UDC 004.934.2

Informatics and Mathematical Methods in Simulation
Vol. 6 (2016), No. 3, pp. 259-269

PRONUNCIATION QUALITY ASSESSMENT BY COMPARISON WITH SAMPLE

G.A. Dobrovolsky, O.A. Todoriko, N.G. Keberle

Zaporizhzhya National University,
66, Zhukovskogo str., Zaporizhzhya, 69600, Ukraine; e-mail: gen.dobr@gmail.com

The task of pronunciation quality assessment by comparison with a reference example usually requires large training set of such examples. Unfortunately, such sets even for widely used human languages are rare. Most annotated speech corpora contain examples of mispronunciation, without reference utterance examples. In this paper we propose an approach to assess pronunciation quality by comparison with a reference example given small set of reference utterance examples. Dynamic time warping with silence model allows to compare reference utterance by teacher/native speaker with student's utterance and to obtain feature sets describing mispronunciation at word and phone level. Student's utterance is then classified as correct or mispronounced using bagging method.

Keywords: computer-aided pronunciation training, language learning, mispronunciation detection, dynamic time warping, bagging.

Problem statement

Computer-Aided Language Learning (CALL) systems [1-2] have gained new attention nowadays, as speech recognition technologies (SRT) widely used in human-computer interaction with search engines can be adapted to distant language learning. Computer-Aided Pronunciation Training (CAPT) systems respond to the demand of SRT client to be understood. There are various technologies to teach reading, listening, and grammar, to improve and expand vocabulary. At the same time, oral speech and correct pronunciation training are harder to automate, and are more to the research, than to the technology, however several pronunciation assessment services already exist [3-4].

The straightforward way to assess pronunciation is to use automatic speech recognition (ASR) system. Current ASR systems are based on supervised machine learning techniques. Training of ASR system requires a large corpus of annotated (manually/automatically) reference data – audio files storing sound of a phoneme/word/phrase/text utterance of a person in a given language. Such a prerequisite causes a bottleneck of direct adoption of ASR system to pronunciation assessment – necessary datasets are only available for the most used languages [5-6], whereas there are 7102 languages spoken in the world [7]. One more bottleneck of ASR system adoption is the vocabulary used. Sufficient datasets are available only for the most common, everyday topics (e.g. British English corpus WJSCAM0 [8] for news). Specific terminology words, professional slang, rare vocabulary words will be substituted by similarly sounding words.

Therefore, there is a need of exploring alternative approaches that do not require large reference data, and do not perform extra operations, e.g. do not perform full ASR.

Related Work

At the early stages, pronunciation quality assessment was performed for the whole phrase with the help of hidden Markov model. Obtained results did not depend on a teacher,

but did not point to the error type [10-12]. To overcome this difficulty the researchers focused on various ways of detection of "problematic" phonemes extracted from utterance examples, and their classification as pronounced correctly or mispronounced [13-16]. The results of such approach have shown increased precision of pronunciation assessment. Approaches to extend ASR system with typical pronunciation errors [17] lead to increased quality of assessment. However, they require a-priori sets of typical pronunciation errors, inherent to language learners of different nationalities. As a result, only those typical errors could be assessed, i.e. person-specific utterances within the same nationality are not taken into account.

Recently, comparison-based approaches to mispronunciation detection [9], [18] appear, attempting to avoid usage of a full ASR system. They differ in the way how classification is performed, and how feature sets of utterances are obtained. In [9] SVMs are used for classification, and utterance feature sets extracted with Gaussian posteriograms (GP) and Mel frequency cepstral coefficients (MFCC) are compared. In [18] classification is done with Gaussian mixed models (GMMs), and deep neural networks (DNNs) are used for extraction of feature sets.

Aim of the paper

In this paper, we propose to use bootstrap aggregating (bagging) algorithm to improve classification of example utterances, taming the problem of small reference datasets. The approach is inspired by the previous success in application of dynamic time warping (DTW) with silence model [9] to mispronunciation detection. However, in [9] support vector machines (SVM) are used in classification of example utterances, which require a large reference dataset for classifier training. Bagging algorithm allows starting pronunciation assessment with a small reference dataset, incrementally adding new references. Such an environment is inherent to a socially-oriented on-line language learning system, where teachers/native speakers can add their utterances of sample phrases, and the system reclassifies students' pronunciation accordingly.

Results

Mispronunciation quality assessment simplified method is based on the following assumptions:

- if a phrase uttered by a student similar to a phrase uttered by a teacher, then the student has a good pronunciation;
- similarity criterion is a distance function between correspondent features' values of conditional phonemes utterance by teacher and student;
- uttered phrase is split into conditional phonemes in assumption that features of the sound change essentially between different conditional phonemes, rather than inside one conditional phoneme;
- silence and pauses between words are not taken into account.

Claimed that student's pronunciation is well-trained if his/her phrase is similar to a teacher phrase. This allows at the beginning only a small set of teacher sample utterances. The benefit of such an approach is its simplicity, incremental pronunciation quality assessment improvement as more correctly pronounced samples (e.g. by students) are put into a sample set.

Sound file preparation stage

Sound file preparation stage is traditional for speech recognition (see Fig.1). First, low-frequency component is removed as not important for speech recognition by means of signal smoothing:

$$x_k = \alpha \cdot x_k + (1-\alpha) \cdot x_{k\text{-}1}, \ |\alpha| < 1, \tag{1}$$

where α - is a parameter, regulating the level of smoothing.

Then, the signal amplitude is mapped to the segment [-1, 1] and the signal is split into frames, F (see Fig. 1). Frames are overlapping fragments of the sound file, having length depending on the frequency of the sound. In our case, as sound was recorded at 22 kHz, and fast Fourier transformation requires 2^n discrete signal values in a frame, frame length was 23 ms (512 values), and overlapping window was 11 ms (256 values).

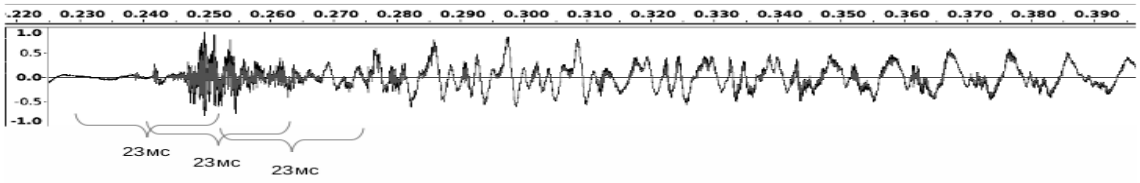

Fig. 1. Dependency of a sound signal on time and frame size explanation

For each frame t of F we calculated MFCC [19] feature set, and additionally energy, entropy, and their first and second derivatives, resulting in a feature set f_t of 42 features.

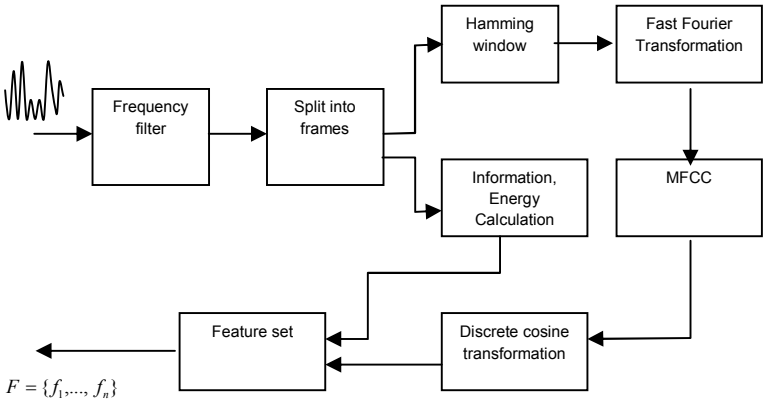

Fig. 2. Sound file preparation steps

For a frame t its energy is evaluated as biased estimate of the variance of the input signal:

$$E_t = \frac{1}{N}\sum_{k=0}^{N-1}\left(x_k - \bar{x}_t\right)^2, t = \overline{1,T}, \tag{2}$$

where $\bar{x}_t$ - is an average value of signal in a frame t, N – quantity of amplitude values in a frame, t – frame number, T – total quantity of frames.

To evaluate entropy we obtain amplitude sweep $[a_{\min}, a_{\max}]$, the resulted segment is split into R parts $[a_0, a_1]$, $[a_1, a_2]$, … , $[a_{R-1}, a_R]$, where $a_0 = a_{\min}$ and $a_R = a_{\max}$, and for each

frame we calculate the quantity of amplitudes, belonging to the segment and obtain frequency histogram. Then, using Shannon's definition of information entropy, we obtain:

$$I = -\sum_{i=1}^{R} p_i \ln(p_i), \tag{3}$$

where p_i – is a signal amplitude share, belonging to the segment $[a_{i-1}, a_i]$.

Usage of Mel Frequency Cepstral Coefficients (MFCC) is one of the standard techniques to obtain features of a sound in ASR systems [19]. MFCC features are obtained with the help of a set of frequency filters, taking into account the peculiarity of a human ear to have different sensibility in different parts of the audio spectrum – almost linear for frequencies below 1 kHz and logarithmic for higher frequencies.

At the first step we calculate signal energy logarithm upon application of each filter

$$S(t,m) = \ln\left(\sum_{n=0}^{N-1} \left|X(t,n)\right|^2 H(m,n)\right), t = \overline{1,T}, m = \overline{0,M-1}, \tag{4}$$

where $X(t,n)$ – is a n-th component of Fourier image in the frame t, $H(m,n)$ – is a n-th component of m-th Mel-Frequency filter, N – window size, M – predefined quantity of Mel filters, T – quantity of frames. Usually in ASR systems $M = 20$, but $M = 12$ is also acceptable.

At the second step we perform discrete cosine transformation of $S(t,m)$ values:

$$c(t,m) = \sum_{m_1=0}^{M-1} S(t,m_1) \cos\left(\frac{m(m_1 - 0{,}5)\pi}{M}\right), t = \overline{1,T}, m = \overline{0,M-1}, \tag{5}$$

We also calculate first and second derivatives to take into account human ear reaction to the spectrum changes in time:

$$\begin{aligned} dc(t,m) &= c(t+2,m) - c(t-2,m), \\ d^2c(t,m) &= c(t+1,m) - c(t-1,m). \end{aligned} \tag{6}$$

The same derivatives are calculated for energy E and information entropy I as well.

Values (2), (3), (5), (6) form a feature set f_t for each frame t, resulting in a feature set of 42 features

$$f_t = \left\langle c(t,m), dc(t,m), d^2c(t,m), E, dE, d^2E, I, dI, d^2I \right\rangle. \tag{7}$$

Preparation of samples

To detect silence we seek frames with minimal information entropy values [20] that are considered as noise. Frames contain informative speech, if its Mahalanobis distance to any of frames considered as noise exceeds a given threshold [21].

Sequence of frames, *F*, is then separated into conditional phonemes, by pair wise comparison of Euclidean distances between correspondent MFCC values of each two neighbor frames $f_t, f_{t=1}$. We assume that sound characteristics change essentially between two different conditional phonemes, rather than within the same phoneme. To calculate Euclidean distances we use MFCC features of the same nature (energies, frequencies etc). Conditional

phonemes set may not coincide with the traditional sound set of the language, and for each specific phrase may differ.

Separation of a sample phrase into words may be performed manually or with the help of some ASR system.

Comparison with sample

After sample and student utterances are prepared as shown in Fig.3, DTW algorithm is used to align two frame sets (see Fig.3).

Given sample $FT = \{ft_1,...,ft_n\}$ and student's $FS = \{fs_1,...fs_m\}$ frame sets, DTW distance matrix Φ is constructed as

$$\Phi(i,j) = D(ft_i, fs_j), i = 1..n, j = 1..m,$$

where D – is Euclidean distance between sample/student frames.

As student utterance is uncertain, with pauses, we use DTW with modified distance function, taking silence frames into account, as in [9].

Silence vector ϕ_{sil} keeps average distances from each frame of FS to each frame of FT, marked as silence,

$$\phi_{\text{sil}}(j) = \frac{1}{r}\sum_{k=1}^{r}\Phi(k,j),$$

where r – is a quantity of frames in FT marked as silence.

Modified distance matrix is then obtained as

$$\Phi'(i,j) = \begin{cases} \min(\Phi(i,j), \phi_{sil}(j)), & i \in B \\ \Phi(i,j), & i \notin B \end{cases},$$

where $\phi_{sil}(j)$ – average distance between *j*-th frame of *FS* and frames of *FT*, marked as silence, *i* – sample frame index, *j* – student frame index, *B* – set of sample frames, where student can (or is allowed to) make a pause.

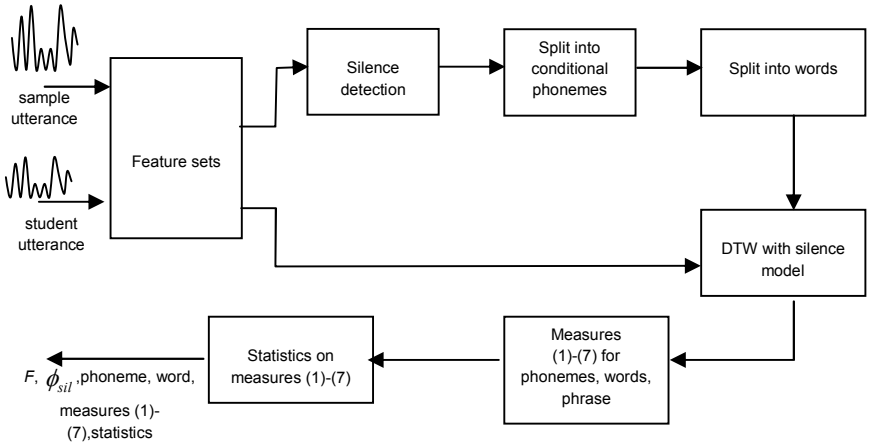

Fig. 3. Comparison steps

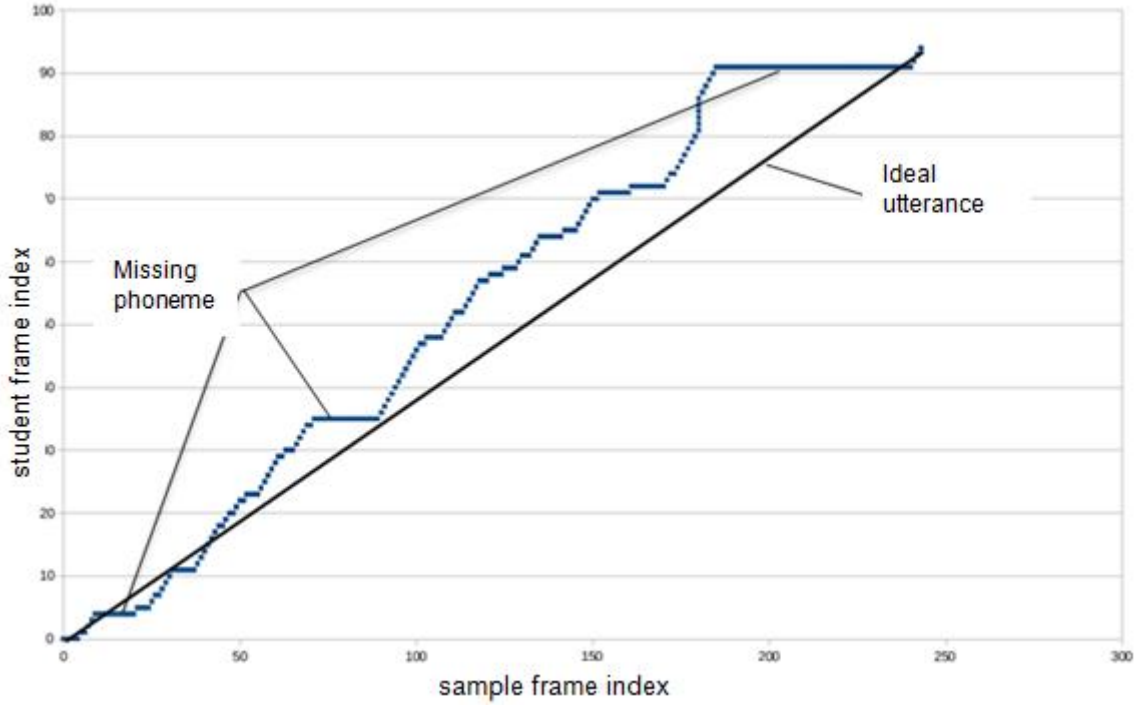

Fig. 4. Sample (ideal) and student utterances of a phrase “This woman has got a good dress”

Given $t_{t\min}, t_{t\max}$ – begin/end indexes of frames of a particular conditional phoneme of sample utterance, $t_{s\min}, t_{s\max}$ – of student utterance, we obtain the following set of measures:

– max/min indexes of student frames t_s given the index of sample frame t_t

$$\begin{aligned} s_0(t_t) &= \min(t_s \mid t_t), \\ s_1(t_t) &= \max(t_s \mid t_t); \end{aligned} \tag{8}$$

– max/min indexes of sample frames t_t given the index of student frame t_s

$$\begin{aligned} t_0(t_s) &= \min(t_t \mid t_s), \\ t_1(t_s) &= \max(t_t \mid t_s); \end{aligned} \tag{9}$$

– average angle of a slope of the graph of a linear function (see Fig. 4)

$$K = \frac{t_{s\max} - t_{s\min}}{t_{t\max} - t_{t\min}}; \tag{10}$$

– deviation from the graph of a linear function

$$C = \sum_{t=t_{t\min}}^{t_{t\max}} \max\left(\left|s_0(t) - t \cdot K\right|, \left|s_1(t) - t \cdot K\right|\right), \tag{11}$$

– maximal deviation from the graph of a linear function

$$D = \max_{t_{t\min} \le t \le t_{t\max}} \left(\max\left(\left|s_0(t) - t \cdot K\right|, \left|s_1(t) - t \cdot K\right|\right)\right), \tag{12}$$

– maximal quantity of student frames correspondent to one sample frame

$$S = \max(s_1(t) - s_0(t)), t_{t_{\min}} \le t \le t_{t_{\max}}; \quad (13)$$

– maximal quantity of sample frames correspondent to one student frame

$$S = \max(t_1(s) - t_0(s)), t_{s_{\min}} \le s \le t_{s_{\max}}. \quad (14)$$

Enlisted measures aim at evaluation of the pronunciation speed and duration of a phoneme utterance.

To measure similarity of utterance of two phonemes of the same length (in frames) we used Euclidean distance between each pair of phonemes $R_1 = \sum_{t=t_{t_{\min}}}^{t_{t_{\max}}} \sum_{s=s_0(t)}^{s_1(t)} \Phi'(t,s)$, and $R_2 = \sum_{t=t_{t_{\min}}}^{t_{t_{\max}}} \Phi'(t, t \cdot K)$.

Classification

Given a feature set f_t and measures set (8)-(14), it is possible to classify student utterance as correct or mispronounced. Classification task was formulated as follows: given a small set of sample utterances and an example utterance, obtain pronunciation quality as similarity measure. Each sound file is presented as two-dimensional array: $F = \{f_t\}, t = \overline{1,T}$, where t – is a frame number, f_t – set of 42 features (7), calculated for the frame t.

To compare different recordings, their durations were equalized with DTW, hence all the sounds were presented as two-dimensional arrays of the same size.

Let classifier be presented with an unknown function $h : F \to \{-1,+1\}$, where "-1" and "+1" are classes correspondent to "mispronounced" and "pronounced correctly". Function h is selected such that if $h(F) < 0$, an example utterance is considered as mispronounced, and if $h(F) > 0$ – as pronounced correctly.

The main problem for mispronunciation detection task was the small set of samples. Most machine learning techniques (Bayesian classifiers, neural networks, hidden Markov's models) require large training sets. Usage of small training sets leads to overfitting problem – classifier simply stores the whole training set, without learning and generalizing, so even slight modification of the sample leads to errors. However, modifications are unavoidable, because sample utterances can be recorded by different people, having different recording devices.

Therefore, only simple classifiers, such as support vector machines (SVM) dealing with small training sets, can be applied. SVM classification technique seeks for hyper plane separating two clusters in multidimensional space, where the most important points are the closest to the borders of clusters. However, in our task it is unknown if such a hyper plane exists, because it is possible, that there is not a plane, but a complex surface (parabolic etc.). We conducted a set of experiments with SVM, as in [9], but on a small training set, and obtained unsatisfactory results.

Hence, we concluded at selection of machine learning ensemble meta-algorithm Bootstrap Aggregating (Bagging) [22].

Bagging

As the sample set is relatively small, classifiers like SVM cannot be used due to large training set required, we propose to use bootstrap aggregating, or bagging algorithm [22] to generate training set for classifier.

The main idea of bagging is to create an ensemble of simple classifiers, each of which is trained on a randomly selected training subset

$$h_q(F, z_q), q = \overline{1, Q},$$

where Q – number of classifiers, z_q – some adjustable parameters, F – audio file feature set. Q is either predefined or adjusted depending on the training results.

After training, we obtain a set of h_q, on average behaving as a $h(F)$ we seek for, and the resulting classifier is averaging all the h_q:

$$h(F) = \frac{1}{Q}\sum_{q=1}^{Q} sign(h_q(F, z_q)),$$

that is a value of comparison between a sample and an example.

Training set construction for bagging

To create training sets three consequent random generators were used. First generator selected a feature index m_q from the feature set f_m (integer from 1 to 42), second – a moment of time t_q. Third generator worked several times – it selected indexes of elements from the set of all utterances, both sample and students', $\{l_{qi}\}, i = \overline{1, I}$.

Training subset is a set of pairs

$$(F_{t_q, m_q}[l_{qi}], class[l_{qi}])$$

where $class[l_{qi}] = 1$, if $F_{t_q, m_q}[l_{qi}]$ is correctly pronounced, $class[l_{qi}] = -1$, if $F_{t_q, m_q}[l_{qi}]$ is mispronounced.

As functions $h_q(F[l_{qi}], z_q)$ we selected linear functions

$$h_q(F[l_{qi}], z_q) = F_{t_q, m_q}[l_{qi}] + z_q, \quad (15)$$

where $F_{t_q, m_q}[l_{qi}]$ is a real number – the value of m_q for feature set f_{t_q}, for l_{qi}-th utterance, z_q - some real number.

To train each classifier $h_q(F, z_q)$ it is necessary to find z_q, minimizing the error

$$ERR_q = \sum_{i=1}^{I} \left| class[l_{qi}] - sign\left(h_q\left(F_{t_q, m_q}[l_{qi}], z_q\right)\right)\right|$$

Classifiers (8) are simple, easy to create and to train. For each classifier $h_q(F, z_q)$ calculated is the frequency of errors, and the most precise classifiers remain, others are

removed. The selection assumes each classifier decided its dominant class, "-1" or "+1", and then class number is averaged.

The benefits of bagging are: there is no overfitting, adding "noise" is a step of classifiers creation; best features are selected automatically at the classifiers selection stage; rather complex surfaces, not just planes in the feature space, can be dealt with bagging.

Experiments

To assess pronunciation quality, calibration of results is needed. To calibrate the system, we use small additional set of utterances by students with good pronunciation grades, confirmed by a teacher. This additional set was used to obtain minimum and maximum permissible values of each feature.

Phoneme or word considered as mispronounced if any of measures (8)-(14) go beyond permissible values. A phrase is considered as mispronounced if any of phonemes or words was mispronounced.

Examples of pronunciation quality assessment are shown in Table 1, where "-f" – female student, "-m" – male student.

Table 1.

Grades of pronunciation quality assessment for the phrase "This woman has got a good dress"

Example	Worst word grade	Worst conditional phoneme	Expert grade
students with good pronunciation grades			
2-f	0.250	0.250	good
3-f	0.250	0.250	good
4-f	0.250	0.389	good
5-f	0.250	0.250	good
6-f	0.250	0.250	good
7-m	0.250	0.250	good
8-m	0.250	0.250	good
9-m	0.250	0.250	good
students			
06-f	1.499	1.033	weak accent
00-m	1.887	3.764	strong accent
05-f	3.949	6.386	strong accent
07-f	3.748	2.499	strong accent
08-f	5.936	6.440	strong accent
08-f	4.343	6.814	strong accent
09-m	2.017	3.976	strong accent
01-f	2.191	9.836	strong accent
03-f	5.247	3.582	missed word
04-f	3.106	7.746	other phrase
10-f	25.487	25.487	one word instead of phrase
11-m	2.666	2.666	one word instead of phrase

Concluding remarks

The paper discusses the possibility to adopt known algorithms, used in ASR systems, to a comparison-based CAPT system. The proposed combination is MFCC-based sound feature set, DTW with silence model and bagging for creation/training pronunciation classifiers given a small sample set. Training is performed for each sample utterance separately, and allows for a small sample set. Adding a new sample does not require the whole system rebuilding, hence the solution is scalable.

Proposed approach evaluates both correctness of pronunciation and duration/number of phonemes. To define proper pronunciation a small training set is enough – nearly 10 samples of each phrase, uttered by different voices and at different rate of speech.

Directions of future work are seen as follows. First, to compare the quality of results on other corpora possessing both sample and student utterances. Second, to apply other classifier types that are tolerant to small sample sets.

References

1. Witt, S.M. Automatic Error Detection in Pronunciation Training: Where we are and where we need to go / S.M. Witt // Proc. IS ADEPT. – 2012. – Vol. 1. – PP. 1-8.
2. Lohiya, S.V. Survey on Computer Aided Language Learning using automatic accent assessment techniques / S.V. Lohiya, M.V. Kamble // Proc. ICPC. – 2015. – PP. 1-4.
3. Duolingo [Electronic resource]. – Access to resource: https://www.duolingo.com
4. Englishtown [Electronic resource]. – Access to resource: http://www.englishtown.com
5. Nuance Recognizer [Electronic resource]. – Access to resource: http://www.nuance.com/for-business/by-solution/customer-service-solutions/solutions-services/inbound-solutions/self-service-automation/recognizer/recognizer-languages/index.htm
6. Google voice search [Electronic resource]. – Access to resource: https://en.wikipedia.org/wiki/Google_Voice_Search
7. Ethnologue: Languages of the World [Electronic resource]. – Access to resource: http://www.ethnologue.com/
8. Robinson, T. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition / T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals // Proc. ICASSP 1995. – IEEE Computer Society. – 1995. – PP. 81-84.
9. Lee, A.A. Comparison-based Approach to Mispronunciation Detection / A. Lee, J. Glass // Proc. SLT Workshop 2012. – IEEE. – 2012. – PP. 382-387.
10. Eskenazi, M. An overview of spoken language technology for education / M. Eskenazi // Speech Communication. – 2009. – 51(10). – PP. 832-844.
11. Delmonte, R. Exploring Speech Technologies for Language Learning, Speech and Language Technologies / R. Delmonte // InTech. – 2011. – PP.71-105.
12. Levis, J. Computer technology in teaching and researching pronunciation / J. Levis // Annual Review of Applied Linguistics. – 2008. – Volume 27. – PP.184-202.
13. Franco, H. Automatic detection of phone-level mispronunciation for language learning / H. Franco, L. Neumeyer, M. Ramos, H. Bratt // Proc. Eurospeech 99. – ICSA. – 1999. – PP. 851-854.
14. Witt, S.M. Phone-level pronunciation scoring and assessment for interactive language learning / S.M. Witt, S. Young // Speech Communication. – 2000. – 30(2-3). – PP. 95-108.
15. Yoon, S.-Y. Landmark-based Automated Pronunciation Error Detection / S.-Y. Yoon, M. Hasegawa-Johnson, R. Sproat // Proc. Interspeech. – ISCA. – 2010. – PP. 614-617.
16. Ai, R. Automatic Pronunciation Error Detection and Feedback Generation for CALL Applications / R. Ai // Lecture Notes in Computer Science. – 2015. – Volume 9192. – PP. 175-186.
17. Harrison, A.M. Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer / A.M. Harrison, W.Y. Lau, H.M. Meng, L. Wang // Proc. Interspeech. – ISCA. – 2008. – PP. 2787-2790.
18. Nicolao, M. Automatic assessment of English learner pronunciation using discriminative classifiers / M. Nicolao, A.V. Beeston, T. Hain // Proc. ICASSP 2015. – IEEE. – 2015. – PP. 5351-5355.
19. Mida, L. Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques / L. Muda, M. Begam, I. Elamvazuthi // Journal of Computing. – 2010. – 2(3). – PP. 138-143.

20. Asgari, M. Voice Activity Detection Using Entropy in Spectrum Domain / M. Asgari, A. Sayadian, M. Farhadloo // Proc. Telecommunication Networks and Applications. – IEEE. – 2008. – PP. 407-410.
21. Dobrovolsky, G.A. Application of Shannon entropy for voice activity detection in noisy sound recordings (in Russian) / G.A. Dobrovolsky, O.O. Todoriko // Herald of Kherson National Technical University. – 3(58). – 2016 (in press).
22. Breiman, L. Bagging Predictors / L. Breiman // Machine Learning. – 1996. – 24(2). – PP. 123-140.

ОЦІНКА ЯКОСТІ ВИМОВИ МЕТОДОМ ПОРІВНЯННЯ З ЕТАЛОНОМ

Г.А. Добровольський, О.О. Тодоріко, Н.Г. Кеберле

Запорізький національний університет
вул. Жуковського, 66, м. Запоріжжя, 69600, Україна; e-mail: gen.dobr@gmail.com

Задача оцінки якості вимови за допомогою порівняння з еталонною вимовою зазвичай потребує великої кількості еталонів. На жаль, підібрати необхідну кількість еталонів навіть для розповсюдженої мови важко, оскільки переважна більшість анотованих корпусів містить лише набори прикладів некоректної вимови, без еталонних прикладів. У даній статті запропоновано один підхід до оцінки якості вимови методом порівняння з еталоном в умовах невеликої кількості еталонних вимов. Метод DTW з урахуванням тиші дозволяє співставити еталонну фразу, яку вимовив вчитель/носій мови, із фразою учня, та отримати набір властивостей вимови рівня слова і фонеми. На цьому наборі властивостей виконується класифікація фрази як коректно/некоректно вимовленої за допомогою методу bagging, який не потребує великої кількості еталонів для навчання.

Ключові слова: комп'ютеризоване навчання вимові, вивчення мови, визначення помилок у вимові, dynamic time warping, bagging

ОЦЕНКА КАЧЕСТВА ПРОИЗНОШЕНИЯ МЕТОДОМ СРАВНЕНИЯ С ЭТАЛОНОМ

Г.А. Добровольский, О.А. Тодорико, Н.Г. Кеберле

Запорожский национальный университет
ул. Жуковского, 66, г. Запорожье, 69600, Украина; e-mail: gen.dobr@gmail.com

Задача оценки качества произношения путем сравнения с эталонным произношением обычно требует большого количества эталонов. К сожалению, подобрать нужное количество эталонов даже для широко распространенных языков трудно, подавляющее большинство аннотированных корпусов содержат лишь наборы примеров неправильного произношения, но не эталонные примеры. В данной статье предлагается один подход к оценке качества произношения методом сравнения с эталоном при условии небольшого количества эталонов. Dynamic Time Warping с учетом тишины позволяет сопоставить эталонную фразу, произнесенную учителем/носителем языка, с фразой ученика, и получить набор свойств произношения уровня слова и фонемы. На основании этого набора свойств выполняется классификация фразы как правильно/неправильно произнесенной с помощью метода bagging, который не требует большого количества эталонов для обучения.

Ключевые слова: компьютеризированное обучение произношению, изучение языка, определение ошибок в произношении, dynamic time warping, bagging