Tactical Language Detection and Modeling of Learner Speech Errors: The case of Arabic tactical language training for American English speakers

Nicolaus Mote, Lewis Johnson
USC / Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292
{mote, johnson}@isi.edu

Abhinav Sethy, Jorge Silva, Shrikanth Narayanan
University of Southern California
Los Angeles, CA 90089
{sethy,jorgesil,shri}@sipi.usc.edu

Abstract

The Tactical Language Training System (TLTS) is a speech-enabled computer learning environment designed to teach Arabic spoken communication to American English speakers (and is described in a companion paper(Johnson et al, 2004)). This paper elaborates upon the modeling and detection of learner speech errors along multiple levels of linguistic details ranging from segmental to lexico-semantic aspects. Detecting learner errors enables providing tailored pedagogical feedback in TLTS.

1 Introduction

This paper describes the techniques used in the Tactical Language Training System (TLTS) for detecting and detecting errors in learner speech. As described in the companion paper in this volume (Johnson et al., 2004), the TLTS incorporates two speech-enabled learning environments: an interactive game called the Mission Practice Environment (MPE) that simulates conversations with native speakers, and an intelligent tutoring systems called the Mission Skill Builder (MSB) for acquiring and practicing communicative skills.

Error modeling in TLTS serves two purposes: first, to recognize what the learner intends to say, and second, to recognize the deviations the learner makes from what he intends to say. The components of the TLTS must each be robust in the face of learner errors. The MSB gives learners detailed feedback on their speech errors, and therefore needs to identify and classify those errors. Confidence of error detection is also important, so that the system avoids giving negative corrective feedback when recognition confidence is low.

Recognition here is accomplished by a Hidden Markov Model ASR system that has been bootstrapped from Modern Standard Arabic speech and enhanced with data from native and learner Lebanese Arabic speech. Our speech recognition engine departs from the standard ASR task in a couple of notable ways: we not only recognize true Arabic words, but also mispronounced and misuttered Arabic words. Secondly, because we are dealing with learner speech, we need deal only with a smaller subset of language that corresponds to what the learner has been taught, so we can safely reduce our recognition vocabulary size. This simplification is necessary, because supplementing our base Arabic recognition grammar with disfluencies upon each item in the original grammar increases the HMM state size to the degree that robust detection would be untenable otherwise.

In the next section we give an overview of the speech recognition system used in the TLTS. Sections 3 and 4 describe the overall methods employed for modeling language errors. Section 5 describes the speech recognition methods used to perform speech error detection and scoring.

2 Speech Recognition Overview

2.1 Training setup and system design

The speech recognition system was implemented using the Cambridge HTK toolkit. The feature space comprised of 12 mel frequency cepstral coefficients extracted at a frame rate of 10ms using a 16 ms hamming window. First and second order differentials plus an energy component were also included. Monophonic models were built for the 37 phones in our Levantine Arabic lexicon. A skip state silence model was also trained. The phone models had three states with eight mixture components.

The system was trained on a modern standard Arabic dataset with around 10 hours of native speech. A mapping from modern standard Arabic phone set to Levantine Arabic phone set was used for this purpose. To support learner speech recognition in the TLTS our initial efforts focused on acoustic modeling for robust speech recognition especially in light of limited domain data availability (Srinivasamurthy & Narayanan, 2003). Non-native speech examples (around one hour) collected from our trial runs of the learning system were also included in the training data.

2.2 USCPers

The transcription system that we use for displaying Arabic text and automated processing of learner speech is based on USCPers (Ganjavi et al, 2003). USCPers is designed for automated processing of languages that use the Arabic script. It was originally developed for Persian but is extended easily to other languages such as Dari, Urdu, and Arabic. The system is based on the ASCII symbol set and provides for both a phonemic based transcription system and a scheme, which explicitly includes vowels to reduce word sense ambiguity. This is particularly important as Arabic script omits soft vowels in the traditional written form. For example, the written forms of the words 'six' and 'lungs' are the same in Arabic while their pronunciations are different due to different vowel placements. The USCPers system allows us to disambiguate between such 'homograms' and makes them amenable to processing by ASR and Natural Language Processing (NLP) systems. We use a direct phonetic mapping from modern standard and Levantine Arabic to USCPers.

2.3 Recognition setup

The speech recognizer is used in two different fashions. For the MSB, the recognizer is constrained to recognize only the pronunciation variants of the utterance currently being taught. In the MPE, the recognizer network is an FSG, which has all the utterances in the MSB as parallel paths.

The MSB uses dynamic loading of networks for ASR. For each utterance being recognized, the schoolhouse signals the ASR to load a different recognition network. The choice of recognition network to use is governed by the progress the learner has made in the MSB curriculum and this allows us to use priors conditioned on the learner progress and other factors which the learner agent models can define.

3 Error Modelling

Learner-speech is problematic at best when processed by commercial ASR systems, but its deviations from native speech are systematic, and thus do lend themselves to error modelling.

To deal with these errors, we represented them by a handful of competing subsystems. These subsystems represent different transformation processes such as learner phonology disfluencies, learner cognitive state errors, learner syntax disfluencies, and others. Automating the construction of grammars for the ASR, this model is applied to our base recognition grammar and augments each fluent word or phrase to be detected with a supplemental top *n* most likely disfluent words or phrases.

The problem of extracting errors from learner speech is nontrivial because of the indefinite and overlapping natures of these errors. First, we must deal with the possibility that the error was falsely detected due to uncertainty in the speech recognition. Additionally, we must deal with the fact that multiple kinds of errors can appear the same. Given a disfluent utterance, for example, we do not know if the learner has misremembered the vocabulary, if he is having difficulty with word morphology, or if he is only having problems producing certain sounds. The ASR produces identical surface forms for all three scenarios, so we must disambiguate between the possible mistakes during the full recognition process.

3.1 Phonological Error Modelling

Phonological errors arise due to interference between the first and second language phoneme sets. A statistically driven noisy-channel model was chosen to represent learner phonological mistakes.

The probability values settled upon by this statistical system are equivalent to those that a rule-based system would implement—mistakes occur most often in sounds that are allophones in the L1 but distinct phonemes in the L2, especially over phonological features that carry low functional load in English (gemination and pharyngealization). The benefit of a statistically-driven process is that it allows for the rapid development of a model without the necessity of gathering expert knowledge.

A side effect of the statistical modelling of phonological errors is that it also captures the influences of orthography upon learner phonology. Because TLTS uses the USCPers transliteration system instead of Arabic script, the learner is additionally influenced in his pronunciation by associations gained from experience with his L1 (English) orthography. Examining the phonological disfluencies harvested by the statistical model, we found many errors that were influenced by English letter and letter-sequence pronunciation rules. In English, two 'e' letters together denote a change in vowel tone, while in Arabic, 'ee' denotes a change in utterance length. Non-alphabetic characters in USCPers (such as '$' and '9') posed additional confusion for learners which was also captured by the phonological modelling subsystem.

3.2 Lexical Error Modelling

Language learners tend to confuse words that are "close" to one another. Words that are learned

sequentially (in the same lesson), words that have similar pronunciation (*ra'iib*—Sergeant and *raa'id*—Major), or that have similar meanings (*$meel*—left, and *yemiin*—right), are all prone to learner confusion. To model this, we have a set of subsystems that approximates each of these closeness properties. When the similarity metric for two words in the ASR lexicon is salient enough, the confused word is included in the recognition grammar.

4 Error Recognition in Context

While ASR of learner speech is quite error-prone compared to mainstream native-speaker ASR, our task in TLTS is made easier by the fact that our system doesn't need to provide as explicit feedback to the learner. Precision is the highest priority in traditional ASR systems, we can take advantage of the inherent ambiguity in pedagogical interactions to lessen our need for this precision. That is, our system doesn't need to provide direct feedback to the learner about what it *thought* it heard—it needs only give the learner the type of feedback that a human language tutor would. Thus, we have leeway in that we can offer more vague responses or encouragement as pedagogical feedback when confidence from the speech recognition is lacking.

Additionally, because the ASR system is sitting in a pedagogical framework, we have other resources to pull from in evaluating the user's speech than just the current speech utterance we are processing. The learner's history of committing an error (or his history of avoiding an error-prone speech unit) can alter our confidence that a user has made an error, or which error among a set of possibilities that he has made. The learner's history of making the mistake in question increases our confidence that we were not mistaken in error classification. In contrast, the learner's history of performing a problematic speech unit correctly lowers our confidence that the learner has made a mistake. We find measuring history of success to be just as important as measuring history of failure—often, a learner will attempt to avoid making a mistake by avoiding the associated phrase altogether.

5 Error Detection During Recognition

The primary goal of the recognizer in the MSB setup is to choose the pronunciation variant which best matches the learner speech. However a problem with this setup which surfaced during our trail runs was that learners frequently speak out of domain words--sometimes even in English. If such cases are not filtered out, pronunciation feedback would be generated for these utterances which do not even correspond to the utterance being taught. This turned out to be a major usability issue. We incorporated a hypothesis rejection pre-processing step to take care of this issue.

5.1 Hypothesis rejection

To identify whether the learner is actually speaking the text he is supposed to we compare the likelihood of the best pronunciation variant with likelihood scores from two other recognizers. The first alternate recognizer (*recalt1*) uses a recognizer network containing all the Arabic utterances that are covered by the tactical language training system. The second recognizer (*recalt2*) is an English phonetic recognizer trained on TIMIT. If the difference in likelihood score of the *recalt2* recognizer and the best pronunciation variant is within a certain limit then the utterance is taken as a valid Arabic utterance. Otherwise it is rejected as an English utterance. If an utterance clears the Arabic/English test, then the scores of *recalt1* and the best pronunciation variant are compared. If the *recalt1* score is higher than the best pronunciation variant by a particular value, we reject the utterance as an out of context Arabic utterance which does not match the text the learner is supposed to speak.

5.2 Tolerance to user errors

In many cases it would be beneficial to let the user continue with advanced training material even though the pronunciation of the current utterances being learnt do not match the canonical pronunciation. To achieve this the ASR provides the learner model in the MSB with a score on how the current utterance matches the canonical pronunciation in addition to returning the best pronunciation variant and the score for the best variant. The learner model then uses this information in conjunction with other factors to decide future learning process.

5.3 Pronunciation scoring

In order to provide the user more accurate feedback on his pronunciation we are working towards incorporating pronunciation scoring schemes described in the current literature (Neumeyer et. al 2000) (Teixeira et. al 2000) (S.M. Witt et. al, 1997). At a segmental level the general approach is to first perform a phone segmentation using HMMs trained with target speakers; second, segmental scores are obtained based on log-likelihood indicators, posterior probabilities of phone segments, probabilities of segment duration and timing; finally, scores are combined in order to achieve higher correlation with human ratings. Our

recent work in comparing Hidden Markov Models promises to provide a good way to normalize these phone level scores. By measuring divergence of phone level models across different phonetic classes within native speaker models we can come up with a language dependent measure of tolerance of error for different phones.

We will use the normalized system score to complement our current feedback system based on pronunciation variants and further help in localizing pronunciation errors.

6 Conclusion

In this paper we have presented an approach to provide learners feedback on their spoken language skills in an Arabic Tactical Language Training (TLT) application. For every utterance in our TLT system, we generated non-native pronunciation variations which are dynamically loaded into an Arabic ASR. Likelihood scores and decoding results from ASR are interpreted in context of the learner's current skills and past performance to provide appropriate feedback. To prevent the ASR from providing false feedback in cases where the user has not spoken the correct utterance or used English instead of Arabic we included a hypothesis rejection module which compares HMM likelihoods from an Arabic recognizer, an English recognizer and pronunciation variants to detect whether the user has spoken the right utterance..

The error feedback system has been tested iteratively with subjects. There were problems with false error detection in the first version, but with the current version subjects report that they find it useful, and there have been no detected instances of learner frustration due to erroneous feedback. Further data collection and analysis is ongoing. We are also working on incorporating suprasegmental features such as pitch contours, stress patterns, syllable durations into the learner feedback module.

7 Acknowledgements

The project team includes, in addition to the authors, CARTE members Carole Beal, Ulf Hermjakob, Catherine LaBore, Andrew Marshall, Yousef Nouhi, Dimitra Papachristou, David Pynadath, Mei Si, Gladys Saroyan, Tania Baban, Nadim Daher, Chirag Merchant and Brett Rutland. From the US Military Academy COL Stephen LaRocca, John Morgan and Sherri Bellinger. From the USC Viterbi School of Engineering Naveen Srinivasamurthy, Joe Tepperman and Larry Kite. From Micro Analysis and Design Ursula Lauper and Anna Fowles-Winkler. From the USC School of Education Harold O'Neil, and from UCLA CRESST Eva Baker. This project is part of the DARWARS initiative sponsored by the US Defense Advanced Research Projects Agency.

References

C. Cucchiarini, H. Strik, and L. Boves, "*Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms*", Speech Communication, 30 (2-s3), 109-119, 2000.

H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen. "*Combination of Machine Scores for Automatic Grading of Pronunciation Quality*". Speech Communications, 30(2–3):121–130, Feb 2000.

Shadi Ganjavi, Panayiotis Georgiou and Shrikanth Narayanan, "*ASCII based transcription systems with the Arabic script: The Case of Persian"*, Proc. IEEE ASRU, (St. Thomas, U.S. Virgin Islands), December, 2003

L. Johnson, S. Marsella, N. Mote, M. Si, H. Vilhjalmsson, S. Wu. 2004. Balanced Perception and Action in the Tactical Language Training System. In Proceedings of the InSTIL Symposium on NLP and Speech Technologies in Advanced Language Learning Systems.

L. Neumeyer, H. Franco, V. Digalakis, and M.Weintraub. "*Automatic Scoring of Pronunciation Quality''*. Speech Communications, 30(2–3):83–93, Feb 2000.

H. Printz, P. Olsen, "Theory and practice of acoustic confusability", ASR 2000, pp 77-84

N. Srinivasamurthy, and S. Narayanan, 2003. Language-adaptive Persian speech recognition. in Proc. Eurospeech 2003.

D. Surendran, and P. Niyogi (2003). Measuring the Functional Load of Phonological Contrasts. University of Chicago Technical Report TR-2003-12.

C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sonmez, "*Prosodic features for automatic text independent evaluation of degree of nativeness for language learners*", Proc. ICSLP, 2000.

A. Wachowicz and B. Scott (1999). Software That Listens: It's Not a Question of Whether, It's a Question of How. *CALICO Journal* 16 (3), 253-276.

S.M. Witt and S.J. Young, "*Language Learning based on Non-native Speech Recognition*", Proc. EUROSPEECH '97, pages 633--636, Rhodes, Greece, 1997