Detection of foreign speakers' pronunciation errors for second language training - preliminary results

Maxine Eskenazi

206 Cyert Hall, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, Pa. 15213, USA
and LIMSI CNRS, BP. 133, 91403 ORSAY CEDEX, FRANCE

ABSTRACT

With the present generation of speech recognizers, dealing with speaker-independent continuous speech and medium-sized vocabularies, the possibilities of applications become larger. Yet some applications have not yet been tried, or have been tried with heavy constraints on the user, due to expected poor recognition performance. And the lack of results to date in the domain of prosody has severely limited use of that information. Researchers may be overly pessimistic. Herein we explore the possibility of using CMU's SPHINX II recognizer and of obtaining correct prosody information in order to implement it in a system to aid in foreign language learning.

1. INTRODUCTION

The use of an automatic system to help a user improve his accent is appealing for two reasons: first, it affords the user more practice time than a human teacher since it is usually more available, and second, the user is not faced with the sometimes overwhelming problem of human judgment of his production of "foreign" sounds. Speech recognition is a natural choice for this type of application (Bernstein 94). Some attempts have already been made to market systems either based on speech recognition (such as Auralang from AURALOG) or on the detection of changes in prosody (the latter coming from work on teaching the hard-of-hearing to speak, such as Video Voice from Micro Video).

In both cases, the type of speech input the speaker can use is very limited, for example, with the speaker choosing from a multiple choice list. This is often due to low expectations of system performance. Yet recent advances, such as (Rypa 96) show that a speech recognizer could be used in ways that allow speakers more freedom. Tests of what is needed to make foreign speakers more intelligible (Rogers 94) are furnishing concrete goals for an automatic system.

2. USING SPEECH RECOGNITION

In most high quality language courses, a year-one student will learn to pronounce the phones of the target language that are not present in his native language. An automatic system that can correct beginning or intermediate students will eventually take the speaker's original language and the target language into account. But even before going into language specifics, we need to know if present recognition systems are good enough to automatically detect phonetic pronunciation errors.

To study this question, we postulate that if we can hold other major sources of variability, such as ambient noise and linguistic content, constant then the difference between a non-native speaker's incorrect phone and a native speaker's correct one will show up as a significant difference in recognition scores.

Taking English (ESL) as our first target, we collected a database of elicited (Eskenazi 91) American English speech where the speaker was prompted to say a precise rejoinder, such as:

teacher: "You didn't want extra insurance! (I did)"
response: " I did want extra insurance."

in the manner in which several well-known oral teaching methods present syntax exercises.

If you are reading these Proceedings on CDROM, you can judge the accents of the speakers yourself by listening to two examples of the sentences used below, [SOUND A96S1.WAV] for speaker msjh, and [SOUND A96S2.WAV] for speaker mkjp.

Ten native speakers of American English (5 male and 5 female) were recorded as well as 20 foreign speakers (one male and one female speaker from all of the following origins:
French, German, Hebrew, Hindi, Italian, Korean, Mandarin, Portuguese, Russian, Spanish.

It should be noted that the non-native speakers had varying degrees of proficiency in English, and this must be taken onto account when interpreting the results. Fairly fluent speakers will make many less phonetic and prosodic errors than new speakers. In order to palliate these differences, the sentences were phonemically labeled, and we asked expert tutors to listen to the sentences and to note where there was an error, what it was, and how they would correct it. The agreement between automatic detection and tutor is our measure of the quality of error detection.

Figure 1 shows the recognition results for native and non-native male speakers when the speech was processed by SPHINX II in forced alignment mode. As past experience has shown, the use of absolute thresholds on phone scores is of little use. We are therefore looking for a non-native's error to have a score,

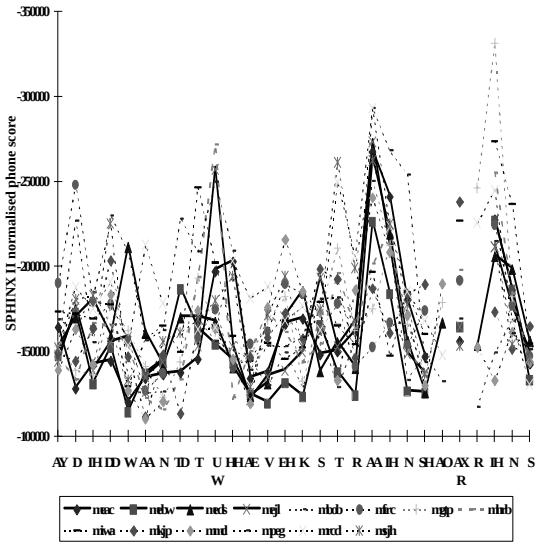

Figure 1: SPHINX II recognition for native and non-native speakers

for a given target phone in a given context, that is significantly distant from the scores that native speakers have. In all graphs, native speakers are shown with solid lines (lines rather than points so the reader can follow the results for a given speaker) and the non-natives have dotted lines. The second letter of the speaker reference indicates the language of origin (see underlined letters in languages above). Phones are on the horizontal axis ("I did want to have extra insurance"). It is to be noted that the following phonological variants were taken into account: for /TD/ of "want" and /AXR/ of "insurance", not all speakers show values since "want" can be pronounced /W AA N/ in this context (geminate); "sur" of "insurance" can be /SH AXR/ or / SH AO R/ (CMU phone notation).

As expected, non-natives are not outliers all of the time, but we can note the following: for /DD/, natives don't release stops here - non-natives did - this is definitely an element contributing to perceived accent, although not one that the tutors had noted as needing correction. This could be considered a minor deviation, that causes a listener to "hear an accent", but not to misunderstand what was said.

For /HH AE V/, many non-natives do not have the /AE/ sound in their native language; they tend to say /EH/, and sometimes /EHF/ instead of /AEV/. For /IH/ of "insurance", differences in vowel quantity are not present in many languages - note that the German speaker is very far off here - it is interesting to note that, compared to scores for native female speakers, the female German speaker shows very similar results. The tutors showed agreement with these outliers.

It would therefore seem, that for this small, yet diverse population, SPHINX II can indeed detect incorrect phones.

3. DETECTING PROSODY ERRORS

But accent does not concern phone errors alone. After year one, pronunciation correction centers almost entirely on prosody. Even when phone targets are not reached, correct prosody guides the flow of speech in a way that affords comprehension. And the measures of what the speaker needs to improve on must be distinct elements that can be practiced and understood. Promising work in this domain includes comparison of language rhythm constructs (Tajima 96) and work on automatic detection of the sentence accent (Sautermeister 96). Yet these studies are language-dependent. While language-dependent knowledge will be essential to the *correction mechanism*, we are looking for a set of detectors that, independently of the target language, can reliably detect the difference between native and non-native speech. This would especially be of value in making it less costly to change the target language of a system.

Before developing *automatic* detectors, we have examined the speech signal to determine whether we can characterise the information that the human tutor uses to detect errors. After examining phone-, syllable-, and word-sized segments, we have developed a set of three measures that agree closely with the human tutors' expertise.

We first measured duration on the speech signal and compared the duration of one voiced segment to the duration of the preceding one ("ration of seg1/seg2" in the figures), making the observations independent of individual variations in speed (and pitch and amplitude).

There are two cases of outliers here that bear mention in Figure 2.: mfrc and miwa have an unusually long "EHK", compared to the following vocalic segment. This can be from a poor attempt at the pronunciation of a lax vowel (not present in their native languages). Mkjp's "to" is longer than his "have", which is quite different from all of the other speakers. This had also been noted by the tutors.

Although it does not concern our present study, it is interesting to observe the extremely small spread of native *and* non-native values at "want"/"to". "Want" is longer than "to" for everyone, with the same relative proportions being respected. This may be due to the fact that prepositions (or non-content words) are not marked

by lengthened duration cues in any language. It would seem that they are simply easy to identify by their short duration compared to the length of the surrounding segments.

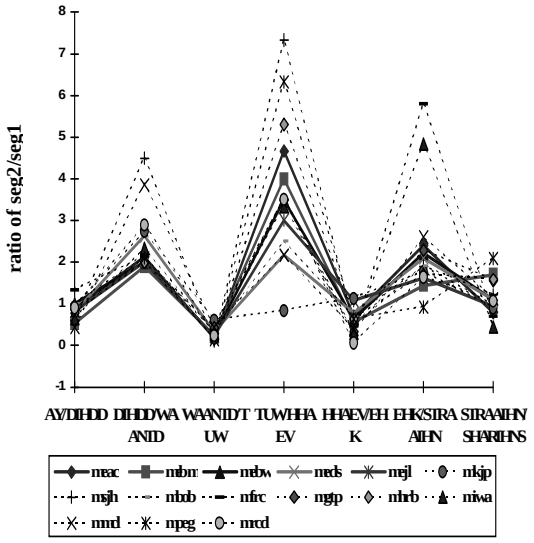

Figure 2: Two-by-two comparison of duration of vocalic segments

Using the vocalic segments we had previously defined, we totaled the number of pitch peaks present in the signal for each segment. Again, we compared results between two neighboring segments.

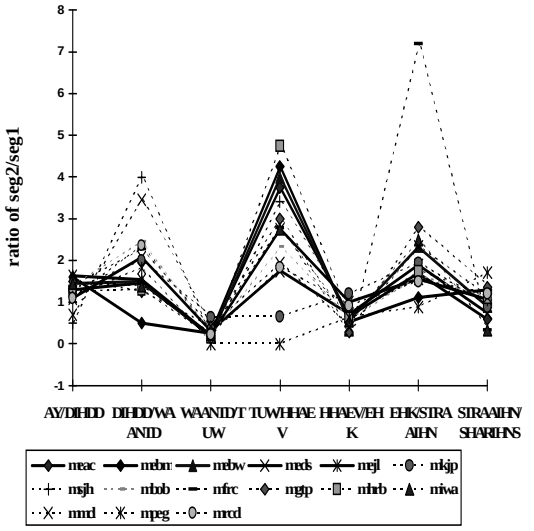

Figure 3: Two-by-two comparison of number of peaks on segment

In Figure 3, we note the following outliers: mfrc raised pitch much higher on "EHK" than on the following vocalic segment - this observation needs to be tempered by the fact that "EHK" is longer (see Figure 2) as well. "Did" for msjh and mmcl is much higher than "want". Mkjp and mpeg show lower pitch for "to", than expected, considering the pitch of "have". Although, for mkjp, we need to relate this to performance on duration as well, it seems that for mpeg, pitch varies independently of duration.

The reduced speaker space at the "want"/"to" point is related to duration and, again, production of non-content words.

Finally, we examined amplitude by taking the average of all of the cepstral C0 values over a given vocalic segment. We then compared them, as before, to the neighboring vocalic segments (Figure 4).

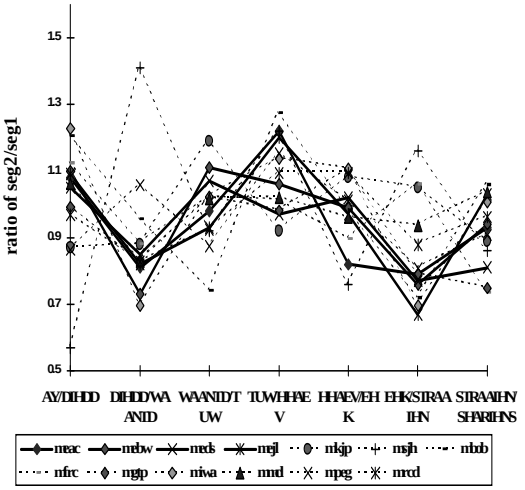

Figure 4: Two-by-two comparison of average amplitude

The general aspect of the curves and of the spead of speaker space is different here, with outliers, such as msjh, having stress displaced within the "I"/"did"/"want" region; mbob displacing stress within "did/"want"/"to", and, among others, msjh within the "ex"/"tra in" region. The speakers' use of changes in amplitude appears to be independent of the two other measures.

4 -CONCLUSIONS

We have shown that speakers' errors in a target language can be detected by an automatic recognizer, for phone errors and that the information is indeed present in the waveform, for prosody. The outliers shown here correspond to areas that human tutors noted as needing improvement.

This gives us a basis on which to build a system capable of giving the user useful feedback on his pronunciation. But that alone will not benefit the user. He must be shown how to produce the correct version, through visual aids, not hearing alone (see, for example, (Strange 94)). The speech cannot be read from a screen. Elicited speech, of the type that we have used here, provides the user with a chance to construct his own sentences while giving the recognizer speech that can be processed by forced alignment.

The speakers will make several mistakes in one sentence, not all related to one another. An effective system will need to choose the most perceptible errors and offer the user help on whichever of these he chooses to work on.

The work herein will be pursued, not only for system construction, but also for its validity on a larger population, and on other target languages.

5. ACKNOWLEDGMENTS

The author would like to thank Raj Reddy, Peggy Heidish, Chengxiang Lu, Lin Chase, and Mosur Ravishankar for their help in the scientific, language tutoring, and technical aspects of the article. This work was partially funded by NSF grant no. IRI9505156.

6: REFERENCES

1. Bernstein, J., “Speech recognition in language education”, Proceedings of CALICO ‘94, p. 37-41.

2. Eskenazi, M., Isard, A., “Characterizing the change from casual to careful style in spontaneous speech”, J. Acoust. Soc. Am., Vol. 89 no. 4. pt. 1, June 1991.

3. Rogers, C., Dalby, J., DeVane, G., “ Intelligibility training for foreign-accented speech: a preliminary study”, J. Acoust. Soc. Am., Vol. 96, no. 4., pt. 2, November 1994.

4. Rypa, M., “Echos: a voice interactive language training system for French”, Proceedings of CALICO ‘96, in press.

5. Sautermeister, P., Lyberg, B., “Detection of sentence accents in a speech recognition system”, J. Acoust. Soc. Am., Vol. 99, no. 4 pt. 2, April 1996, p. 2493.

6. Strange, W., “Speech perception by second language learners”, J. Acoust. Soc. Am., Vol. 99, no. 4 pt. 2, April 1996, p. 2493.

7. Tajima, K., Dalby, J., Port, R., “Foreign-accented rhythm and prosody in reiterant speech”, J. Acoust. Soc. Am., Vol. 99, no. 4 pt. 2, April 1996, p. 2493.

Filename:	ICSLP6.DOC
Directory:	C:\WINWORD
Template:	C:\WINWORD\TEMPLATE\NORMAL.DOT
Title:	
Subject:	
Author:	Maxine
Keywords:	
Comments:	
Creation Date:	04/23/96 10:33 AM
Revision Number:	105
Last Saved On:	05/03/96 2:21 PM
Last Saved By:	Maxine
Total Editing Time:	784 Minutes
Last Printed On:	05/03/96 2:23 PM

As of Last Complete Printing

Number of Pages:	4
Number of Words:	1,903 (approx.)
Number of Characters:	10,851 (approx.)