Using ASR Algorithms for Developing Educational Phonetic Software

R.K. Potapova, M.Yu. Ordin

Department of Applied and Experimental Linguistics
Moscow State Linguistic University
potapova@linguanet.ru
m_ordin@hotmail.com

Abstract

This paper presents ideas, methodology and progress report of the project aimed at developing educational software for pronunciation training. The software is supposed to be used by people studying English as a foreign language. It may turn out handy in distant education courses and for additional phonetic drilling when the human teacher is not available. The system which is being developed at Moscow State Linguistic University (MSLU) pursues three major goals: firstly, detecting and classifying phonetic and phonological errors of users, secondly, assessing user's pronunciation, and thirdly, explaining the errors and formulating recommendations how to correct the errors made by the user during the training session. The work on the described project is conducted at the Department of Applied and Experimental Linguistic of MSLU and is financially supported by the grant of the Ministry of Education of the Russian Federation.

1. Introduction

Educational software applications aimed at increasing phonetic competence may greatly benefit from incorporating speech components which would realize review, assessment and correction procedures. That would enable students to work at pronunciation on their own, because the smart computer systems can take up some functions of the teacher.

As it was stated in [Potapova 2002; 2003], CALL systems have emerged as an alternative to traditional methods of supplementing or replacing direct student – teacher interaction because the integration of sound, voice interaction, text, video, animation has made ir possible to develop interactive learning environments that may enhance traditional model of language learning. But the actual impact of CALL in the field of foreign language education has been significant so far.

Some reasons explaining such an insignificant contribution of educational software are the following: lack of a unified theoretical framework for designing and evaluating CALL systems, absence of evidence for methodological benefits of computers in language learning, technological negligence of educators on the one hand and computer specialists on the other. Intelligent, user-adaptive CALL systems which could perform not as mere diagnostic tools, but also provide feedback mechanisms capable of focussing the learner's attention on the domain that needs practicing [Potapova 2002; 2003].

Educational systems aimed at pronunciation training would greatly benefit from incorporating speech technology. Several attempts have been undertaken which demonstrated how it could be implemented and proved efficiency of such incorporation [Petrushin 2002; Kawai 1999].

To develop speech components for such a system, it is required to find the solution to three tasks.

The first task is to single out errors in pronunciation typical of Russian students speaking English. By the word error we will further mean deviations from the standard realization of sound segments produced by native speakers. All the deviations can be divided into two main groups: phonological and phonetic. Phonological errors are those errors which hinder communication process and lead to misunderstanding. Phonetic errors do not introduce difficulties into communication process, but constitute the set of typical deviations usually called a foreign accent.

The second task to be solved is to invent the appropriate methods for measuring the degree of deviation of the input signal from the sample model stored in the system.

The third task is to develop the module responsible for providing the user with clear instruction on how to correct mistakes and improve pronunciation.

The project thus is summed up to the following: adaptation of the algorithms used in the field of recognition of hearing patterns (that is the algorithms of speech recognition) to be implemented while building speech components for educational linguistic software of phonetic profile. We will need the algorithms realizing various types of speech signal analysis, including feature extraction and parameterization; algorithms for comparing observation vectors (one representing the features of the input signal – user's pronunciation, the other representing the vector of the sample model variant of pronunciation of the same speech segment); language modelling algorithms, etc.

2. Methodology of Developing the System

Each segment of the speech flow can be presented as a bundle of features/ The bundle includes both acoustic and articulatory features, and it seems reasonable to divide the whole bundle into two sets. The acoustic features which constitute different sets describing speech segments differ greatly. The sets of features characterizing sounds (as actual phoneme realizations) are determined by the factors of coarticulation. That is why it is impractical to try to find the correlations between acoustic and articulatory images of the sounds. As it was proved within generative approach to sound system of the Language, the same feature can extend longer than the sound duration or the change of features may occur during pronouncing one and the same sound. The feature may disappear while the sound is still being pronounced or new

feature can add while articulating of the sound is not over yet. So, we believe that it is more reasonable to find the correlations between acoustic and articulatory features instead of trying to match acoustic and articulatory images of the whole phoneme realizations. Besides, it is easier to describe assimilation processes on the level of features, because these are the features which influence each other and cause variations in speech.

Now we can define the methodology of developing the desired educational system more precisely.

The first step is to build a database of the most frequent pronunciation mistakes typical of Russian native speakers learning English,* and the database of speech segments of interfered speech which cause ASR systems to stumble.

The second step while developing such a system then is to present the models as sequences of overlapping features.

The third step is to adapt algorithms of speech recognition (particularly feature extraction and parameterization algorithms) to check which sequences of acoustic features are present in the input signal, and to verify that the degree of explicitness of the acoustic features is approximately equal to that in the model signal.

The fourth step is to single out articulatory features correlating with acoustic ones so that the system could transfer the set of sequences of acoustic features into the set of articulatory gestures.

The fifth step is to build a knowledge base to be used in the module of the system aimed at formulating recommendations how the user should change the pronunciation in order to make the input signal match the model stored in the system.

It should be taken into consideration that the instructions on changing articulation gestures like "make the vowel in the word fronter" or "the tap in this word is supposed to be retroflex" are not very suitable for general audience. In case the system is supposed to be implemented widely (intended to be used not by linguists only, but as a drilling machine ready to be implemented into different educational institutions where phonetics is not focussed on in details and for individual practice), simpler instructions excluding specific terminology is to be worked out. If the user substitutes sound /a/ in the word *"cart"* for the Russian /Y/ or English /^/ like in the word "cut" or for any other sound, the instructions like "make the vowel lower in pitch" will work much better than "make the vowel more retracted" or "lower the tongue and pull it back." Such recommendations will make the system more user-friendly and agreeable with a more diverse audience. This factor should be taken into account while building the knowledge base and developing the module of generating instructions.

3. Preliminary Results

Below some preliminary results of the work on this project are presented.

To study the state of the art of recognising interfered speech, the database of the most frequent errors both on the side of the system and on the side of the user is required. Here we are going to present the part of the database of mistakes made by the system of automatic speech recognition when it was processing foreign speech (this part of the database was built by the students of the Department of Applied and Experimental Linguistics.)

The total number of speakers comprises 40 students (native language – Russian) majoring in Applied Linguistics and studying English professionally at the university level for at least 2,5 years. We tried the programme ViaVoice (IBM) intended to be used as voice dictation software. The publicistic text (a newspaper article) was chosen to read into the computer. At first, a short training session was conducted with each participant to adapt the system to the speakers. To digitise the signal the default settings of the programme and ordinary microphones provided with the licensed versions were used. Then all the participants were asked to read into the computer the same article.

Most errors occurred in one-syllable words. They turned out to be much more difficult to recognise than multi-syllable words. Error rate on the consonants is much higher than the error rate on the vowels.

Phoneme /p/ at the beginning of the word in the position proceeding /I/ was substituted for /t/, and /t/ at the end of the words was substituted for /p/ after unvoiced fricatives and diphthongoids.

Russian speakers tend to make the final voiced consonants unvoiced, and ViaVoice almost always substituted /g/ for /k/ and /®/ for /±/.

Most difficult sounds for the system to recognise turned out plosive consonants, both voiced and unvoiced.

Below we present the diagrammes representing error distribution on the segmental level. The first diagramme represents error distribution on the male speaker sample, the other–on the female speaker sample (see figure 1 and figure 2)

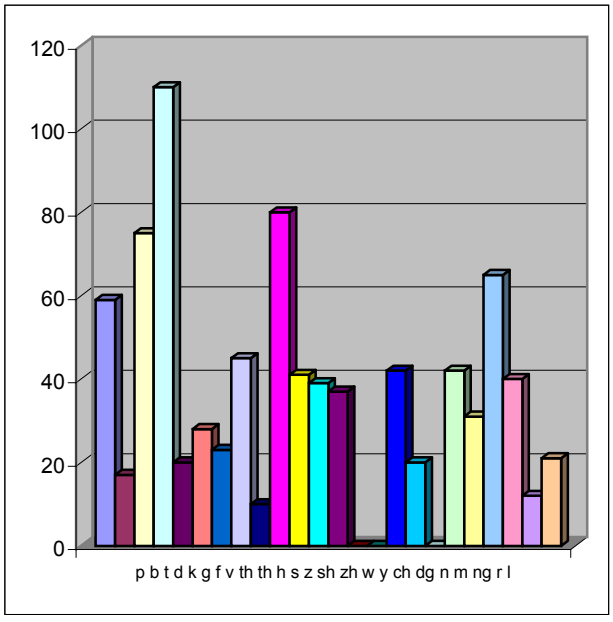

Figure 1: distribution of errors in male speech on segmental level

* The theoretical description of all types of Russian-English interference was given in [Potapov 2003 (a); Potapov 2003 (b)].

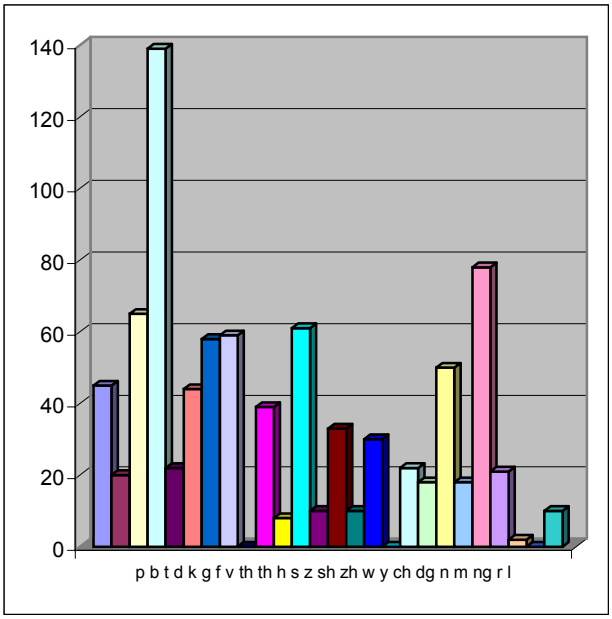

Figure 2: distribution of errors in female speech on segmental level

Errors in recognizing vowels were verified in strong and reduced positions separately. Errors in consonants recognition were verified at the beginning of the word, in the middle of the words and at the end of the words separately. The following tables demonstrate error distribution in vowel and consonant recognition in different texts (of publicistic style, all texts were taken from full-sized British newspapers) read by male (first table) and female (second table) speakers.

Table #1: distribution of errors in male speech on segmental level

	Text #1	Text #2	Text #3	Text #4
Vowels in strong position (as in words:)				
Cut		11	10	9
Path				
Bit	3	13	5	9
Beat	11		42	4
Pot	33			24
Port				
Bird				
Book	8			28
Moon		11		6
Bet	10	9	9	24
Sat	12	5	26	12
Kite	20		11	17
Kate	17		23	33
Cow	20	16	20	28
Low	18	9		11
Fear			20	
Consonants				
p	B 33			
t	M 10	M 4	M 7	B 43 / M 13
d	B 8 / M 34	M 6	B 22 / M 9	B 38 / M 35
k	B 8		B 5 / M	
			9	
g	M 13		M 17	
f	M 15	B 4	E 7	
v	M 40		B 8	
›		B 11		E 1
ð	B 5	B 27	M 21	B 30
h	B 13		B 17	B 15
s	M 17	B 8	M 3	E 13
z	M 2	M 12	M 4	E 17
•				
¥				
w		B 11	B 16	B 21
y				B 22
n	M 8	M 9	B 25 / M 7	M 18
m				M 60

Table #2: distribution of errors in female speech on segmental level

	Text #1	Text #2	Text #3	Text #4
Vowels in strong position (as in words:)				
Cut				
Path	10	5	14	
Bit	6	13	5	9
Beat	12	9	14	10
Pot	7	4	9	7
Port	9	4		15
Bird	10			28
Book				
Moon	11	7	13	18
Bet	7		14	14
Sat	8	3		6
Kite	8	8	11	
Kate	17	13	3	12
Cow	7			11
Low	11	139		14
Fear				
Consonants				
p	B 13	B 3 / M 15		B 11 / M 6
t	M 10	M 14	M 4	B 22 / M 8
d	B 15 / M 34	B 9 / M 11	B 12 / M 5	B 42 / M 14
k	B 12		M 17	
g	M 16		M 22	B 13
f	M 21	B 10	M 12	E 18
v	M 27		B 14	E 21
›				
ð	B 12	B 9	M 11	B 9
h			B 7	B 3
s	E 23	B 13	M 6	M 23
z		M 6	M 1	E 7
•	B 24	M 9		M 4
¥				
w	B 12	B 4	B 3	B 11
y	M 21			
n	M 13	M 23	M 12	M 33
m	M 1		M 5	M 21

Letters in the table stand for the position of the consonant in the word: M denotes that the consonant stands in the middle of the word, B – at the beginning of the word, and E – at the end of the word. The words in the first column denote the vowels that is estimated in the line. The figures in the line stand for the syllabic vowels in the corresponding words. The figures in the tables are given in percents. Percentage was calculated by dividing the total number of mistakes in recognizing a certain error by the total number of phoneme realizations in the texts read during the testing session.

4. Conclusions

Recognizing interfered speech (accented speech, foreign speech) demands additional algorithms to be followed, because xenophones, substitutions, insertions and omissions cause the ASR modules to stumble unless special means of increasing robustness are not undertaken on the stage of construction. Such means include specific methods of adaptation (taking into account that interfered speech is characterized by a significantly wider degree of variability than speech of native speakers), extending the system's vocabulary, employing rules of interference, using hybrid statistical methods based on incorporating neuron networks into Markov chains, etc.

The errors of ASR systems are to a great degree determined by the laws of interference, that is, by the peculiarities of pronunciation of a native language of the speaker which are transferred into the target language. Consequently, speech components of educational system will benefit greatly from taking the interference rules of a pair of languages into account. On the other hand, this will limit the number of potential users of the system to the speakers of a particular language.

Male and Female varieties of speech seem to pose slightly different sets of troubles for ASR system ViaVoice.

Splitting the flow of speech into the sequences of overlapping features seems more reasonable and promising in the domain of developing educational software of phonetic profile than representing the flow of speech as a sequence of bundles of features, because the feature may extend longer than one segment or it may change or be substituted by another feature while uttering of the segment (sound) is not over yet. Besides, it is easier to find correlation between acoustic and articulatory features than between acoustic and articulatory images representing the whole bundles of features (phonemes and phonemic variations).

5. Acknowledgements

We would like to thank the student of the Department of Applied and Experimental Linguistic of MSLU (Moscow) who participated in the experiments and help to build the database of errors. We would like to express special thanks to Polukarov Anton and Polukarova Helen for their help.

The project is supported by Ministry of Education of the Russian Federation. The Head of the Project is R.K. Potapova.

6. References

[1] Petrushin V.A. Student Response for Spoken Language Learning: A Case Study of Learning Chinese Tones. IEEE International Conference on advanced Learning Technologies, 2002.

[2] Potapova R.K. Novije informatsionije tekhnologiji i lingvistika, Moskva, 2002 (in Russian).

[3] Potapova R.K. Modern CALL Systems with Elements of acoustic Feedback. Proceedings of SPECOM 2003, Moscow, 2003.

[4] Potapova R.K., Ordin M.Yu. Articulation models in Educational Software with Embedded ASR components. Proceedings of SPECOM 2003, Moscow, 2003.

[5] Potapov V.V. On language Contrastive-Comparative Analysis of English and Russian Phonetic Systems. Proceedings of SPECOM 2003. Moscow, 2003. (a)

[6] Potapov V.V. The American English Interference in Russian on the Segmental Level. Moscow, 2003. (b)