Phoneme Level Non-Native Pronunciation Analysis by an Auditory Model-based Native Assessment Scheme

Christos Koniaris and Olov Engwall

Centre for Speech Technology, School of Computer Science & Communication
KTH - Royal Institute of Technology, Stockholm, Sweden
[koniaris, engwall]@kth.se

Abstract

We introduce a general method for automatic diagnostic evaluation of the pronunciation of individual non-native speakers based on a model of the human auditory system trained with native data stimuli. For each phoneme class, the Euclidean geometry similarity between the native perceptual domain and the non-native speech power spectrum domain is measured. The problematic phonemes for a given second language speaker are found by comparing this measure to the Euclidean geometry similarity for the same phonemes produced by native speakers only. The method is applied to different groups of non-native speakers of various language backgrounds and the experimental results are in agreement with theoretical findings of linguistic studies.

Index Terms: second language learning, auditory model, distortion measure, perceptual assessment, phoneme.

1. Introduction

When learning to speak a second language (L2), it may be a challenge to master the pronunciation of unfamiliar phonemes. One important reason for this is that the learner's auditory perception of the L2 phonemes is not sufficiently accurate. According to the dispersion principle and the auditory enhancement hypothesis [1], this is manifested in the learner's production through either large variations between attempts to produce the same phoneme (i.e., low precision) or consistent replacement of the target phoneme by another (i.e., low accuracy), often the most similar one in the native language (L1). Repeated practice is required to overcome both of these problems and computer-assisted pronunciation training may have an important role in this training, provided firstly that the computer program can analyze which phonemes the learner needs to practice and secondly that pronunciation errors can be detected and signaled to the learner. In this paper, we are focusing on the first issue, with the objective to make an automatic diagnostic evaluation of the phonemes that require additional practicing. That is, we are not addressing real-time mispronunciation detection of single utterances while the learner is using such a program. Instead, we are working with previously collected recordings in which the learner has produced the difficult phonemes several times in different settings. This means that the diagnostic evaluation can focus both on low precision [2] and consistent low accuracy. We propose a *language independent*, auditory model-based method to automatically identify phonemes that are repeatedly mispronounced by *individual* non-native speakers.

Commonly, pronunciation error detection has been formulated as a classification problem. In [3], the goodness of pronunciation (GOP) algorithm was presented to calculate the likelihood ratio of a phoneme realization to its canonical pronunciation. In [4], four different classifiers were examined to account for mispronunciation detection: one GOP-based, one combining cepstral coefficients with linear discriminant analysis, and two acoustic-phonetic classifiers. In [5], the problem was addressed with a support vector machine framework, with pronunciation space models to improve performance. None of these methods are based on auditory perception, e.g., psychoacoustic models of the periphery [6]. This may lead to unwanted decisions on the L2 pronunciation, since the statistical methods may not correspond to the judgment of a native human listener.

In [7, 8], a novel method to select perceptually relevant acoustic features, called auditory model-based feature selection, was presented for robust speech recognition. In this paper, we convert the method to be used in analyzing the pronunciation of L2 learners. The fundamental principle of our method is built upon measuring the similarity of the Euclidean geometry of the auditory representation for a group of native speakers and the speech signal's power spectrum for individual non-native speakers for each phoneme. This is motivated by the property of the human auditory periphery to provide a relatively good separation of sound classes, and the assumption that little information relevant for phoneme separation is lost in the mapping from the acoustic domain to the perceptual domain. Next, we calculate the geometric similarity of the native-speakers' power spectrum domain and perceptual domain for each phoneme. By comparing the measures for the non-native speaker and the native speakers, we find, quantitatively, the phonemes that are mispronounced by the L2 speaker.

2. Exploiting psychoacoustic knowledge

Models of the auditory periphery are used in various cases to study the perceptual processing of a sound or simply, to explain how it functions. The hearing system plays a vital role in human perception. The electrical signals produced by the hair cells in the inner ear travel through the auditory nerve to the brain. A sound is then considered to be perceived by the time these electrical signals reach the auditory cortex of the brain, where a cognitive processing is performed.

We consider a psychoacoustic model [6] that uses a series of auditory filters for computing the distortion. This agrees with recent findings about the spectral integration property of the human auditory system. The outer and middle ear are represented by a band-pass filter followed by a gammatone filterbank that models the basilar membrane of the inner ear. The total distortion measure is calculated as a summation of the distortion detectability provided by each auditory filter f multiplied by the effective duration L_e of the input signals.

2.1. A weighted square Euclidean dissimilarity measure

For our method, we need to measure the Euclidean dissimilarity between the speech frequency and the perceptual domains assuming a bijective and distance preserving mapping between these two domains.

Allowing small distortions, we quantify the dissimilarity of the Euclidean geometry of the magnitude spectrum of speech and the auditory periphery output. The approach is based on an analogous measure, introduced in [7], dealing with the selection of optimal acoustic feature subsets.

Starting by considering the power spectrum of the speech signal, we define a distortion measure in this domain as $\Phi : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+$, where $\mathbb{R}^+$ are the non-negative real numbers. Let $\mathbf{x}_i \in \mathbb{R}^N$ be the N-dimensional periodogram of the speech signal frame $i \in \mathbb{Z}$ and $\hat{\mathbf{x}}_{i,j}$ be the j'th perturbation of $\mathbf{x}_i$. A Euclidean norm-based measure is then

$$\Phi(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j}) = \parallel \mathbf{x}_i - \hat{\mathbf{x}}_{i,j} \parallel^2 . \tag{1}$$

Analogously, we define a perceptual-domain distortion measure as $\Upsilon : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+$. Let $\mathbf{y}(\mathbf{x}_i)$ and $\mathbf{y}(\hat{\mathbf{x}}_{i,j})$ be the auditory model output signals, where $\mathbf{y} : \mathbb{R}^N \rightarrow \mathbb{R}^M$ is a mapping of $\mathbf{x}_i$ and $\hat{\mathbf{x}}_{i,j}$, respectively, to the M-dimensional perceptual domain. Then

$$\Upsilon(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j}) = \parallel \mathbf{y}(\mathbf{x}_i) - \mathbf{y}(\hat{\mathbf{x}}_{i,j}) \parallel^2 . \tag{2}$$

We wish to find the geometric dissimilarity between the perceptual-domain distances and the speech frequency-domain distances using the following weighted square Euclidean measure of dissimilarity

$$\mathcal{A} = \frac{1}{\mathcal{I}} \sum_{i \in \mathcal{I}} \frac{1}{\mathcal{J}_i} \sum_{j \in \mathcal{J}_i} [\Upsilon(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j}) - \Phi(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j})]^2 . \tag{3}$$

Note that the scaling factor λ used in [7] to weigh out scaling mismatches between the two domains is not used here, since we now deal with speech data from different speakers. Indexes $i \in \mathcal{I}$ and $j \in \mathcal{J}_i$ represent a finite frame sequence and a finite set of acoustic perturbations, respectively.

2.2. Approximating perceptual distortion measure

Our approach to computing Eq. (2), i.e., the distortion in the auditory-model domain, utilizes the perturbation analysis and the sensitivity matrix [9]. Assume $\Upsilon(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j})$ to be known and suppose $\Upsilon(\mathbf{x}_i, \mathbf{x}_i) = 0$ to form a minimum. Additionally, assume that $\Upsilon(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j})$ is differentiable in $\hat{\mathbf{x}}_{i,j}$. Then, for sufficiently small perturbations $\hat{\mathbf{x}}_{i,j} - \mathbf{x}_i$, the following approximation can be made

$$\Upsilon(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j}) \approx [\hat{\mathbf{x}}_{i,j} - \mathbf{x}_i]^T \mathbf{D}_\Upsilon(\mathbf{x}_i) [\hat{\mathbf{x}}_{i,j} - \mathbf{x}_i], \tag{4}$$

where $\mathbf{D}_{\Upsilon,\mu\nu}(\mathbf{x}_i) = \left. \frac{\partial^2 \Upsilon(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j})}{\partial \hat{x}_\mu \partial \hat{x}_\nu} \right|_{\hat{\mathbf{x}}_{i,j} = \mathbf{x}_i}$ is the sensitivity matrix. This matrix, calculated using the auditory model introduced above, is a diagonal matrix with the diagonal element for row and column f given by

$$\mathbf{D}_{\Upsilon,ff}(\mathbf{x}) \approx 2 C_s L_e \sum_q \frac{\frac{1}{N} \sum_f \mathcal{F}(f)}{\frac{1}{N} \sum_f \mathcal{F}(f) |x(f)|^2 + C_a}, \tag{5}$$

where $\mathcal{F}(f) = |h_{om}(f)|^2 |\gamma_q(f)|^2$. h_{om} is the outer and middle ear transfer function and γ_q is the q'th gammatone filter. Constants C_s and C_a are calibrated based on measurement data, the integer g labels the gammatone filter and, finally, $\mathcal{G}$ is the set of gammatone filters considered. Eq. (5) shows how an error (small change) in the speech signal is reflected in the perceptual domain.

3. Native perceptual assessment scheme

Despite the variation in the acoustic signal between L1 speakers for the same phoneme, native speakers can still easily distinguish the acoustic properties of their L1 phonemes. They may, on the other hand, have difficulties handling corresponding speech uttered by foreign speakers. We perform the sensitivity analysis described in Sec. 2.1 to assess the degree of difference in the production of a phoneme by a non-native speaker compared to native speakers.

Let $\mathbf{p}$ be a phoneme class. Usually, the mapping from $\mathbf{p}$ to the speech or the perceptual domain is given, instead of the distortion criterion per se. Hence, consider the mapping $\mathbf{x}$ to the speech frequency domain. Assuming the mapping $\mathbf{x}$ to be analytic, the Taylor series is used to make a local approximation around $\mathbf{p}$:

$$\mathbf{x}(\hat{\mathbf{p}}) \approx \mathbf{x}(\mathbf{p}) + \mathbf{J}_p [\hat{\mathbf{p}} - \mathbf{p}], \tag{6}$$

where[1] $\mathbf{J}_p = \left. \frac{\partial \mathbf{x}(\mathbf{p})}{\partial \hat{\mathbf{p}}} \right|_{\hat{\mathbf{p}} = \mathbf{p}}$. We now consider a way to relate native to non-native speech.

3.1. Expressing non-native speech into native space

Considering native speech only, Eq. (6) can be written as

$$\mathbf{x}^{nat}(\hat{\mathbf{p}}) \approx \mathbf{x}^{nat}(\mathbf{p}) + \mathbf{J}_p^{nat} [\hat{\mathbf{p}} - \mathbf{p}]. \tag{7}$$

Accordingly, the non-native speech frequency representation $\mathbf{x}^L$ of a language group of speakers L can be expressed as

$$\mathbf{x}^L(\hat{\mathbf{p}}) \approx \mathbf{x}^L(\mathbf{p}) + \mathbf{J}_p^L [\hat{\mathbf{p}} - \mathbf{p}], \tag{8}$$

or, by using Eq. (7), it can be related to the native speech signal as

$$\mathbf{x}^L(\hat{\mathbf{p}}) \approx \mathbf{x}^L(\mathbf{p}) + \mathbf{W}_p^L \, [\mathbf{x}^{nat}(\hat{\mathbf{p}}) - \mathbf{x}^{nat}(\mathbf{p})], \tag{9}$$

where $\mathbf{W}_p^L = \mathbf{J}_p^L \, [\mathbf{J}_p^{nat}]^{-1}$. Then, the power spectrum distortion measure $\Phi_p^L(\mathbf{x}_i^{nat}, \hat{\mathbf{x}}_{i,j}^{nat})$, Eq. (1), for the non-native speech signal is approximated as

$$\Phi_p^L(\cdot\,,\,\cdot) \approx [\mathbf{x}_i^{nat} - \hat{\mathbf{x}}_{i,j}^{nat}]^T \, [\mathbf{W}_p^L]^T \, \mathbf{W}_p^L \, [\mathbf{x}_i^{nat} - \hat{\mathbf{x}}_{i,j}^{nat}], \tag{10}$$

where $i \in \mathcal{I}, j \in \mathcal{J}_i$.

3.2. Finding matrix $\mathbf{W}_p^L$

Phenomena such as duration or silence mismatch between the native and non-native speech signal preclude the computation of the $\mathbf{W}_p^L$ on a frame basis. Furthermore, mathematical reasons, e.g., non-invertible matrices, make this effort impossible. Therefore, as a compromise, the $\mathbf{W}_p^L$ matrix is calculated by considering a common matrix for all frames i of a certain phoneme class p for a specific foreign language group of speakers L.

We assume both native and non-native speech signals to follow a Gaussian distribution. Hence, Eq. (9) can be expressed as $\mathcal{N}(\mu_p^L, \ \mathbf{\Sigma}_p^L) \sim \mathcal{N}(\mathbf{W}_p^L \, \mu_p^{nat}, \ \mathbf{W}_p^L \, \mathbf{\Sigma}_p^{nat} \, [\mathbf{W}_p^L]^T)$, where μ_p^L, μ_p^{nat} are the mean vectors of the distortion in non-native

[1] $\hat{\mathbf{p}}$ is a theoretical notion. It can be considered to be the phoneme that is expressed by the mapping $\hat{\mathbf{x}}$, i.e., the "distorted" phoneme class as represented by a distorted speech signal.

and native speech signals for a phoneme class p, respectively and $\mathbf{\Sigma}_p^L$, $\mathbf{\Sigma}_p^{nat}$ their covariance matrices.

The Schur decomposition of the above two covariance matrices gives

$$\mathbf{\Sigma}_p^{nat} = \mathbf{V}_p^{nat}\, \mathbf{S}_p^{nat}\, [\mathbf{V}_p^{nat}]^T, \tag{11}$$

for the native language group and

$$\mathbf{\Sigma}_p^{L} = \mathbf{V}_p^{L}\, \mathbf{S}_p^{L}\, [\mathbf{V}_p^{L}]^T, \tag{12}$$

for the non-native language group L, respectively. Assuming the following distributions

$$\begin{aligned}
Z &\sim \mathcal{N}([\mathbf{V}_p^{nat}]^T\, \mu_p^{nat},\ [\mathbf{V}_p^{nat}]^T\, \mathbf{\Sigma}_p^{nat}\, \mathbf{V}_p^{nat}) \\
Q &\sim \mathcal{N}([\mathbf{S}_p^{nat}]^{-\frac{1}{2}}\, \mu_Z,\ [\mathbf{S}_p^{nat}]^{-\frac{1}{2}}\, \mathbf{\Sigma}_Z\, [\mathbf{S}_p^{nat}]^{-\frac{T}{2}}), \\
K &\sim \mathcal{N}([\mathbf{S}_p^{L}]^{\frac{1}{2}}\, \mu_Q,\ [\mathbf{S}_p^{L}]^{\frac{1}{2}}\, \mathbf{\Sigma}_Q\, [\mathbf{S}_p^{L}]^{\frac{T}{2}}), \\
\Psi &\sim \mathcal{N}(\mathbf{V}_p^{L}\, \mu_K,\ \mathbf{V}_p^{L}\, \mathbf{\Sigma}_K\, [\mathbf{V}_p^{L}]^T),
\end{aligned} \tag{13}$$

and performing a Schur decomposition in each of them, it can be proved that matrix $\mathbf{W}_p^L$ is given by

$$\mathbf{W}_p^L = \mathbf{V}_p^L\, [\mathbf{S}_p^L]^{\frac{1}{2}}\, [\mathbf{S}_p^{nat}]^{-\frac{1}{2}}\, [\mathbf{V}_p^{nat}]^T. \tag{14}$$

3.3. The algorithm

A significant component of the automatic evaluation of the pronunciation of non-native speakers is the sensitivity matrix $\mathbf{D}_{\Upsilon_p^{nat}}(\mathbf{x}_i)$ given by Eq. (5), which was calculated for each speech segment i, and phoneme p for the native speakers. A set of 100 vectors $\hat{\mathbf{x}}_{i,j}$ was computed by adding 30 dB SNR i.i.d. Gaussian noise to $\mathbf{x}_i$. Eq. (14) was used to compute the matrix $\mathbf{W}_p^L$ over all frames of each phoneme and language group. Next, the dissimilarity measure $\mathcal{A}_p^L$ was calculated using the native perceptual distortion measure $\Upsilon_p^{nat}(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j})$ given by Eq. (4), and the non-native speech frequency distortion measure $\Phi_p^L(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j})$ given by Eq. (10). Then, the corresponding dissimilarity measure for the native speakers $\mathcal{A}_p^{nat}$ was calculated using again $\Upsilon_p^{nat}(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j})$ and the native speech frequency distortion measure $\Phi_p^{nat}(\mathbf{x}_i, \hat{\mathbf{x}}_{i,j})$ of Eq. (1). Finally, the *native-perceptual assessment degree* Θ_p^L was computed for every phoneme and L1 background as

$$\Theta_p^L = \frac{\mathcal{A}_p^L}{\mathcal{A}_p^{nat}}. \tag{15}$$

Θ_p^L is a normalized ratio that shows the degree of the dissimilarity between the native perceptual outcome and the non-native power spectrum as compared to the native-only case.

4. Speech data

A speech corpus, sampled at 16 kHz, was recorded to be used in our experiments. It was designed for L2 learners of Swedish as a part of the computer-assisted language learning program Ville [10], consisting of an embodied conversational agent that acts as a virtual language tutor for Swedish. 37 (23 male and 14 female) speakers of different language backgrounds (cf. Table 1) took a test twice within one month's time, before and after practising at home. The test lasted 30 minutes and consisted of exercises in which the participants repeated single words and sentences of varying complexity after Ville. For the purpose of our method, 11 (9 males and 2 females) Swedish speakers were also recorded once each.

Table 1: *List of participants (par.) and L1 backgrounds (bkgr.)*

L1 bkgr.	(par.)files	L1 bkgr.	(par.)files	L1 bkgr.	(par.)files
Eng.(US)	(2) 318	*Russian*	(4) 583	*Arabic*	(1) 164
German	(2) 249	*Greek*	(3) 393	*Chinese*	(5) 832
French	(3) 347	*Spanish*	(5) 882	*Persian*	(6) 987
Polish	(2) 317	*Turkish*	(4) 604	*Swedish*	(11) 888

The data were cleaned from extra-linguistic content, e.g., coughs, long pauses and hesitation phenomena, such as repetitions and fillers ("um", "uh", "eh" etc). When necessary, e.g., in the case of deletions and insertions, the accompanying text file was adjusted to the actual content. A phone-level transcription was automatically generated from the speech signal and the text file using an HMM-based aligner [11]. These phone-level transcription files were used to separate the speech data into phoneme categories. The material contains all Swedish phonemes, but five of the most challenging, /æ/, /œ/, /ɳ/, /ɖ/, and /ɭ/ were excluded from our study due to the small number of occurrences in the database. The speech signal was pre-emphasized and the output was windowed by a Hamming window of 25 ms with an overlap of 10 ms. A discrete Fourier transform of 512 points was applied to the windowed frame to compute the signal's power spectrum.

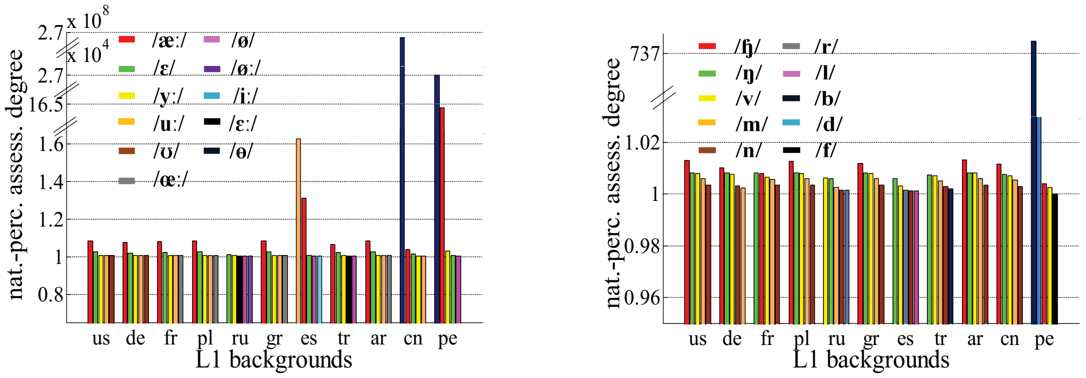
5. Results and Discussion

We present our experimental results and discuss our findings in comparison with a linguistic study of problematic Swedish phonemes for different L1 groups [12]. Since we only consider the acoustic signal of the uttered phonemes to evaluate the foreign accent, our work is not a comprehensive linguistic study. Grammatical or syntactical errors, as well as errors due to context, are excluded.

Table 2 lists the phonemes found to be difficult for the different groups of non-native speakers in our study. The results of our perceptual-based method are, in general, in agreement with previous linguistic observations [12]. For each L2 speaker group, the first line shows, in order, the most deviating vowels according to our method. Correspondingly, the second line shows the problematic consonants. Divergences from the theoretical findings are reported in parentheses, most of which are for lower Θ_p^L. Fig. 1(a-b) illustrates the degree Θ_p^L per L2 speaker group for the five most mispronounced (a) vowels, and

Table 2: *Problematic phonemes per language background. The phonemes are shown in decreasing order, starting from the one with the highest Θ_p^L (see Eq. (15)). Phonemes that differ from the linguistic study findings are listed in parentheses.*

L1 background		phonemes
English	vowels	æː, ɛ, yː, uː, ʊ, œː, ɛː, ø, ɵ, øː, (iː), ɑː, (ə), eː, e, ɔ, a, ʉː
(US)	consonants	ɧ, ŋ, (v), m, n, (b), r, (d), l, k, ʂ, t
German	vowels	æː, (ɛ), yː, uː, (ʊ), ɛː, (ø), œː, øː, iː, ə, ɑː
	consonants	ɧ, ŋ, v, n, (m), b, r, d, (l), k, ʂ, t, p, (h), f, ɕ, s
French	vowels	æː, ɛ, yː, uː, œː, (ʊ), ɛː, (ø), ɵ, øː, iː, ə, ɑː, eː, e, ɔ, (a)
	consonants	ŋ, ɧ, (v), m, n, b, r, (l), d, ʂ, k, t, p, h, ɕ, g
Polish	vowels	æː, (ɛ), yː, uː, œː, (ʊ), ɛː, ø, ɵ, øː, iː, ɑː, ə, eː, (e), (ɔ)
	consonants	ɧ, ŋ, v, (m), n, b, (r), d, (l), k, ʂ, t, p, s, (f)
Russian	vowels	ɛ, yː, ɛː, ø, øː, iː, ɑː, ɵ, eː, e, ʉː, a, (ɔ), ʏ, (ə), (i), œː
	consonants	v, ŋ, (m), (n), (r), d, (l), h, b, k, t, g, (s), f, j
Greek	vowels	æː, (ɛ), yː, uː, œː, (ʊ), ɛː, ø, ɵ, øː, iː, ɑː, ə, eː, e, (ɔ)
	consonants	ɧ, ŋ, (v), m, n, b, (r), d, l, ʂ, k, t, p, ɕ, (f), s
Spanish	vowels	uː, æː, ɛ, ø, iː, ɵ, eː, e, øː, (a), ʉː, ɑː, (i), ə, ʏ, (ɔ), oː
	consonants	ŋ, v, (r), n, (l), b, t, ʂ, (f), k, g, d, j, p, ɕ, s, h
Turkish	vowels	æː, (ɛ), yː, (ɛː), (ø), uː, ɵ, ʊ, øː, iː, ɑː, eː, (ə)
	consonants	ŋ, v, (m), n, b, r, l, d, k, (ʂ), t, p, f, h, g, ʈ, j, s
Arabic	vowels	æː, ɛ, yː, uː, œː, (ʊ), ɛː, ø, ɵ, øː, iː, ɑː, ə, eː, e, (ɔ), a, ʉː
	consonants	ɧ, ŋ, v, (m), (n), (b), r, d, (l), k, ʂ, t, p
Chinese	vowels	ɵ, æː, ɛ, yː, uː, ɛː, ø, øː, iː, ɑː, eː, e, ɔ, (ə), a, ʉː, (i), oː
	consonants	ɧ, ŋ, v, m, n, b, r, l, d, k, t, f, g, ʈ, p, j, (h), (s)
Persian	vowels	ɵ, æː, yː, (ɛ), ø, øː, ʏ, (i), ʉː, eː, a, (e), (ɑː), oː, (ɔ)
	consonants	b, d, (ɧ), v, (f), g, (h), t, s, (j), ɕ, p, ʈ, k, l, r



(a) vowels

(b) consonants

Figure 1: *The value of* Θ_p^L *for the five most problematic vowels/consonants for each L2 learning group as listed in Table 2. (us: English (US), de: German, fr: French, pl: Polish, ru: Russian, gr: Greek, es: Spanish, tr: Turkish, ar: Arabic, cn: Chinese, pe: Persian.)*

(b) consonants. The two figures reveal difficulties in Swedish phonemes for each L2 group, including (but not limited to): 1) Most groups, but foremost, Persian and Spanish speakers, have problems with the more open r-allophone /æː/. 2) Chinese and Persian speakers face difficulties to producing the rounded /ɵ/. 3) Spanish speakers mispronounce the long /uː/. 4) Persian speakers often make voicing errors in /b/. 5) Almost all groups have problems with the "*sje-sound*", /ɧ/, the more or less uniquely Swedish rounded velar fricative. 6) In addition, most speakers are inclined to mispronounce the velar nasal /ŋ/. For many phonemes, Θ_p^L has a relatively low value as shown in Fig. 1. This can be explained by the nature and the context of the test (repeating after a native speaker and guided by a text prompt).

Many of these differences in the results shown in Table 2 may be due to methodology. The aim in [12] was to make an inventory of pronunciation errors for different groups of L2 speakers and their importance with respect to native listeners. It was performed through linguistic analysis and subjective observations. We, on the contrary, focus on an objective acoustic evaluation of the phone. This means that phenomena such as epenthesis, elision, coarticulation, or phoneme position are not considered in our work. Moreover, whereas the goal in [12] was to identify common difficulties for a group of speakers with the same L1, our method is aimed at identification of problematic phonemes for single speakers. It is hence a desirable built-in attribute of our auditory-based method that it identifies the problematic phonemes for the tested speaker(s), rather than general problems associated with the L1. The method therefore captures properties of the individual speaker(s), such as fluency in the L2 or social background, which may have more influence on the speaker's accent than the L1 background.

6. Conclusions

We presented a machine-driven method to quantitatively assess the non-native speakers' phonemes pronunciation based on an auditory model. The method estimates native perception of non-native speech and compares the outcome to native perception of native speech. It hence indicates phonemes for which native speakers would perceive a mispronunciation by a non-native speaker. The results are verified by theoretical linguistic studies. It is our intention to integrate the new method in a CAPT framework. An additional path for future work is to investigate if and how the method can be expanded to mispronunciation detection of single utterances, that is to identify in real time, when the learner practices with the system, when an error occurs.

7. Acknowledgments

This work is supported by the Swedish Research Council project 80449001 Computer-Animated LAnguage TEAchers (CALATEA). The authors wish to thank Saikat Chatterjee from the Communication Theory Lab, KTH for constructive suggestions.

8. References

[1] R. L. Diehl and K. R. Kluender, "On the Objects of Speech Perception", *Ecological Psychology*, 1(2):121-144, Jan. 1989.

[2] C. Koniaris and O. Engwall, "Perceptual Differentiation Modeling Explains Phoneme Mispronunciation by Non-Native Speakers", *in IEEE Int. Conf. Acoust., Speech, Sig. Proc., Prague, Czech Republic*, 5704-5707, May 2011.

[3] S. M. Witt and S. Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning", *Speech Communication*, 30(2-3):95-108, Feb. 2000.

[4] H. Strik, K. Truong, F. de Wet and C. Cucchiarini, "Comparing Different Approaches for Automatic Pronunciation Error Detection", *Speech Communication*, 51(10):845-852, Oct. 2009.

[5] S. Wei, G. Hu, Y. Hu and R.-H. Wang, "A New Method for Mispronunciation Detection Using Support Vector Machine based on Pronunciation Space Models", *Speech Communication*, 51(10):896-905, Oct. 2009.

[6] S. van de Par, A. Kohlrausch, G. Charestan and R. Heusdens, "A New Psychoacoustical Masking Model for Audio Coding Applications", *in IEEE Int. Conf. Acoust., Speech, Sig. Proc., Orlando, FL, USA*, 2:1805-1808, May 2002.

[7] C. Koniaris, M. Kuropatwinksi and W. B. Kleijn, "Auditory-Model Based Robust Feature Selection for Speech Recognition", *J. Acoust. Soc. Amer.*, 127(2):EL73-EL79, Feb. 2010.

[8] C. Koniaris, S. Chatterjee and W. B. Kleijn, "Selecting Static and Dynamic Features Using an Advanced Auditory Model for Speech Recognition", *in IEEE Int. Conf. Acoust., Speech, Sig. Proc., Dallas, TX, USA*, 4342-4345, Mar. 2010.

[9] W. R. Gardner and B. D. Rao, "Theoretical Analysis of the High-Rate Vector Quantization of LPC Parameters", *IEEE Tr. Speech, Audio Proc.*, 3(5):367-381, Sep. 1995.

[10] P. Wik and A. Hjalmarsson, "Embodied Conversational Agents in Computer Assisted Language Learning", *Speech Communication*, 51(10):1024-1037, Oct. 2009.

[11] K. Sjölander, "An HMM-based System for Automatic Segmentation and Alignment of Speech", *in Fonetik*, 93-96, Jun. 2003.

[12] R. Bannert, "Problems in Learning Swedish Pronunciation and in Understanding Foreign Accent", *Folia Linguistica*, 18(1-2):193-222, Jan. 1984.