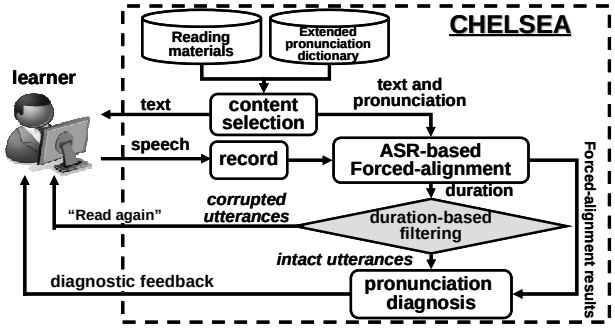

Figure 1: CHELSEA: A CAPT system equipped with a filtering process that separates intact utterances from corrupted ones.

Confidence measures have been used in earlier work to verify that an input utterance has appropriate content for the speech application [4]. For example, a phone-dependent confidence measure is used for utterance rejection in [5]. In [6], the generalized word posterior probability is computed for each word and utterance rejection is performed based on a combination of word scores. Phone duration has been used as feature for computing confidence measures in ASR applications for embedded and noise environments [7, 8] as well as verifying selected utterances in a language learning application [9]. In this work, the forced alignment nature of our CAPT approach (explained in the next section) can exaggerate the phone duration variations in corrupted

utterances. We will investigate the use of a statistical phone duration model to filter for intact utterances that can further undergo detailed phonetic analysis for mispronunciation detection and diagnosis.

2. THE CHELSEA CAPT SYSTEM

CHELSEA is a CAPT system developed specifically for Chinese learners of English, i.e. learners whose primary language (L1) is Chinese (either Cantonese or Putonghua) and secondary language (L2) is English. The system incorporates specially designed text prompts to elicit L2 utterances that may contain mispronunciations due to negative language transfer. Mispronunciation detection is achieved through forced phonetic alignment by the automatic speech recognizer, between the recorded utterance and the extended (canonical and variants) pronunciations of the text prompt.

The recognizer uses acoustic models trained with (US) English speech from the TIMIT corpus. The models are cross-word triphone HMMs (3 emitting states, 12 Gaussian mixtures, 13 PLP+Δ+ΔΔ, with cepstral mean normalization). The recognizer's vocabulary consists of an extended pronunciation dictionary (EPD). The EPD augments canonical word pronunciations extracted from a standard pronunciation dictionary with non-native pronunciation variants typical of Chinese learners of English [3, 10]. These variants are generated via phonological rules or a data-driven approach [3, 10, 11] in order to capture the common mispronunciations of Chinese learners. This approach for CAPT can be applied to any language pairs by providing the appropriate rules.

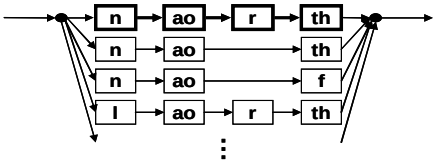

Figure 2: Example EPD showing the canonical pronunciation (in bold) and pronunciation variants for the word "north".

The recognition grammar in CHELSEA is generated dynamically from the words in the text prompt by pronunciation lookup from the EPD. An illustration is provided in Figure 2 for the word "north". When presented with an input utterance, the system will *forced-align* it with all the possible pronunciations of the text prompt, with reference to the EPD. Should the best alignment be one of the variants (as opposed to the canonical pronunciation), the system will be able to pinpoint the location(s) and type(s) of the mispronunciation(s), as shown in Figure 3. This forced alignment procedure also generates phone boundaries, from which phone durations may be obtained.

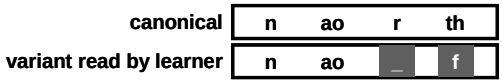

Figure 3: An illustration of captured mispronunciations that are highlighted in the system's feedback for the user – more specifically, /r/ is deleted and /th/ is misread as /f/.

3. DURATIONS OF FORCED-ALIGNED PHONES

We need to filter for intact utterances (i.e. those of reasonable quality and relevance) that can be appropriately subjected to detailed phonetic analysis for mispronunciation detection and diagnosis. We assume that the phones in an intact utterance should largely carry their respective inherent durations. Our filtering approach references the durations obtained by the recognizer through forced alignment.

As an illustration, Figure 4 shows the word "north" in a sentence which is one of the system's text prompts. If the learner utters the text prompt with correct pronunciation as in (a), the phone durations should resemble their inherent values. In (b), the learner mispronounces the word and the best alignment selects the pronunciation variant that is among those predicted in the EPD. The phone durations in the forced alignment should also resemble their inherent values.

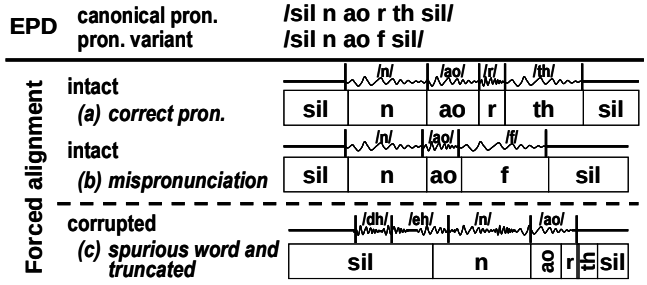

Figure 4: This figure illustrates forced alignment between an input utterance and the best-matching phone sequence from the extended pronunciation dictionary (EPD). Forced alignment produces reasonable phonetic durations for an intact utterance. On the other hand, phonetic durations of corrupted utterances tend to be overly long or short.

In (c), the input utterance (with a phone sequence of /dh eh n ao/) does *not* correspond in any way to the prompted text (that includes the word "north" with reference phoneme sequence /n ao r th/). Forced alignment makes the best effort possible to align the input utterance with one of the pronunciations in the EPD. This results in the frames of spurious phones (e.g. /dh/ and /eh/ that do not appear in the pronunciation of "north") being absorbed by the SILENCE segment or a non-silence phone segment(s) (e.g. /eh/ being absorbed into the /n/ segment). The latter causes lengthening of the absorbing phone segment. As for missing phones (e.g. /r/ and /th/ that occurs in the word "north" but are absent in the input utterance) in the EPD pronunciation that do not correspond to any acoustic frames, they tend to be assigned very short durations by the alignment algorithm. Hence, if forced alignment produces phone durations that are overly long or short, as compared with their inherent values, it may suggest that the input utterance is *not intact* and should not be subjected to further detailed phonetic analysis. As such, we can design a filtering approach based on phonetic durations to identify intact utterances that are analyzed phonetically for further mispronunciation detection and diagnoses.

4. THE CU-CHLOE CORPUS

We have designed and collected the Chinese University CHinese Learners Of English (CU-CHLOE) corpus. This contains English recordings from 100 speakers (50 male and 50 female) whose mother tongue is Cantonese. Each speaker records (i) The Aesop's Fable "The North Wind and the Sun" (NW), which includes six sentences and has a good coverage of English phones; as well as (ii) a set of 20 phonemic sentences (PS) that are specially designed by experienced English teachers to cover common English mispronunciations.

We divide the corpus into disjoint training and testing sets. Recordings of the NW from 50 speakers (25 male and 25 female) are used for training parameters and tuning decision thresholds. Recordings of both the NW and PS of the remaining 50 speakers are used for testing.

Anecdotal observations when the CHELSEA system was demonstrated to general users shows several main types of "corruption" that causes an input utterance to be filtered out: (i) the input utterance does not follow the prompted text (e.g. due to side conversations); (ii) the user did not speak the complete prompt; (iii) the recording is truncated (e.g. the user presses the stop button prematurely) and (iv) the input utterance includes a restart (e.g. "I said … I said that he is a good student"). Based on such observations, we augment the test set by simulating the respective types of corruptions by: shuffling among text prompts and their corresponding utterances, taking a partial initial segment of an utterance, truncating some of the test utterances, as well as duplicating the initial part of an utterance to simulate a restart. Figure 5 illustrates the organization of our experimental corpus:

Speaker ID		*F01 M01 … F25 M25*	*F26 M26 … F50 M50*
Intact utterances	NW	25M + 25F, 6 utt. each (300 utt. in total)	25M + 25F, 6 utt. each (300 utt. in total)
	PS		25M + 25F, 20 utt. each (1000 utt. in total)
Simulated non-intact utterances	NW	25M + 25F,6 utt. each, 4 types (1200 utt. in total)	25M + 25F, 6 utt. each, 4 types (1200 utt. in total)
	PS		25M + 25F, 20 utt. each 4 types (4000 utt. in total)
		Training set	Test set

Figure 5: Organization of the CU-CHLOE corpus for our investigation. The normal corpus data are used as intact utterance. Corrupted utterances are simulated with 4-types of corruptions: *shuffled*, *partial*, *truncated*, and *restarted*.

5. MODELING PHONE DURATIONS WITH THE GAMMA DISTRIBUTION

Phone durations vary across speakers and utterances and have often been modeled statistically by the Gamma distribution [7, 12, 13]. We verified this based on the corpus statistics of the NW recordings in the CU-CHLOE training set which contains speech data from non-native speakers (see Figure 6). The duration distributions of certain phones (especially consonants) tend to fit well with the exponential distribution – a special case of Gamma.

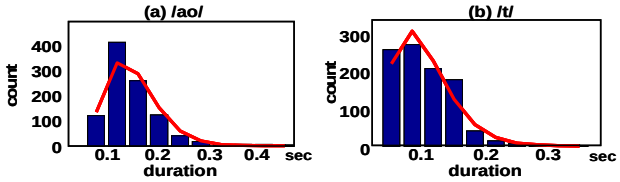

Figure 6: Phone duration histograms from the CU-CHLOE corpus: (a) for the vowel /ao/, (b) for the plosive /t/. Corpus statistics (bars) are fitted with Gamma distributions (lines).

We have also examined the phone duration statistics of the "shuffled" corpus. We observed a concentration of phone with durations near zero millisecond, while the remaining phones exhibit a general exponential distribution. This confirms our speculation that when the input utterance does not correspond to the text prompt, forced alignment tends to produce phone durations that are overly short or long Hereafter we adopt the Gamma distribution in the model and anti-model for phone durations.

6. FILTERING FOR INTACT UTTERANCES BY PHONE POSTERIOR PROBABILITIES

Our baseline filtering approach scores an utterance with the normalized product of phone posterior probabilities [11, 14]. The equation is:

$$\frac{1}{\sum_{\forall r \in S} dur(r)} \log\left(\prod_{\forall p \in S} \frac{a(\bar{O}_{t_s,t_e} \mid p)}{\sum_{\forall q \in L} a(\bar{O}_{t_s,t_e} \mid q)} \right) \qquad (1)$$

where a is the acoustic score returned by the ASR, O is the acoustic observation, t_s is the starting time, t_e is the ending time, $dur(r)$ is the duration of the phone r, S is the set of phones in the utterance and L is the set of phones in the language. This gives an EER of 21.48% over the test set.

7. FILTERING FOR INTACT UTTERANCE BY PHONE DURATION MODELS

7.1. Phone Sequence Duration Model

We devise a phone sequence duration model by estimating the joint duration probabilities for the phones in the utterance and incorporating length normalization:

$$\frac{1}{\|S\|} \log\left(\prod_{p \in S} P(dur(p) \mid p) \right), \qquad (2)$$

where S is the set of phones in the utterance, $\|S\|$ is the number of phones in the utterance. The statistical phone duration model P is the Gamma distribution with trained parameters (based on the training set). Evaluation on the test set gives 22.62% in EER.

7.2. Likelihood Ratio Test between the Phone Duration Model and an Anti-model

We also incorporate an anti-model in phone duration scoring, with the aim to increase the discriminative power of the phone duration model. A likelihood ratio test (LRT) is applied as shown in Equation (3):

$$\frac{1}{\|S\|} \log\left(\prod_{p \in S} \frac{P(dur(p) \mid p)}{P_{anti-model}(dur(p))} \right). \qquad (3)$$

7.2.1. Highest-scoring Competing Phone as Anti-model

One method of realizing an anti-model is to find the phone (among all the alternatives in the inventory) that maximizes the observed phone duration probability, as follows:

$$\frac{1}{\|S\|}\log\left(\prod_{p\in S}\frac{P(dur(p)\mid p)}{\max\limits_{q\in L,\,p\neq q}\left(P(dur(p)\mid q)\right)}\right) \qquad (4)$$

where L is the set of phones in the language. This approach achieves a test set EER of 17.33%.

7.2.2. The "Catch-All" Anti-model

Another method of realizing an anti-model is to train a Gamma distribution based on a "shuffled" training set (here we use the NW subset). The rationale is to obtain a "catch-all" anti-model for use in the LRT where each phone duration model is trained with non-corresponding phonetic segments in the utterance. This method achieves an EER of 17.16% over the test set.

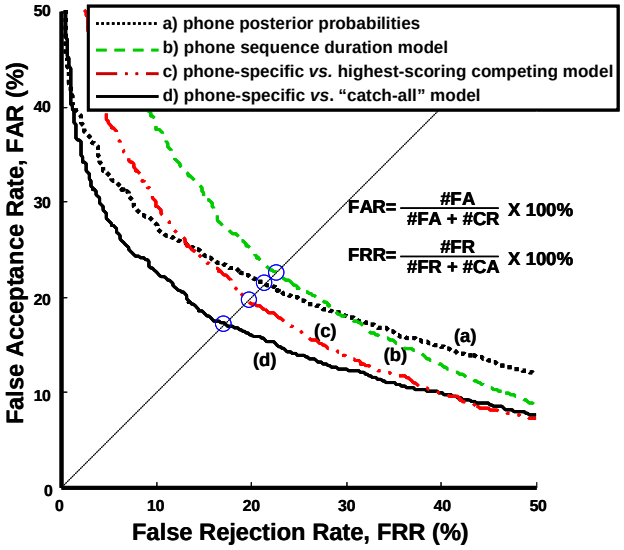

Figure 7: ROCs of the tested approaches for identifying corrupted utterances (test set). LRT between phone-specific duration model and anti-model performs the best.

8. OBSERVATIONS AND ANALYSIS

Results in Figure 7 indicate that the use of an anti-model in the LRT offers additional discriminative power in filtering to yield favorable performance. In particular, the catch-all anti-model performs better than the one that uses the highest-scoring competing phone. This is because the former method can better target the phone durations resulting from forced alignment with corrupted utterances.

9. CONCLUSIONS

We explore the use of phone duration modeling to filter for intact utterances that are input into a CAPT system. The experimental corpus contains non-native English speech from 100 Chinese learners. The Gamma distribution is verified to achieve a good fit with the data and is adopted as the phone duration model. Filtering methods based on a variety of phone duration models, anti-models and likelihood ratio tests (LRT) were investigated. The best filtering performance (i.e. in rejecting corrupted utterances) is attained with the LRT involving the phone-specific duration model and an anti-model that is specially trained with simulated corrupted utterances. Evaluation is conducted with test data that involves 50 speakers which are disjoint from the set of training speakers. A relative improvement (reduction of EER) of 20% is achieved in comparison with the baseline method that uses phone posterior probabilities. In additional to CAPT system, this proposed filtering approach can also be applied to other applications that need to verify the recorded speech content to a user prompt.

10. ACKNOWLEDGMENTS

This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies. We also thank Dr. Frank Soong of MSRA.

11. REFERENCES

[1] Kawai G. and Hirose K., "A Call System Using Speech Recognition to Teach the Pronunciation of Japanese Tokushuhaku," STiLL-1998, pp. 73-76, 1998.

[2] Kawahara T. *et al.*, "Practical Use of English Pronunciation System for Japanese Students in the CALL Classroom," INTERSPEECH2004, pp. 1689-1692, 2004.

[3] Meng H. *et al.*, "Deriving Salient Learners' Mispronunciations from Cross-Language Phonological Comparisons," ASRU2007.

[4] Jiang H., "Confidence measures for speech recognition: a survey," Speech Communication, vol. 45, pp. 455-470, 2005.

[5] Cohen M. *et al.*, "A phone-dependent confidence measure for utterance rejection," ICASSP1996, pp. 515-518.

[6] Lo W. K. and Soong F., "Generalized posterior probability for minimum error verification of recognized sentences," ICASSP2005, pp. 85-55.

[7] Pellorn B. L. and Hansen J. H. L., "A duration-based Confidence Measure for Automatic Segmentation of Noise Corrupted Speech," ICSLP1998, 1998.

[8] Goronzy S. *et al.*, "Phone-duration-based Confidence Measures for Embedded Applications," ICSLP2000, 2000.

[9] Doremalen J., *et al.*, "Utterance Verification in Language Learning Applications," SLaTE2009, 2009.

[10] Harrison A. *et al.*, "Improving Mispronunciation Detection and Diagnosis of Learners' Speech with Context-Sensitive Phonological Rules Based on Language Transfer," INTERSPEECH2008, pp. 2787-2790.

[11] Lo W. K. *et al.*, "Decision Fusion for Improving Mispronunciation Detection using Language Transfer Knowledge and Phoneme-dependent Pronunciation Scoring," ISCSLP2008.

[12] Levinson S. E., "Continuously Variable Duration Hidden Markov Models for Speech Analysis," ICASSP1986, pp. 1241-1244.

[13] Ramus F., "Acoustic Correlates of Linguistic Rhythm: Perspectives," SpeechProsody2002, pp. 115-120.

[14] Witt S. M. and Young S. J., "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning," Speech Communication, 30:95-108, 2000.