PHONE-LEVEL SCORING IN COMPUTER AIDED PRONUNCIATION TRAINING FOR TATAR LANGUAGE

Mykola Sazhok[1]*, Aidar Khusainov*[2]*, Valentyna Robeiko*[1]

[1]Int. Research/Training Center for Information Technologies and Systems, Kyiv, Ukraine
[2]Institute of Applied Semiotics of the Academy of Sciences of Republic of Tatarstan, Kazan, RF

ABSTRACT

This work is part of research aimed to develop speech technology for the Tatar language. First results for pronunciation quality automatic assessing are presented. Speakers are allowed for reading a predefined set of words or sentences and the system tries to produce a reasonable score. The pronunciation score for the entire utterance is constructed counting for both the HMM-based log-likelihood acoustic measure and the estimated duration of the phoneme segment. The performance of the proposed algorithms by measuring how well the machine-produced scores correlate with human judgments are evaluated on a speech corpus. Results and further research are discussed.

1. INTRODUCTION

Computer-assisted language learning (CALL) systems is potentially beneficial for both the student and teacher. Typical foreign language instruction courses focus mainly on reading, writing and listening comprehension, much less effort is devoted to teaching correct pronunciation. Among reasons is that it requires more expensive resources, such as extensive individual practice with tutors who are rather natives of the target language. CALL systems are aimed to provide continuous feedback to the student in a self-studying environment. Accurate measuring of student's pronunciation quality is extremely required for CALL system effectiveness during the interactive teaching process in order to enable immediate detection and correction of errors.

The basic pronunciation scoring paradigm uses hidden Markov models (HMMs) to generate phonetic segmentations of the student's speech [1, 2]. From these segmentations, we use the HMMs to obtain spectral match and duration scores. The effectiveness of the different machine scores is evaluated based on their correlation with human grades on a large database.

For Tatar language, an available speech corpus [3] includes records from 251 speakers with the total duration of 7 hours (continuous speech is over 5.5 hours in length). The estimated HMM parameters allowed for 89% recognition accuracy for 1300 isolated words. Therefore, we expect that forced alignment will be satisfactory for phoneme scoring. We also may judge about proper automatic pronunciation grading since all speakers in the corpus are ranked by their pronunciation skills.

2. PRONUNCIATION SCORING

The different pronunciation scoring algorithms studied are all based on phonetic time alignments generated using various HMM-based toolkits [4]. These HMMs have been trained using the database of native speakers. The front-end extracts mel-frequency cepstral coefficients (MFCC). To generate the alignments for the student's speech we must know the text read by the student. From these alignments and statistical models obtained from the native speech, probabilistic scores are derived for the student's speech. The statistical models used to do the scoring are all based on phone units and, as such, no statistics of specific sentences or words are used.

Phone log-posterior probabilities was combined linearly with phoneme duration probabilities since measurements of duration exposes almost no correlation with individual phone quality, which is typical for pinpoint error detection level [5, 6].

A set of context-dependent models along with the HMM phone alignment are used to compute an average posterior probability by the training set. For each segment at time samples $t=t_i:(t_i+l_i-1)$ corresponding to the phoneme q_i , after estimation the frame-based posterior probability $P(q_i|y_t)$ [4], we can evaluate the posterior acoustic score:

$$\hat{a}_i=\frac{1}{l_i}\sum_{t=t_i}^{t_i+l_i-1}\log P(q_i|y_t)\ . \tag{1}$$

The posterior-based score for a whole sentence Q is defined as the average of the individual posterior scores over the N phoneme segments in a sentence:

$$\hat{a}=\frac{1}{N}\sum_{i=1}^{N}\hat{a}_i\ . \tag{2}$$

For each phone q_i we can estimate mean μ_i and variance σ_i of posterior scores, therefore, we approximate the log-based acoustic score r_i for the observed phone segment as

$$r_i = \begin{cases} 1, & \text{if } a_i > \mu_i \\ e^{-\frac{(a_i-\mu_i)^2}{2\sigma_i^2}}, & \text{otherwise.} \end{cases} \quad (3)$$

where a_i is the result of application (1) to the control set.

The procedure to compute the duration-based phone score is as follows. First, from the phoneme-level alignment we measure the phone duration in frames. To obtain the corresponding phone-segment-duration score, the log-probability of the duration is computed using a discrete distribution of durations for the corresponding phone. The discrete duration distributions were previously trained from alignments generated for the native speaker training data.

We did not "normalize" the phoneme duration proceeding from the assumption that most common tempo rates are well presented in training set.

2. DATA AND KNOWLEDGE BASE

The Data and Knowledge base contains a speech corpus, basic phoneme alphabet and letter-to-phoneme conversion means [3].

The speech corpus was accomplished with human pronunciation grades. Pronunciation skills for all 251 speakers were graded by human experts on a scale of 1–5 ranging the categories from 'poor' to 'excellent'. We assume, such a speaker level scoring means that a more skilled speaker pronounced most sentences better, on average. Beside the distribution of human assigned grades, Table 1 exposes significant majority of female, especially among well skilled speakers.

Table 1. Distribution of human assigned grades

Grade	<3	3	4	5
Female speakers	1	11	45	128
All speakers	6	16	64	165
All speakers (%)	2.4	6.4	25.5	65.7

We planned to apply context-dependent phone model. This allowed for reducing the basic phoneme alphabet to 32 units (9 vowels and 23 consonants). More hypothetical consonants are stipulated by the co-articulation, which is approximated with phoneme-triphone model quite accurately.

Less than 30 *find-replace-and-move* rules were constructed for letter-to-phoneme conversion [7], which was used to form the pronunciation vocabulary.

3. EXPERIMENTS

HMM acoustic parameters were estimated for over 8000 physical units, 32 or less Gaussian mixture components were used. The pronunciation scoring module implements estimation for acoustic score (1)–(3) and duration-based score and produces the final machine score after their weighting. The optimal weight was estimated experimentally.

As it follows from Table 1, female speakers make up the majority of 74%. Therefore, we formed the control set only from female speakers ranked below 4 accomplished with 5 speakers graded 4 or 5 that made total 16 test speakers.

In Table 2 we illustrate sentence-by-sentence comparison for two speakers with different human-assigned grades. Here we can see that the system detected better pronunciation skills correctly in 80% of sentences.

Table 2. *Good* and *Average* speaker comparison.

	Score for speakers graded:		Indicator
Sentence Id	*Good* (4)	*Average* (3)	for *Good*
1132	-1.75	-2.17	1
1133	-2.14	-2.09	-1
1134	-1.73	-0.67	-1
1135	-1.91	-2.47	1
1136	-1.99	-4.33	1
1137	-2.16	-2.10	-1
1138	-1.89	-2.34	1
1139	-1.93	-2.52	1
1140	-1.99	-2.62	1
1141	-1.91	-2.05	1
1142	-2.45	-2.23	-1
1143	-2.10	-3.44	1
1144	-2.07	-2.15	1
2008	-2.30	-2.27	-1
2009	-2.88	-4.29	1
2024	-2.37	-1.84	-1
2087	-1.92	-2.14	1
2150	-2.17	-2.36	1
2151	-2.35	-1.92	-1
2152	-3.95	-1.99	-1
2153	-1.68	-1.94	1
2154	-3.32	-2.74	-1
2155	-2.06	-2.06	-1
2156	-2.20	-2.35	1
2157	-2.06	-2.08	1
2158	-1.95	-1.92	-1
2160	-2.42	-2.47	1
2163	-2.62	-2.40	-1
4142	-2.15	-2.65	1
4143	-2.51	-2.20	-1
4144	-2.88	-2.48	-1
4145	-2.27	-3.81	1
4146	-1.92	-3.16	1
4147	-2.08	-2.36	1
4148	-2.12	-2.56	1
Totals:	-78.18	-85.17	7

The results for 16 test speakers are shown in Table 3. We can see that the machine log-based score is greater for speaker with better spoken language skills in most cases.

Table 3. Speaker pronunciation scoring summary.

Speaker Id	Phone segments	Machine score	Human assigned grade
16	1221	-2.31	5
90	1069	-2.53	4
206	1174	-2.64	5
146	1196	-2.66	3
44	1175	-2.68	5
91	1016	-2.69	3
108	1123	-2.78	3
33	1038	-2.78	3
31	1179	-2.80	3
66	964	-2.95	3
6	1362	-2.98	3
135	1271	-3.05	2
141	1343	-3.06	3
5	1254	-3.32	3
13	1221	-3.33	3
8	1065	-3.70	3

Figure 1 illustrates the correspondence between machine-produced pronunciation scores and human assigned grades. Here, the smoothing window length is equal to 4.

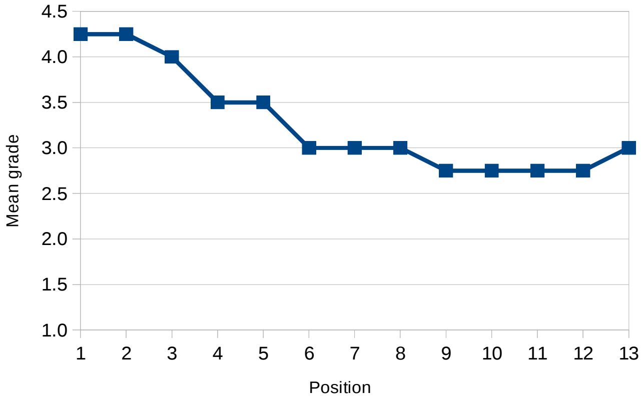

Figure 1. Smoothed human assigned grade by speaker position ordered by the machine-produced score.

The pronunciation training client-server based demonstration system is available at [8]. The user records a phrase proposed by the system and uploads it to the server. After processing, the 10-grade based score is exposed to the user. If the pronounced phrase is irrelevant to the proposed text the zero score is returned.

4. CONCLUSION

The proposed phone-level scoring includes both acoustic and temporal features. The developed CALL system for Tatar pronunciation training demonstrates promising performance.

Future research will be concentrated on extending the analyzed segment to word and entire phrase.

REFERENCES

[1] Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M. Automatic evaluation and training in English pronunciation. In: Proc. ICSLP'1990. Kobe, Japan, pp. 1185-1188.

[2] S. Witt, S. Young. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication, 30((2/3)):95– 108, 2000.

[3] A. Khusainov, J. Suleymanov., M. Sazhok, V. Robeiko. Data and knowledge base development for Tatar speech recognition. In: These of Int. Conference MegaLing'2013. Kyiv, Ukraine. http://megaling.ulif.org.ua/attachments/article/326/sazhok_2013--tatar-asr--megaling.pdf (in Russian).

[4] H. Franco, L. Neumeyer, V. Digalakis, O. Ronen. Combination of machine scores for automatic grading of pronunciation quality. Speech Communication 30 (2000), pp. 121-130.

[5] Y. Kim, H. Franco, L. Neumeyer. Automatic pronunciation scoring of specific phone segments for language instruction. In Proc. Internat. Conf. Acoust. Speech Signal Process., ICASSP 97. Munich, pp. 1471-1474.

[6] M. Peabody. Methods for Pronunciation Assessment in Computer Aided Language Learning. PhD Thesis, MIT, 2011, 176 p.

[7] M. Sazhok, V. Robeiko. Bidirectional Text-To-Pronunciation Conversion with Word Stress Prediction for Ukrainian. In: Proc. All-Ukrainian Int. Conference on Signal/Image Processing and Pattern Recognition, UkrObraz'2012, Kyiv, Ukraine, pp. 43-46.

[8] http://www.cybermova.com/technology/CALL.html