Spoken Language Understanding Software for Language Learning

Hassan Alam, Aman Kumar, Fuad Rahman, Rachmat Hartono, Yuliya Tarnikova

BCL Technologies, 990 Linden Drive, Santa Clara, California, USA
(hassana, amank, fuad, rachmat, yuliyaT)@bcltechnologies.com

Abstract

In this paper we describe a preliminary, work-in-progress Spoken Language Understanding Software (SLUS) with tailored feedback options, which uses interactive spoken language interface to teach Iraqi Arabic and culture to second language learners. The SLUS analyzes input speech by the second language learner and grades for correct pronunciation in terms of supra-segmental and rudimentary segmental errors such as missing consonants. We evaluated this software on training data with the help of two native speakers, and found that the software recorded an accuracy of around 70% in law and order domain. For future work, we plan to develop similar systems for multiple languages.

Index Terms: Speech Processing, Speech Recognition, Intonation, Language learning, Support Vector Machine

1. Introduction

In this study we developed a preliminary, work-in-progress Spoken Language Understanding Software (SLUS) with tailored feedback options, which uses interactive spoken language interface to teach Iraqi Arabic and culture. The SLUS analyzes input speech by the second language learner and grades for correct pronunciation in terms of intonation and rudimentary segmental errors such as missing consonants. Arabic language itself has many features that cause difficulties for strategies developed for processing Romance and Germanic languages. Due to the nature of the challenges posed by less-studied languages such as Arabic, the sophistication of computer-based models of Arabic speech, and especially of dialectical speech, has lagged behind that of the European languages. In order to build such a system we developed a comprehensive model of Iraqi Arabic against which the student's performance is measured. This model includes many aspects: (1) an acoustic model; (2) a dictionary or vocabulary model; (3) a grammar model; and (4) a model of common errors or "disfluencies". In traditional (not computer-assisted) instructions, these models take the form of written descriptions and examples of sounds, vocabulary lists, and grammatical rules; and the student's performance in the language is graded by human instructors. For computer-based language instruction, all of these must be cast as explicit databases and mathematical models, so that they can be used to automatically grade student performance, to identify errors, and to evoke appropriate and believable responses from simulated tutors.

In order to test new methodologies for creating Language Models we created a corpus by transcribing and recording the scenarios in both Modern Standard Arabic and in the Iraqi dialect that is most prevalent in central and southern Iraq. Using the test sentences from the corpus and an acoustic analysis software, preliminary prosodic and intonational models (based on Pierrehumbert and Beckman (1988)) were developed for the target language to create training data with acoustic features. We used COTS SRI speech recognition engine (DynaSpeak) for speech-to-text processing. We prototyped and performed (1) evaluation of stress and pitch contours of the input speech, (2) addition of phonetic information to SRI's DynaSpeak, and (3) re-ranking of the ASR output using a Support Vector Machine (SVM). We evaluated this software on training data with the help of two native speakers, and found that the software recorded an accuracy of around 70% in law and order domain. The language models developed could be used to build a prototype tutoring system for multiple languages.

2. Methodology

We integrated Automatic Speech Recognition (ASR) and speech analysis software that recognizes disfluencies in speech that may be exhibited by learners of a second language. A Support Vector Machine (SVM) is used to recognize the "most probable" utterance intended by non-native students with disfluent speech, while an analysis of stress, fundamental frequencies, and phonetic features provides feedback on errors in the student's "accent".

The spoken dialects of Arabic differ from Modern Standard Arabic in many important ways:

- Most of the dialects use subject-verb-object (SVO) word order, whereas classical Arabic and MSA use verb-subject-object (VSO) word order.
- Each dialect, while maintaining much of the MSA vocabulary (with likely variations) includes vocabulary that is unique to the dialect.
- The dialects use a simplified grammar, such as omitting the "dual" form of nouns, and only using the singular and plural.

In addition to the difficulties that come from multiple dialects, the Arabic language itself has many features that cause difficulties for strategies developed for processing Romance and Germanic languages. These include:

- Four of the consonants are differentiated from others only by pharyngealization (a constriction of the throat). Other sounds are formed by glottal stop or as pharyngeal affricates. These sounds are very difficult for Americans and other Western persons to distinguish and create.
- Short vowels are normally not included in written Arabic. Especially in dialectical Arabic, there is little or no agreement on the correct transcription of these short vowels.
- Many language features that are formed in Western languages as "function words" (the, of, to, was, not, etc.) are formed by morphological variations in

Arabic words, including not only prefixes and affixes, but also variations in the middle of words.

- Proper social usage of the language requires adherence to protocols for addressing new acquaintances and gaining their trust.
- Non-verbal cues and intonation are used very differently than in America and Western Europe. Even when words are spoken with the correct pronunciation, their meaning and intent can be misunderstood if these other cues do not appear to be "in sync" with what is spoken. This can lead to a failure to establish trust.

To address the lack of available prosodic models, we developed prototype tools to analyze and model stress and pitch contours. Using the 53 test sentences (details in section 2.1), and the Praat acoustic analysis software (freely available at http://www.fon.hum.uva.nl/praat/, and widely used in the academic and industrial world), preliminary prosodic and intonational models were developed for the target language.

Based on Pierrehumbert's model (1980) (as subsequently modified in Pierrehumbert and Beckman (1988)), we developed a linguistically-driven prosodic and intonational model that related the F0 (fundamental frequency - the acoustic correlate of accent) pattern derived from the syntactic structure of the target language. This model has two level tones, High (H) and Low (L), as primitives employed in types of tonal events: pitch accent and boundary tones. Observations thus far:

- One of the most frequent characteristics concerning interrogative utterances is rising of the pitch of all or part of the utterance whether or not the utterance finishes with a final rise.
- There is no local pitch characteristic for straightforward yes/no questions
- Focalization is manifested by a rising pitch movement, which may spread over the emphasized word

2.1. Iraqi Corpus

We hired native Iraqi Arabic speakers to create Iraqi Arabic corpus. We recorded the voices of several native speakers, as indicated in Table 1. These passages were additionally transcribed phonetically, to provide additional language model data.

	Iraqi Dialect
Total words/sentences	600/53
Voice recordings per passage	1 native male, 1 native female

Table 1. Iraqi Arabic corpus details

2.2. Acoustic Modeling

To address the current lack of prosodic models, we developed a model of lexical stress and boundary tones for Iraqi Arabic. We examined boundary tones at word, phrase, and sentence boundaries. Cantineau (1960) maintains that in Arabic, word stress never plays any distinctive role; however, he did not specifically study Iraqi Arabic. From the generative phonologists' point of view accent plays a distinctive role with some minimal pairs. The accent in a syllable varies based on the length or its phonological weight. The types of rules that we evaluated include:

a) if the last syllable of the word is over heavy (CVVC, CVCC), the syllable gets the accent. (C=consonant, V=vowel)

b) if (a) does not apply and if the penultimate is heavy this syllable gets the accent

c) if (a) and (b) do not apply, the antepenultimate receives the accent

English Translation	Arabic Script	Roman/Phonetic Transcription
Where are you coming from?	من وين جايين؟	**min wen ja-een**

Based on these evaluations, we built a prototype acoustic model for deciphering supra-segmental information for Iraqi Arabic.

2.3. Selection of the ASR engine

As part of our research, we compared several Arabic ASR Engines, such as systems by Sakhr (sakhr.com), AppTek,(apptek.com) Aramedia (Aramedia.com), Sehda (sehda.com), SRI International (Dynaspeak – sri.com). Based on their performances, we narrowed the selection of ASR engines to Apptek's PlainSpeech and SRI International's DynaSpeak. Samples from the 53 test sentences were sent to both organizations. The performance of SRI's Dynaspeak was clearly superior. The correct transcription of the input speech was identified as one of the top two candidates for each of the sentences that was tested. An example of the ASR output is summarized in Figure 1.

Iraqi sentence: من وين جايين؟

Romanized phonetic transcription of Iraqi sentence: min wen ja-iin?

Buckwalter transcription of Iraqi sentence: mIn wyn jAIn

Transcription (Buckwalter Notation)	DynaSpeak Ranking	BCL's SVM Re-ranking
mjn hwin jAn	0	1
mIn wyn jAIn *(correct)*	1	0
mna y$ jHA	2	2
mhIn $y jwAin	3	3

Figure 1. Example Iraqi sentence used to test ASR.

2.4. Enhancement of the ASR engine

Despite the superior performance of DynaSpeak, there are deficiencies for this application that need to be addressed:

- For some sample sentences, DynaSpeak ranked the correct selection second, not first.
- The output is a transliteration of the written Arabic, not a full phonetic transcription.

- The output does not annotate stress or pitch contours.

To improve performance, we included (1) stress and pitch contour information, (2) additional phonetic information to DynaSpeak, and (3) re-ranking of the ASR output using a Support Vector Machine (SVM) algorithm. Figure 2 shows how these enhancements were added to DynaSpeak to provide two confidence scores. Note that these enhancements treat the ASR as a "black box", and other commercial ASR engines and could be ported to multiple languages.

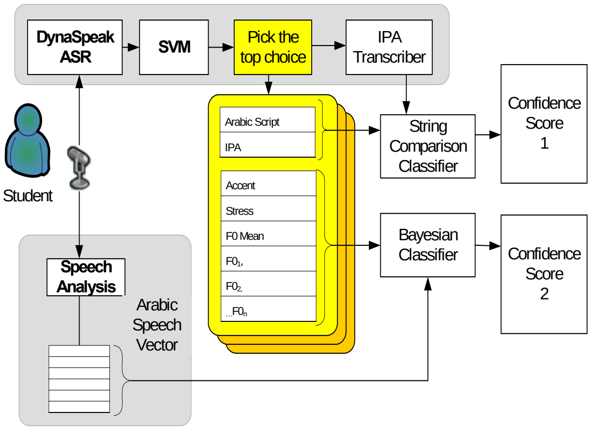

Figure 2. Run-time speech recognition and analysis system.

Below is shown one of the input sentences for a native speaker speaking one of the 53 sample sentences.

Figure 3 shows Praat analysis of the input speech, including , from top down:

- The raw waveform
- The spectrogram, with fundamental pitch (f0)
- The pitch contour

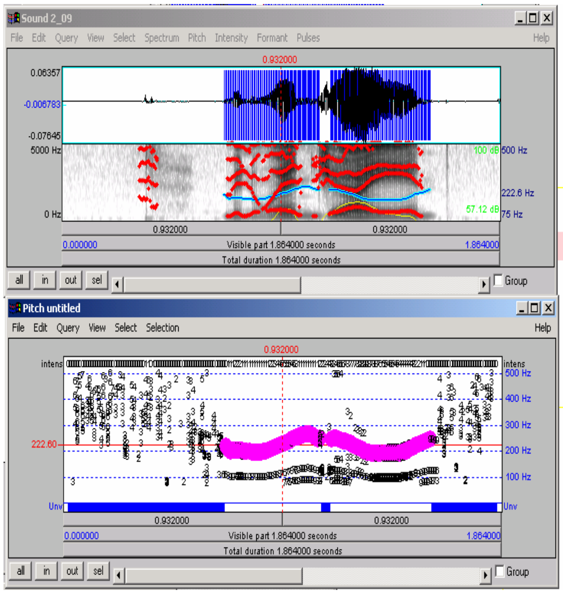

Figure 3. Spectrogram and Pitch Contour created using Praat.

The pitch contour shows the characteristic rising tone of the question. When compared to the reference intonation model, features like this can be used to provide information to the student on the intonation of his/her utterances. Other observations made during this study include (1) there is a local pitch characteristic for straightforward yes/no questions, and (2) focalization is manifested by a rising pitch movement, which may spread over the emphasized word. For this study, phonetic information was provided external to the ASR, by matching the ASR output sentence to its phonetic transcription.

The other enhancement that we prototyped was the re-ranking of the ASR selections using a Support Vector Machine (SVM). BCL's implementation of the SVM combined the signal and the phonetic features. Some of the features we used included:

1. ***$(Score_1 - Score_2) - (Score_2 - Score_3)$*****,** where ($Score_1$ is the highest score, the peak), $score_2$ is the next highest score etc. after normalization.
2. ***$(Score_1 - Score_2)$ / <phonetic difference between 1st and 2nd choice>***

In order to calculate phonetic distance we find arrays of phonemes for both phrases, and then find a mapping of one array to the other, which gives minimal score. The overall score of a mapping is equal to the following:

Σ distances between all pairs of phonemes mapped to each other + 0.5 * all unmapped vowels + all unmapped consonants

The distance between two phonemes is calculated as follows:

- If one of them is vowel, and the other is consonant, then the distance is 1.

- If both same, then the distance is derived from *0.5 * sqrt (sum of squared differences of features)*

The vowel features include the 'Front-Back' and 'Low-High'. The consonant features include the Place of Articulation (POA) and the Manner of Articulation (MOA). Each feature is mapped to a number from [0, 1] interval. For the "front-back" feature, the 'Front' is mapped to 0, and the 'Back' to 1. For the "low-high" feature, the 'Low' is mapped to 0, and the 'High' to 1. For the case of 'POA', the feature 'Voiceless biliabial' is mapped to 0 and the feature 'Voiced glottal' is mapped to 1. Finally, for 'MOA', the feature 'Stop' is mapped to 0, and 'Glide' - to 1. The scoring and mapping process described here, which provide significant enhancement in performance, are totally novel and have not been used in other work identified in our survey of related methods.

3. Integrated Pilot Test

We did a pilot test using a male American English (AE) speaker with limited Arabic training (one year in college). We briefly describe the process here:

1. Input speech using DynaSpeak
2. Speech to Text output of the input speech
3. BCL SVM module processes the output choices. Support Vector Machine (SVM) re-ranking of the DynaSpeak ASR output, based on global measure

of phonetic distance of the top candidates identified by the ASR

4. Praat analyzes the input speech of the student for prosodic features

5. BCL’s Language module compares the values from the speech lexicon built in the training phase and suggests if the input speech is accepted or not

In an example case the student had to speak the given sentence (*min wen ja-een?*) three times and still he did not get the correct pronunciation. The system found the following problem areas:

- Segmental errors (corresponds to missing consonant or vowel)
- Wrong intonation (wrong F0 contour readings, based on the spectral phase envelop on the training data)

4. Conclusions and Result

In this paper we presented a proof-of-concept method that analyzes Iraqi Arabic input speech by a second language learner for correct pronunciation in terms of intonation and rudimentary segmental errors such as missing consonants. We evaluated this software on training data with the help of two native speakers and one AE speaker, and found that the software recorded an overall accuracy of around 70% in law and order domain. Based on the preliminary findings, it is our understanding that the suggested method can be ported to multiple languages.

5. Acknowledgements

This work was partially funded by the US government (DOD) under contract No.: N61339-06-C-0006. This study was done in partnership with Dr. Ed Sims of Vcom3D Inc.

6. References

[1] Cantineau, Jean (1960): Etudes de linguistique arabe, Paris, C. Klincksieck.Encyclopaedia of Islam, 2nd ed., Leiden/London.

[2] Pierrehumbert, J. (1980). The Phonology and Phonetics of English Intonation, Ph.D. dissertation, MIT.

[3] Pierrehumbert, J. and M. Beckman. (1988). Japanese Tone Structure. (Linguistic Monograph Series 15).

[4] Rahman, Fuad., Yuliya Tarnikova, Aman Kumar, Hassan Alam (2004): Second Guessing a Commercial 'Black Box' Classifier by an 'In House' Classifier: Serial Classifier Combination in a Speech Recognition Application. Multiple Classifier Systems 2004: 374-383

[5] Silverman, K., Beckman M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J.. (1992). TOBI: A standard for labeling English prosody. Proceedings of the International Conference on Spoken Language Processing (ICSLP), Banff, Canada.