An approach to measure pronunciation similarity in second language learning using radial basis function kernel

Christos Koniaris

University of Gothenburg, Centre for Language Technology
Department of Philosophy, Linguistics and Theory of Science
Dialogue Technology Lab, Gothenburg, Sweden

christos.koniaris@gu.se

ABSTRACT
This paper shows a method to diagnose potential mispronunciations in second language learning by studying the characteristics of the speech produced by a group of native speakers and the speech produced by various non-native groups of speakers from diverse language backgrounds. The method compares the native auditory perception and the non-native spectral representation on the phoneme level using similarity measures that are based on the radial basis function kernel. A list of ordered problematic phonemes is found for each non-native group of speakers and the results are analyzed based on a relevant linguistic survey found in the literature. The experimental results indicate an agreement with linguistic findings of up to 80.8% for vowels and 80.3% for consonants.

KEYWORDS: pronunciation error detection, similarity measure, radial basis function kernel, phoneme, second language learning.

Christos Koniaris 2014. An approach to measure pronunciation similarity in second language learning using radial basis function kernel. *Proceedings of the third workshop on NLP for computer-assisted language learning.* NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107: 74–86.

1 Introduction

Second language (L2) speakers are generally having trouble with certain phonemes of the target language that do not exist in the sound system of their native language (Flege, 1995; Guion et al., 2000). It is therefore common practice to include speech sounds from their first language (L1) or ignore unfamiliar ones (Piske et al., 2001) while practicing a new language. Within a computer-assisted language learning (CALL) program, the task of automatic *pronunciation error detection* (PED) is to find effective techniques to diagnose and detect mispronunciations in order to assist L2 learners to improve their oral capabilities.

In (Neumeyer et al., 1996; Franco et al., 1997; Neumeyer et al., 2000) a system used for performing automatic speech recognition (ASR) is turned into an automatic pronunciation scoring system, in which several different scores, e.g., hidden Markov models (HMM) phone log-likelihood, are compared to human listeners' evaluation. The experiments show that certain scores, such as the log-posterior and the normalized duration correlate well with human ratings. Scoring is also the main characteristic of the goodness of pronunciation (GOP) proposed in (Witt and Young, 2000), which measures the quality of pronunciations of non-native speakers. The idea is to score each phone of an utterance depending on how close the pronunciation of the non-native speaker is to that of native speakers. A method that combines knowledge from acoustic-phonetic, linguistic, and from expert listeners is presented in (Park and Rhee, 2004), in which the analysis of the results is done by finding the correlation of human listeners and machine-based rating. In (Truong et al., 2005), a set of classification approaches based on linear discriminant analysis (LDA) and decision trees is presented. These classifiers are used to analyze the mispronunciations of second language learners of Dutch. In (Tepperman and Narayanan, 2008), the research is oriented in introducing articulatory information in PED by reformulating the hidden-articulator Markov models (HAMM) (Tepperman and Narayanan, 2005) and deriving new articulatory-based features for classification. In (Strik et al., 2009), four different classification systems are examined: a GOP-based, one combining cepstral coefficients and LDA, a method based on the work described in (Weigelt et al., 1990), which is an algorithm that discriminates voiceless fricatives from voiceless plosives, and an LDA-acoustic-phonetic feature classifier. It is found that the two LDA-based classification systems perform better in mispronunciation detection. In (Wei et al., 2009), the authors use support vector machines (SVM) to model phones with several parallel acoustic models that represent the variation in pronunciation at various proficiency levels. This approach seems to achieve better results in comparison to more traditional posterior probability based methods.

Since the pronunciation of a phone is not only related to its acoustics, aspects, such as fluency, syllable structure, word stress, intonation, prosody or segmental quality may also be considered for investigation of pronunciation errors. For example, the work that is presented in (Delmonte, 2000) concerns a prosodic module of a CALL system called SLIM. This module deals with phonetic and prosodic problems both at the word but also at the segmental level. Prosodic measures based on F0, power and duration of L2 and L1 speech are used in (Yamashita et al., 2005) within a multiple regression framework to predict the prosodic proficiency of L2 learners. In (Raux and Kawahara, 2002), a probabilistic algorithm is applied to derive intelligibility from error rates and also define a function of error priority to indicate which errors are most critical to intelligibility. Finally, in (Xu et al., 2009), linguistic knowledge obtained from the non-native speakers' most common mistakes, and pronunciation space constructed using revised log-posterior probability vectors is considered along with an SVM classifier.

In this paper, a PED method based on psychoacoustic knowledge from a spectral auditory model (van de Par et al., 2002) is presented that models the native perception to evaluate non-native pronunciations based on acoustic and auditory processing of the speech sounds. The fundamental assumption is based on the ability of the human auditory system to distinguish speech sounds of various type. The method compares the acoustic and auditory-perceptual characteristics of uttered phones on a frame-by-frame basis. In doing so, it utilizes a similarity measure based on radial basis function kernel or RBF kernel, which is compared with a Euclidean distance measure that was used in (Koniaris and Engwall, 2011; Koniaris et al., 2013). The motivation for this arrives from the fact that the data become sparse in a high dimensional space and hence choosing RBF kernel seems a more suitable solution since it is considered more appropriate for such conditions (Braun et al., 2008). Roughly speaking, the method performs a comparison between speech sounds generated by a group of native speakers with the corresponding speech sounds generated by different L2 groups of speakers. This is done separately for each phoneme category and the uttered phones are transformed into their auditory representations for the native speech, and into their spectrum representations for the non-native speech. In each domain, a distortion measure based on the RBF kernel is computed for each speech frame and then the two distortion measures are explored – considering all the frames – to investigate, quantitatively, the similarities between the native and the non-native phones.

The paper is organized as follows. Section 2 presents the method and implementation issues, Section 3 discusses the experiments and the findings and finally Section 4 provides conclusions.

2 Method

The underlying idea behind the pronunciation error detection method that is described here is based on the auditory ability of a native speaker to discriminate the mispronounced phonemes produced by L2 speakers while hearing them speaking. The diagnostic evaluation of the pronunciation errors is done on the speech signal level by comparing the similarities between the auditory perceptual domain of the native speech and the power spectrum domain of the non-native speech. It is assumed that a non-native acoustic representation will have very similar characteristics to native provided that the non-native speech is produced without significant mispronunciation. On the other hand, if the non-native speech suffers from severe pronunciation errors then the two representations, of the L2 and L1 speakers, will differ a lot and thus the measured similarities will become minimal (Koniaris et al., 2013).

In short, the approach tries to measure the distortion in a set of phones that belong to a specific phoneme, produced by a group of native speakers n and compare it to that of non-native speakers of some specific language background ℓ. For this, it is assumed that some form of acoustic representation $\mathbf{x}$ is extracted from the speech signal $\mathbf{s}$ of a phone $\mathbf{p}$ to evaluate the distortion measure ϕ in the corresponding transformed domain, where $\phi : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+$, with $\mathbb{R}^+$ denoting the non-negative real numbers and N indicating the dimensionality of the vector $\mathbf{x}$. Then, the RBF kernel-based similarity measure is,

$$\phi(\mathbf{s}_i, \hat{\mathbf{s}}_{i,j}) = e^{\gamma ||\mathbf{x}(\mathbf{s}_i) - \mathbf{x}(\hat{\mathbf{s}}_{i,j})||^2}, \tag{1}$$

where $i \in \mathbb{Z}$ is the index of the considered speech frame, $\hat{\mathbf{s}}_{i,j}$ is the j'th perturbation of $\mathbf{s}_i$ that is used to compute distortion and $\gamma = -\frac{1}{2\sigma^2}$. It is noted that σ will determine the size of the considered area around $\mathbf{s}_i$. An analogous measure is defined for the auditory perception domain where the speech signal is transformed into the auditory model output representation $\mathbf{y}$. Again, a RBF kernel-based distortion measure is computed in the auditory domain $\upsilon : \mathbb{R}^M \times \mathbb{R}^M \rightarrow \mathbb{R}^+$, where M is the dimensionality of the internal representation $\mathbf{y}$, as

$$\upsilon(\mathbf{s}_i, \hat{\mathbf{s}}_{i,j}) = e^{\gamma \|\mathbf{y}(\mathbf{s}_i) - \mathbf{y}(\hat{\mathbf{s}}_{i,j})\|^2}. \tag{2}$$

The above distortion measures of Eqs. (1) and (2) are then compared using the following similarity measure

$$\mathscr{A} = \frac{1}{\mathscr{I}} \sum_{i \in \mathscr{I}} \frac{1}{\mathscr{J}_i} \sum_{j \in \mathscr{J}_i} \left[\upsilon(\mathbf{s}_i, \hat{\mathbf{s}}_{i,j}) - \phi(\mathbf{s}_i, \hat{\mathbf{s}}_{i,j}) \right]^2, \tag{3}$$

where $i \in \mathscr{I}$ and $j \in \mathscr{J}_i$ represent a finite frame sequence and a finite set of acoustic perturbations, respectively. This measure is used to find mispronunciations as described in (Koniaris et al., 2013), i.e., by computing the distortion measure $\upsilon(\mathbf{s}_i, \hat{\mathbf{s}}_{i,j})$ using only native speech and the spectral distortion measure $\phi(\mathbf{s}_i, \hat{\mathbf{s}}_{i,j})$, calculated separately for non-native (and thus computing $\mathscr{A}_\ell$) and native speech (and thus computing $\mathscr{A}_n$). Finally, the *native-perceptual assessment degree (nPAD)* is computed for every phoneme and L1 background as

$$nPAD = \frac{\mathscr{A}_\ell}{\mathscr{A}_n}, \tag{4}$$

which is a normalized ratio that shows the degree of the similarity between the native perceptual outcome and the non-native speech signal representation, as compared to the native-only case. The higher the nPAD value is, the more problematic the L2 phoneme is.

2.1 Practical implementation

Considering $\mathbf{p}$ to be a phone that a speaker has produced, the speech power spectrum $\mathbf{x}(\mathbf{p})$ can be seen simply as a function of $\mathbf{p}$ that maps this phone onto the spectral domain. If additionally is considered a small area around $\mathbf{p}$, a local approximation is possible using the Taylor series expansion, thus

$$\mathbf{x}(\hat{\mathbf{p}}) \approx \mathbf{x}(\mathbf{p}) + \mathbf{J}_\mathbf{x}[\hat{\mathbf{p}} - \mathbf{p}], \tag{5}$$

where $\mathbf{J}_\mathbf{x} = \left. \frac{\partial \mathbf{x}(\mathbf{p})}{\partial \hat{\mathbf{p}}} \right|_{\hat{\mathbf{p}} = \mathbf{p}}$ and $\hat{\mathbf{p}}$ is the perturbed phone. Assuming that the small distortion $[\hat{\mathbf{p}} - \mathbf{p}]$ remains the same independently of the language background of the speaker, Eq. (5) can be used either for native speech $\mathbf{x}_n$ or non-native speech $\mathbf{x}_\ell$ of a language background ℓ. This means that is possible to find a linearized relation between these two and compute the speech power spectrum distortion in a non-native subspace into the native speech power spectrum domain. Thus,

$$\mathbf{x}_\ell(\hat{\mathbf{p}}) \approx \mathbf{x}_\ell(\mathbf{p}) + \mathbf{W}_\ell \left[\mathbf{x}_n(\hat{\mathbf{p}}) - \mathbf{x}_n(\mathbf{p}) \right], \tag{6}$$

where $\mathbf{W}_\ell = \mathbf{J}_{\mathbf{x}\ell} \left[\mathbf{J}_{\mathbf{x}n} \right]^{-1}$. Eq. (6) implies that a different $\mathbf{W}_\ell$ should be calculated for each frame. However, the duration of phones or silence mismatches between the native and non-native speech signal prevent such computation. In addition, the matrices are non-invertible. Therefore the estimation of $\mathbf{W}_\ell$ is done by considering a common matrix for all frames i of a specific L2 group of speakers ℓ. In speech processing is often assumed that a speech signal follows a Gaussian distribution. Thus, Eq. (6) can be expressed as $\mathcal{N}(\boldsymbol{\mu}_\ell,\ \boldsymbol{\Sigma}_\ell) \sim \mathcal{N}(\mathbf{W}_\ell\, \boldsymbol{\mu}_n,\ \mathbf{W}_\ell\, \boldsymbol{\Sigma}_n\, [\mathbf{W}_\ell]^T)$, where $\boldsymbol{\mu}_\ell$, $\boldsymbol{\mu}_n$ are the mean vectors of the distortion in non-native and native speech signals, respectively and $\boldsymbol{\Sigma}_\ell$, $\boldsymbol{\Sigma}_n$ their covariance matrices.

Considering a matrix decomposition (e.g., eigendecomposition), the two covariance matrices can be expressed as

$$\mathbf{\Sigma}_\zeta = \mathbf{V}_\zeta \, \mathbf{S}_\zeta \, [\mathbf{V}_\zeta]^T, \tag{7}$$

where $\zeta = n$ for the native language group, and $\zeta = \ell$ for the non-native language group. Next, assuming the following distributions

$$\begin{aligned} Z &\sim \mathscr{N}([\mathbf{V}_n]^T \boldsymbol{\mu}_n, \; [\mathbf{V}_n]^T \mathbf{\Sigma}_n \mathbf{V}_n) \\ Q &\sim \mathscr{N}([\mathbf{S}_n]^{-\frac{1}{2}} \boldsymbol{\mu}_Z, \; [\mathbf{S}_n]^{-\frac{1}{2}} \mathbf{\Sigma}_Z \, [\mathbf{S}_n]^{-\frac{T}{2}}), \\ K &\sim \mathscr{N}([\mathbf{S}_L]^{\frac{1}{2}} \boldsymbol{\mu}_Q, \; [\mathbf{S}_L]^{\frac{1}{2}} \mathbf{\Sigma}_Q \, [\mathbf{S}_L]^{\frac{T}{2}}), \\ \Psi &\sim \mathscr{N}(\mathbf{V}_L \boldsymbol{\mu}_K, \; \mathbf{V}_L \mathbf{\Sigma}_K \, [\mathbf{V}_L]^T), \end{aligned} \tag{8}$$

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

$$\mathbf{W}_\ell = \mathbf{V}_\ell \, [\mathbf{S}_\ell]^{\frac{1}{2}} \, [\mathbf{S}_n]^{-\frac{1}{2}} \, [\mathbf{V}_n]^T. \tag{9}$$

Then, the power spectrum distortion measure for the non-native speech signal is calculated as

$$\phi_\ell(\mathbf{x}_{\ell_i}, \hat{\mathbf{x}}_{\ell_{i,j}}) \cong \phi_\ell(\mathbf{x}_{n_i}, \hat{\mathbf{x}}_{n_{i,j}}; \mathbf{x}_{\ell_i}, \hat{\mathbf{x}}_{\ell_{i,j}}) \approx [\mathbf{x}_{n_i} - \hat{\mathbf{x}}_{n_{i,j}}]^T \, [\mathbf{W}_\ell]^T \, \mathbf{W}_\ell \, [\mathbf{x}_{n_i} - \hat{\mathbf{x}}_{n_{i,j}}], \tag{10}$$

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

As mentioned above, a small area is considered around each phone. In practice, this is done by allowing small perturbations, i.e., adding 30 dB SNR independent and identically distributed (i.i.d.) Gaussian noise to each $\mathbf{x}_i$ and generate a set of 100 vectors $\hat{\mathbf{x}}_{i,j}$ for the native speech data n as well as for non-native speech data of all language backgrounds ℓ. All data from native speech are used to calculate the perceptual distortion measure Eq. (2) on a frame by frame basis by exploiting auditory information from the psychoacoustic model presented in (van de Par et al., 2002). Analogously, all data from non-native speech of each language group ℓ are used to compute Eq. (1) and, separately, all data from native speech, too. Next, the similarity measure $\mathscr{A}_\ell$ is calculated using the native perceptual distortion and the non-native spectral distortion measures and also the corresponding similarity measure for the native speakers $\mathscr{A}_n$ using the native perceptual and spectral measures. Then for each phoneme class, the RBF kernel-based *nPAD* Θ_ℓ^{rbf} is computed for every L2 background using Eq. (4). Finally, a Euclidean-based *nPAD* Θ_ℓ, described in (Koniaris and Engwall, 2011; Koniaris et al., 2013), is calculated by considering Euclidean distances in Eqs. (1) and (2), i.e., $\phi(\mathbf{s}_i, \hat{\mathbf{s}}_{i,j}) = \| \, \mathbf{x}(\mathbf{s}_i) - \mathbf{x}(\hat{\mathbf{s}}_{i,j}) \, \|^2$ and $\upsilon(\mathbf{s}_i, \hat{\mathbf{s}}_{i,j}) = \| \, \mathbf{x}(\mathbf{s}_i) - \mathbf{x}(\hat{\mathbf{s}}_{i,j}) \, \|^2$, respectively.

3 Experiments

This section describes the experiments and discusses the findings of the RBF kernel-based approach in relation to the Euclidean-based approach and the theoretical linguistic survey presented in (Bannert, 1984).

3.1 Speech data

The speech data were recorded with a sampling frequency of 16 kHz consisting of 23 phonetically rich single words and 55 sentences of varying complexity and length. The utterances were specifically designed for L2 learners of Swedish that were using a CALL program (Wik and Hjalmarsson, 2009). The collection of the data was done through a desktop microphone while

L1 bkgr.	male/female	utt.	L1 bkgr.	male/female	utt.	L1 bkgr.	male/female	utt.
Eng.(US)	1/1	318	*Russian*	1/3	583	*Arabic*	0/1	164
German	2/0	249	*Greek*	3/0	393	*Chinese*	2/3	832
French	3/0	347	*Spanish*	4/1	882	*Persian*	3/3	987
Polish	0/2	317	*Turkish*	4/0	604	*Swedish*	9/2	888

Table 1: Distribution of the total number of male and female speakers and the number of utterances (utt.) for each language background (L1 bkgr.).

students repeated a word or sentence after the virtual language tutor, the main character of the program. The procedure was simple; first, the animated agent produced an utterance – a pre-recorded natural speech produced by a native speaker – accompanied by a subtitle text and the student repeated afterwards.

The total number of participants was 37 of which 23 were male students and 14 female, from 11 different language backgrounds as it is shown in Table 1. The data recordings took place twice within one month's time, before and after practicing at home. The duration of each recording session was approximately 30 minutes. In addition, 9 male and 2 female Swedish speakers without regional accent varieties were also recorded once each. Non-linguistic information, such as coughs, long pauses, repetitions or fillers was excluded from the final corpus used for experiments. Each speech file was accompanied by a text file, the content of which was adjusted to the actual utterance, thus any deletion or insertion that may have occurred was not considered into the text file. A phone-level transcription was then automatically generated from the speech signal and the text file, using an HMM-based aligner (Sjölander, 2003). These phone-level transcription files were used to separate the speech data into phoneme categories. The material contained all Swedish phonemes, but the two short and more open pre-r allophones /æ/, /œ/ and the retroflexes /ɳ/, /ɖ/, and /ɭ/ were not considered in the experiments because the number of occurrences in the database was not sufficiently large.

For each language background, the speech data were divided into different phoneme categories according to the phone-level transcription files. The speech signal was first pre-emphasized and then windowed every 25 ms with an overlap of 10 ms using a Hamming window. A discrete Fourier transform of 512 points was applied to the windowed frame to compute the signal's power spectrum.

3.2 Results

This section deals with the experiments and results of the described method. The goal is to identify a list of the most problematic phonemes for a given group of L2 speakers using previously recorded data. Hence, the experiments are done offline and the error detection was not made on an utterance basis but on the whole data for each phoneme category. The method is focusing on repeated mispronunciations made by the L2 speakers that deviate from the L1 speakers. Only the speech signal is considered without further linguistic or paralinguistic information. The list of problematic phonemes for each language group is then compared to a linguistic study (Bannert, 1984).

Table 2 lists the vowels identified by the PED algorithms as being problematic for the different groups of non-native speakers. For each L2 speaker group, the first line shows, in decreasing order, the most deviating vowels according to the Euclidean-based nPAD Θ_ℓ. Correspondingly,

L1 bkgr.	nPAD ver.	detected phonemes	missed phonemes accord. to Bannert (1984)
English	Θ_ℓ	æ̲ː̲, ɛ̲, y̲ː̲, u̲ː̲, ʊ, œ̲ː̲, ɛ̲ː̲, ø̲, ɵ̲, ø̲ː̲, (iː), ɑː, (ə), e̲ː̲, e̲, ɔ, a̲, ʉ̲ː̲	ʏ̲, o̲ː̲
(US)	$\Theta_\ell^{rbf\,250}$	e̲ː̲, ɛ̲ː̲, æ̲ː̲, ɑː, ʏ̲, ø̲ː̲, ɛ̲, a̲, ʉ̲ː̲, (i), y̲ː̲, (ə), (iː), ɔ, ɵ̲, ʊ, o̲ː̲, e̲	u̲ː̲, œ̲ː̲, ø̲
	$\Theta_\ell^{rbf\,500}$	æ̲ː̲, ɛ̲ː̲, e̲ː̲, ɑː, ɛ̲, ø̲ː̲, ʏ̲, a̲, ʉ̲ː̲, (i), y̲ː̲, (ə), (iː), e̲, œ̲ː̲, ɔ, ɵ̲, ʊ	u̲ː̲, o̲ː̲, ø̲
	$\Theta_\ell^{rbf\,1000}$	æ̲ː̲, ɛ̲ː̲, e̲ː̲, ɛ̲, ɑː, ø̲ː̲, ʏ̲, a̲, ʉ̲ː̲, œ̲ː̲, e̲, (i), ø̲, (ə), y̲ː̲, (iː), ɔ, ɵ̲	u̲ː̲, o̲ː̲, ʊ
German	Θ_ℓ	æː, (ɛ), y̲ː̲, u̲ː̲, (ʊ), ɛː, (ø), œː, øː, i̲ː̲, ə, ɑ̲ː̲	ʉ̲ː̲, ɵ̲, ʏ
	$\Theta_\ell^{rbf\,250}$	æː, (eː), ʏ, øː, ɑ̲ː̲, ʉ̲ː̲, ɛː, (i), y̲ː̲, (a), ə, i̲ː̲	u̲ː̲, ɵ̲, œː
	$\Theta_\ell^{rbf\,500}$	æː, (eː), ʏ, øː, ɑ̲ː̲, ɛː, ʉ̲ː̲, (a), (i), y̲ː̲, ə, i̲ː̲	u̲ː̲, ɵ̲, œː
	$\Theta_\ell^{rbf\,1000}$	æː, øː, ʏ, (eː), ɑ̲ː̲, ɛː, (a), ʉ̲ː̲, (i), œː, y̲ː̲, ə	u̲ː̲, ɵ̲, i̲ː̲
French	Θ_ℓ	æ̲ː̲, ɛ, y̲ː̲, u̲ː̲, œː, (ʊ), ɛː, (ø), ɵ̲, øː, i̲ː̲, ə, ɑ̲ː̲, eː, e, ɔ, (a)	ʉ̲ː̲, o̲ː̲, ʏ
	$\Theta_\ell^{rbf\,250}$	eː, ɛː, æ̲ː̲, ʏ, øː, ɑ̲ː̲, ʉ̲ː̲, (a), (i), y̲ː̲, i̲ː̲, ə, ɛ, ɔ, o̲ː̲, (ʊ), ɵ̲	u̲ː̲, œː, e
	$\Theta_\ell^{rbf\,500}$	æ̲ː̲, eː, ɛː, ʏ, øː, ɑ̲ː̲, (a), ʉ̲ː̲, ɛ, (i), y̲ː̲, ə, i̲ː̲, œː, ɔ, e, o̲ː̲	u̲ː̲, ɵ̲
	$\Theta_\ell^{rbf\,1000}$	æ̲ː̲, ɛː, eː, øː, ʏ, ɑ̲ː̲, ɛ, (a), œː, ʉ̲ː̲, (i), (ø), e, y̲ː̲, ə, i̲ː̲, ɔ	u̲ː̲, ɵ̲, o̲ː̲
Polish	Θ_ℓ	æ̲ː̲, (ɛ), y̲ː̲, u̲ː̲, œː, (ʊ), ɛ̲ː̲, ø̲, ɵ̲, ø̲ː̲, i̲ː̲, ɑ̲ː̲, ə, e̲ː̲, (e), (ɔ)	ʉ̲ː̲, ʏ̲, o̲ː̲, a
	$\Theta_\ell^{rbf\,250}$	e̲ː̲, ɛ̲ː̲, æ̲ː̲, ɑ̲ː̲, ʏ̲, ø̲ː̲, (ɛ), a, ʉ̲ː̲, (i), y̲ː̲, i̲ː̲, ə, (ɔ), o̲ː̲, ɵ̲	u̲ː̲, ø̲, œː
	$\Theta_\ell^{rbf\,500}$	æ̲ː̲, ɛ̲ː̲, e̲ː̲, ɑ̲ː̲, (ɛ), ø̲ː̲, ʏ̲, a, ʉ̲ː̲, (i), y̲ː̲, ə, (e), i̲ː̲, œː, (ɔ)	u̲ː̲, ɵ̲, o̲ː̲, ø̲
	$\Theta_\ell^{rbf\,1000}$	æ̲ː̲, ɛ̲ː̲, (ɛ), e̲ː̲, ɑ̲ː̲, ø̲ː̲, ʏ̲, a, œː, ʉ̲ː̲, (e), (i), ø̲, ə, y̲ː̲, i̲ː̲	u̲ː̲, o̲ː̲, ɵ̲
Russian	Θ_ℓ	ɛ, y̲ː̲, ɛ̲ː̲, ø̲, ø̲ː̲, i̲ː̲, ɑ̲ː̲, ɵ̲, e̲ː̲, e, ʉ̲ː̲, a, (ɔ), ʏ̲, (ə), (i), œː	u̲ː̲, æ̲ː̲, oː
	$\Theta_\ell^{rbf\,250}$	ʏ̲, ɛ̲ː̲, ʉ̲ː̲, ø̲ː̲, y̲ː̲, e̲ː̲, ɑ̲ː̲, i̲ː̲, a, ɵ̲, (ɔ), æ̲ː̲, (i), (ə), œː, e, u̲ː̲	ø̲, oː, ɛ
	$\Theta_\ell^{rbf\,500}$	ʏ̲, ɛ̲ː̲, ʉ̲ː̲, ø̲ː̲, y̲ː̲, e̲ː̲, ɑ̲ː̲, a, œː, i̲ː̲, æ̲ː̲, ø̲, ɵ̲, e, (ɔ), (i), (ə)	u̲ː̲, oː, ɛ
	$\Theta_\ell^{rbf\,1000}$	ʏ̲, ɛ̲ː̲, ø̲ː̲, ʉ̲ː̲, œː, ø̲, ɑ̲ː̲, e̲ː̲, y̲ː̲, a, i̲ː̲, e, æ̲ː̲, ɛ, ɵ̲, (ɔ), (i)	u̲ː̲, oː
Greek	Θ_ℓ	æː, (ɛ), y̲ː̲, u̲ː̲, œː, (ʊ), ɛ̲ː̲, ø̲, ɵ̲, ø̲ː̲, i̲ː̲, ɑ̲ː̲, ə, e̲ː̲, e, (ɔ)	ʉ̲ː̲, ʏ̲, o̲ː̲
	$\Theta_\ell^{rbf\,250}$	ɛ̲ː̲, e̲ː̲, æː, ɑ̲ː̲, ʏ̲, ø̲ː̲, (ɛ), (a), ʉ̲ː̲, (i), y̲ː̲, i̲ː̲, ə, (ɔ), ɵ̲, o̲ː̲	u̲ː̲, ø̲, œː, e
	$\Theta_\ell^{rbf\,500}$	æː, ɛ̲ː̲, e̲ː̲, ɑ̲ː̲, (ɛ), ʏ̲, ø̲ː̲, (a), ʉ̲ː̲, (i), y̲ː̲, ə, e, i̲ː̲, œː, (ɔ)	u̲ː̲, ɵ̲, o̲ː̲, ø̲
	$\Theta_\ell^{rbf\,1000}$	æː, ɛ̲ː̲, (ɛ), e̲ː̲, ɑ̲ː̲, ø̲ː̲, ʏ̲, (a), œː, e, ʉ̲ː̲, (i), ø̲, ə, y̲ː̲, i̲ː̲	u̲ː̲, ɵ̲, o̲ː̲
Spanish	Θ_ℓ	u̲ː̲, æː, ɛ̲, ø̲, i̲ː̲, ɵ̲, e̲ː̲, e̲, ø̲ː̲, (a), ʉ̲ː̲, ɑ̲ː̲, (i), ə, ʏ̲, (ɔ), o̲ː̲	y̲ː̲, ɛ̲ː̲, œ̲ː̲
	$\Theta_\ell^{rbf\,250}$	ʉ̲ː̲, æː, e̲ː̲, (a), ɛ̲ː̲, i̲ː̲, ʏ̲, (i), ɵ̲, ɑ̲ː̲, ø̲ː̲, ə, (ɔ), e̲, œ̲ː̲, o̲ː̲, (ʊ)	u̲ː̲, y̲ː̲, ɛ̲, ø̲
	$\Theta_\ell^{rbf\,500}$	ʉ̲ː̲, (a), e̲ː̲, ʏ̲, i̲ː̲, (i), ɛ̲ː̲, ɑ̲ː̲, ɵ̲, œ̲ː̲, æː, ø̲ː̲, e̲, ə, (ɔ), ɛ̲, ø̲	u̲ː̲, y̲ː̲, o̲ː̲
	$\Theta_\ell^{rbf\,1000}$	ʉ̲ː̲, (a), e̲ː̲, ʏ̲, (i), i̲ː̲, œ̲ː̲, e̲, ø̲, ø̲ː̲, ɛ̲, ɵ̲, ɑ̲ː̲, ə, æː, ɛ̲ː̲, (ɔ)	u̲ː̲, y̲ː̲, o̲ː̲
Turkish	Θ_ℓ	æ̲ː̲, (ɛ), y̲ː̲, (ɛː), (ø), u̲ː̲, ɵ̲, ʊ̲, ø̲ː̲, i̲ː̲, ɑː, e̲ː̲, (ə)	e̲, ʉ̲ː̲, o̲ː̲, œː
	$\Theta_\ell^{rbf\,250}$	e̲ː̲, (ɛː), (ʏ), ø̲ː̲, æ̲ː̲, (a), ʉ̲ː̲, (i), ɑː, i̲ː̲, y̲ː̲, (ɔ), (ə)	e̲, u̲ː̲, ʊ̲, ɵ̲, o̲ː̲, œ̲ː̲
	$\Theta_\ell^{rbf\,500}$	e̲ː̲, (ʏ), æ̲ː̲, (ɛː), ø̲ː̲, (a), ɑː, (i), ʉ̲ː̲, (ɛ), i̲ː̲, y̲ː̲, (ɔ)	e̲, u̲ː̲, ʊ̲, ɵ̲, o̲ː̲, œ̲ː̲
	$\Theta_\ell^{rbf\,1000}$	e̲ː̲, æ̲ː̲, (ʏ), (ɛː), ø̲ː̲, (a), (ɛ), ɑː, (i), ʉ̲ː̲, (o), œ̲ː̲, e̲	u̲ː̲, ʊ̲, ɵ̲, o̲ː̲, y̲ː̲, i̲ː̲
Arabic	Θ_ℓ	æ̲ː̲, ɛ, y̲ː̲, u̲ː̲, œː, (ʊ), ɛ̲ː̲, ø̲, ɵ̲, ø̲ː̲, i̲ː̲, ɑ̲ː̲, ə, e̲ː̲, e, (ɔ), a, ʉ̲ː̲	ʏ̲, o̲ː̲
	$\Theta_\ell^{rbf\,250}$	e̲ː̲, ɛ̲ː̲, æ̲ː̲, ɑ̲ː̲, ʏ̲, ø̲ː̲, ɛ, a, ʉ̲ː̲, (i), y̲ː̲, i̲ː̲, ə, (ɔ), o̲ː̲, (ʊ), ɵ̲, e	u̲ː̲, ø̲, œː
	$\Theta_\ell^{rbf\,500}$	æ̲ː̲, e̲ː̲, ɛ̲ː̲, ɑ̲ː̲, ɛ, ø̲ː̲, ʏ̲, a, ʉ̲ː̲, (i), y̲ː̲, ə, e, i̲ː̲, œː, (ɔ), (ʊ), ɵ̲	u̲ː̲, o̲ː̲, ø̲
	$\Theta_\ell^{rbf\,1000}$	æ̲ː̲, ɛ̲ː̲, e̲ː̲, ɛ, ɑ̲ː̲, ø̲ː̲, ʏ̲, a, œː, ʉ̲ː̲, e, (i), ø̲, ə, y̲ː̲, i̲ː̲, (ɔ), ɵ̲	u̲ː̲, o̲ː̲
Chinese	Θ_ℓ	ɵ̲, æː, ɛ, y̲ː̲, u̲ː̲, ɛː, ø, ø̲ː̲, iː, ɑː, eː, e, ɔ, (ə), a̲, ʉ̲ː̲, (i), o̲ː̲	ʏ̲, œː
	$\Theta_\ell^{rbf\,250}$	ɵ̲, ɑː, ʉ̲ː̲, eː, (i), ɛː, ø̲ː̲, æː, iː, ɔ, a̲, y̲ː̲, o̲ː̲, (ə), u̲ː̲, e, (ʊ), œː	ʏ̲, ø, ɛ
	$\Theta_\ell^{rbf\,500}$	ɵ̲, ɑː, æː, ʉ̲ː̲, ɛː, eː, ø̲ː̲, (i), a̲, ɔ, iː, y̲ː̲, o̲ː̲, œː, e, (ə), u̲ː̲, ɛ	ʏ̲, ø
	$\Theta_\ell^{rbf\,1000}$	ɑː, æː, ɛː, ø̲ː̲, ʉ̲ː̲, eː, (i), a̲, œː, ɔ, e, iː, ɛ, o̲ː̲, y̲ː̲, ø, (ə), u̲ː̲	ʏ̲, ɵ̲
Persian	Θ_ℓ	ɵ̲, æ̲ː̲, y̲ː̲, (ɛ), ø̲, ø̲ː̲, ʏ̲, (i), ʉ̲ː̲, e̲ː̲, a̲, (e), (ɑː), o̲ː̲, (ɔ)	i̲ː̲, u̲ː̲, ɛ̲ː̲, œ̲ː̲, ə
	$\Theta_\ell^{rbf\,250}$	ɵ̲, æ̲ː̲, (e), (ɛ), ə, ɛ̲ː̲, a̲, ø̲ː̲, e̲ː̲, ʏ̲, (i), ʉ̲ː̲, (ɑː), u̲ː̲, œ̲ː̲	i̲ː̲, y̲ː̲, o̲ː̲, ø̲
	$\Theta_\ell^{rbf\,500}$	æ̲ː̲, ɵ̲, (e), (ɛ), ø̲ː̲, ə, ɛ̲ː̲, œ̲ː̲, ʉ̲ː̲, ø̲, ʏ̲, e̲ː̲, a̲, (i), (ɑː)	i̲ː̲, u̲ː̲, y̲ː̲, o̲ː̲
	$\Theta_\ell^{rbf\,1000}$	æ̲ː̲, ø̲ː̲, ø̲, œ̲ː̲, ɛ̲ː̲, (e), (ɛ), ɵ̲, ʏ̲, ʉ̲ː̲, (i), ə, e̲ː̲, (ɑː), a̲	i̲ː̲, y̲ː̲, o̲ː̲, u̲ː̲

Table 2: Problematic vowels per language background. To the left, the vowels are shown in decreasing order, starting from the one with the highest nPAD. Phonemes that differ from the linguistic study findings are in parentheses, and the seriously problematic according to Bannert (1984) are underscored. To the right, the missed vowels.

	Θ_ℓ	Θ_ℓ^{rbf250}	Θ_ℓ^{rbf500}	$\Theta_\ell^{rbf1000}$
Better performance in no. of language groups	3	2	3	***4***
Mismatches with theory (total)	***19.2%***	22.0%	20.9%	***19.2%***
Seriously problematic phonemes missed (total)	***21.3%***	22.0%	26.0%	25.2%
Mismatches in top 5 phonemes	8	***7***	8	***7***
Seriously problematic captured in top 5 phonemes	34	***35***	34	34

Table 3: Summary of findings for vowels.

the second, third and fourth lines show the results of the RBF kernel-based nPAD Θ_ℓ^{rbf} for $\sigma^2 = 0.002$, 0.001 and 0.0005, respectively. These are shown in table as Θ_ℓ^{rbf250}, Θ_ℓ^{rbf500} and $\Theta_\ell^{rbf1000}$, respectively because γ in Eqs. (1) and (2) becomes 250, 500 and 1000, respectively. As ground truth is considered the linguistic survey described in (Bannert, 1984). False rejections according to (Bannert, 1984) are indicated in parentheses and false accepts according to (Bannert, 1984) are listed in the right-most column. Some phonemes are shown underscored. These are the seriously problematic phonemes according to (Bannert, 1984), i.e., they are totally mispronounced by the non-native speakers. Generally, the nPAD methods capture most of the common errors made by each language group when its members are trying to learn Swedish. The Euclidean-based nPAD Θ_ℓ is better for American English, Spanish and Turkish speakers. The RBF kernel-based nPAD Θ_ℓ^{rbf250} is better for Polish, Θ_ℓ^{rbf500} is better for French and Chinese and $\Theta_\ell^{rbf1000}$ is better for Russian, Greek, Arabic and Persian speakers. For the German speaking group both Θ_ℓ^{rbf250} and Θ_ℓ^{rbf500} perform equally well.

Table 3 summarizes the findings of the approaches for vowels. The Euclidean-based measure achieves a lower percentage of mismatches with the theoretical linguistic findings and also misses less seriously mispronounced vowels compared to the RBF kernel-based measures. On the other hand, Θ_ℓ^{rbf250} captures the most seriously problematic vowels of all methods when looking only at the top 5 vowels of the list of problematic ones and also has the least mismatches with Bannert, again when only the five most problematic phonemes according to the method are considered. Θ_ℓ^{rbf500} seems not achieving better performance compared to the rest of the methods according to the table list and finally, $\Theta_\ell^{rbf1000}$ is generally performing better in more groups of L2 speakers, has the least mismatches with theory (as Θ_ℓ does, too) and also has less mismatches when only the top 5 most problematic vowels are considered (the same as Θ_ℓ^{rbf250}).

The reported findings show clearly that for many groups of L2 speakers the open pre-r allophone /æː/ is very problematic as it appears most of times at the top of the problematic vowels. Another vowel that appears problematic is /ɛː/, which is often not pronounced with a long duration as it is supposed but rather short, often being replaced by /ɛ/. Generally, it is revealed that most of the foreign speakers face difficulties when trying to produce the Swedish long vowels. Hence, /ʉː/, /ɑː/ and /eː/ are vowels that both the tested methods and Bannert's linguistic survey diagnose as seriously problematic for most of the L2 groups.

Table 4 lists the consonants that are diagnosed as being mispronounced by the L2 groups. As in Table 2, the first row shows, in decreasing order, the most deviating consonants according to the Euclidean-based nPAD Θ_ℓ and the following rows the three RBF kernel-based nPADs Θ_ℓ^{rbf250}, Θ_ℓ^{rbf500} and $\Theta_\ell^{rbf1000}$, respectively. Θ_ℓ is better for French, Greek, Spanish and Turkish speakers. $\Theta_\ell^{rbf1000}$ is better for Persian speakers while all three RBF kernel-based nPADs are

L1 bkgr.	nPAD ver.	detected phonemes	missed phonemes accord. to Bannert (1984)
English	Θ_ℓ	ɧ̲, ŋ̲, (v), m, n, (b), r̲, (d), l̲, k, ʂ, t	s̲, ɕ̲, ʈ
(US)	Θ_ℓ^{rbf250}	ʂ, ɕ̲, s̲, (j), ɧ̲, r̲, l̲, (g), (d), ŋ̲, k, t	m, n, ʈ
	Θ_ℓ^{rbf500}	ʂ, ɕ̲, s̲, (j), r̲, l̲, ɧ̲, (g), ŋ̲, k, (d), t	m, n, ʈ
	$\Theta_\ell^{rbf1000}$	ʂ, ɕ̲, s̲, (j), r̲, l̲, (g), k, ŋ̲, ɧ̲, t, (d)	m, n, ʈ
German	Θ_ℓ	ɧ̲, ŋ̲, v̲, n, (m), b̲, r̲, d̲, (l), k, ʂ̲, t, p, (h), f, ɕ, s̲	g̲, ʈ̲, j
	Θ_ℓ^{rbf250}	ʂ̲, ɕ, s̲, r̲, (l), ɧ̲, g̲, ŋ̲, d̲, k, t, b̲, (h), f, v̲, n, p	ʈ̲, j
	Θ_ℓ^{rbf500}	ɕ, ʂ̲, s̲, r̲, (l), g̲, ŋ̲, d̲, k, ɧ̲, t, b̲, (h), f, v̲, n, p	ʈ̲, j
	$\Theta_\ell^{rbf1000}$	ɕ, ʂ̲, s̲, r̲, (l), g̲, ŋ̲, k, d̲, t, (h), b̲, f, v̲, n, ɧ̲, j	ʈ̲, p
French	Θ_ℓ	ŋ̲, ɧ, (v), m, n, b̲, r̲, (l), d̲, ʂ, k̲, t̲, p̲, h, ɕ̲, g̲	s, ʈ
	Θ_ℓ^{rbf250}	ʂ, ɕ̲, (j), r̲, (l), g̲, ŋ̲, s, d̲, ɧ, b̲, k̲, h, t̲, m, (v)	p̲, ʈ, n
	Θ_ℓ^{rbf500}	ʂ, ɕ̲, (j), r̲, (l), g̲, ŋ̲, s, d̲, b̲, k̲, h, t̲, m, ɧ, n	p̲, ʈ
	$\Theta_\ell^{rbf1000}$	ʂ, ɕ̲, (j), r̲, (l), ŋ̲, s, g̲, k̲, d̲, b̲, h, t̲, m, n, (v)	p̲, ʈ, ɧ
Polish	Θ_ℓ	ɧ̲, ŋ̲, v̲, (m), n, b̲, (r), d̲, (l), k̲, ʂ̲, t̲, p̲, s̲, (f)	g̲, h̲, ɕ, ʈ
	Θ_ℓ^{rbf250}	ʂ̲, s̲, ɕ, (j), ɧ̲, (l), (r), g̲, ŋ̲, d̲, t̲, k̲, b̲, (f), v̲	h̲, p̲, n, ʈ
	Θ_ℓ^{rbf500}	ʂ̲, s̲, ɕ, (j), (r), (l), ɧ̲, g̲, ŋ̲, k̲, t̲, b̲, d̲, (f), v̲	h̲, p̲, n, ʈ
	$\Theta_\ell^{rbf1000}$	ʂ̲, s̲, ɕ, (j), (r), (l), g̲, ŋ̲, k̲, t̲, ɧ̲, b̲, d̲, (f), v̲	h̲, p̲, n, ʈ
Russian	Θ_ℓ	v̲, ŋ̲, (m), (n), (r), d̲, (l), h, b̲, k̲, t̲, g̲, (s), f, j	p̲, ɧ̲, ʂ̲, ɕ, ʈ
	Θ_ℓ^{rbf250}	j, (l), (s), (r), ŋ̲, g̲, v̲, k̲, f, (m), (n), t̲, b̲, d̲, p̲	ɧ̲, ʂ̲, ɕ, h, ʈ
	Θ_ℓ^{rbf500}	j, (s), (l), (r), ŋ̲, g̲, v̲, k̲, (m), f, (n), t̲, b̲, d̲, p̲	ɧ̲, ʂ̲, ɕ, h, ʈ
	$\Theta_\ell^{rbf1000}$	j, (s), (l), (r), ŋ̲, g̲, v̲, k̲, (m), (n), f, t̲, b̲, d̲, p̲	ɧ̲, ʂ̲, ɕ, h, ʈ
Greek	Θ_ℓ	ɧ̲, ŋ̲, (v), m, n̲, b̲, (r), d̲, l, ʂ̲, k̲, t̲, p̲, ɕ̲, (f), s̲	g̲, h̲, t̲
	Θ_ℓ^{rbf250}	ʂ̲, ɕ̲, (j), s̲, ɧ̲, (r), l, g̲, ŋ̲, d̲, t̲, b̲, k̲, (f), (v), m	h̲, n̲, p̲, ʈ
	Θ_ℓ^{rbf500}	ʂ̲, ɕ̲, (j), s̲, (r), l, g̲, ɧ̲, ŋ̲, t̲, d̲, k̲, b̲, (f), (v), m	h̲, n̲, p̲, ʈ
	$\Theta_\ell^{rbf1000}$	ʂ̲, ɕ̲, (j), s̲, (r), l, g̲, ŋ̲, t̲, k̲, d̲, b̲, ɧ̲, (f), (v), m	h̲, n̲, p̲, ʈ
Spanish	Θ_ℓ	ŋ̲, v̲, (r), n, (l), b̲, t̲, ʂ̲, (f), k̲, g̲, d̲, j, p̲, ɕ̲, s̲, h̲	ɧ̲, m, ʈ
	Θ_ℓ^{rbf250}	(l), (r), ʂ̲, ŋ̲, (f), v̲, n, t̲, k̲, g̲, p̲, d̲, b̲, m, h̲, j, ʈ	s̲, ɧ̲, ɕ̲
	Θ_ℓ^{rbf500}	(l), ʂ̲, (r), ŋ̲, (f), v̲, n, t̲, k̲, p̲, g̲, d̲, m, b̲, j, h̲, ʈ	s̲, ɧ̲, ɕ̲
	$\Theta_\ell^{rbf1000}$	(l), ʂ̲, (r), ŋ̲, (f), v̲, n, t̲, k̲, g̲, p̲, d̲, m, b̲, j, h̲, ʈ	s̲, ɧ̲, ɕ̲
Turkish	Θ_ℓ	ŋ̲, v̲, (m), n̲, b̲, r, l, d̲, k̲, (ʂ), t̲, p̲, f, h, g̲, ʈ̲, j, s	ɧ̲, ɕ̲
	Θ_ℓ^{rbf250}	(ʂ), j, l, r, ŋ̲, g̲, k̲, t̲, b̲, s, v̲, f, d̲, (m), p̲, n̲, h, ʈ̲	ɧ̲, ɕ̲
	Θ_ℓ^{rbf500}	(ʂ), j, l, r, ŋ̲, g̲, k̲, t̲, s, b̲, v̲, f, (m), d̲, n̲, p̲, h, ʈ̲	ɧ̲, ɕ̲
	$\Theta_\ell^{rbf1000}$	(ʂ), j, l, r, ŋ̲, k̲, g̲, t̲, s, b̲, v̲, (m), d̲, f, n̲, p̲, h, ʈ̲	ɧ̲, ɕ̲
Arabic	Θ_ℓ	ɧ̲, ŋ̲, v̲, (m), (n), (b), r̲, d, (l), k̲, ʂ̲, t̲, p̲	f̲, s̲, ɕ̲, ʈ
	Θ_ℓ^{rbf250}	ʂ̲, ɕ̲, s̲, (j), ɧ̲, r̲, (l), (g), d, ŋ̲, k̲, t̲, (b)	f̲, p̲, v̲, ʈ
	Θ_ℓ^{rbf500}	ɕ̲, ʂ̲, s̲, (j), r̲, (l), ɧ̲, (g), k̲, d, ŋ̲, t̲, (b)	f̲, p̲, v̲, ʈ
	$\Theta_\ell^{rbf1000}$	ɕ̲, ʂ̲, s̲, (j), r̲, (l), k̲, (g), ŋ̲, t̲, d, ɧ̲, (b)	f̲, p̲, v̲, ʈ
Chinese	Θ_ℓ	ɧ̲, ŋ̲, v̲, m, n̲, b, r̲, l̲, d, k̲, t̲, f, g, ʈ, p̲, j̲, (h), (s)	ʂ̲, ɕ̲
	Θ_ℓ^{rbf250}	ɧ̲, l̲, r̲, ŋ̲, j̲, g, f, k̲, b, v̲, m, n̲, t̲, ʈ, p̲, d, (h), (s)	ʂ̲, ɕ̲
	Θ_ℓ^{rbf500}	l̲, r̲, ɧ̲, ŋ̲, j̲, g, k̲, f, b, v̲, m, n̲, t̲, ʈ, p̲, d, (h), (s)	ʂ̲, ɕ̲
	$\Theta_\ell^{rbf1000}$	l̲, r̲, ŋ̲, j̲, g, k̲, ɧ̲, b, f, m, v̲, n̲, t̲, ʈ, p̲, d, (h), (s)	ʂ̲, ɕ̲
Persian	Θ_ℓ	b, d, (ɧ), v̲, (f), g, (h), t̲, s̲, (j), ɕ̲, p̲, ʈ̲, k̲, l, r	n̲, ŋ̲, ʂ̲, m
	Θ_ℓ^{rbf250}	b, (f), (ɧ), v̲, g, t̲, (h), p̲, n̲, l, ʈ̲, r, d, ŋ̲, k̲, m	s̲, ʂ̲, ɕ̲
	Θ_ℓ^{rbf500}	b, (f), (ɧ), v̲, g, t̲, (h), p̲, n̲, l, k̲, ʈ̲, r, ŋ̲, d, m	s̲, ʂ̲, ɕ̲
	$\Theta_\ell^{rbf1000}$	(f), v̲, (ɧ), g, t̲, p̲, (h), n̲, k̲, b, l, ʈ̲, ŋ̲, r, d, m	s̲, ʂ̲, ɕ̲

Table 4: Problematic consonants per language background. To the left, the consonants are shown in decreasing order, starting from the one with the highest nPAD. Phonemes that differ from the linguistic study findings are in parentheses, and the seriously problematic according to Bannert (1984) are underscored. To the right, the missed consonants.

	Θ_ℓ	Θ_ℓ^{rbf250}	Θ_ℓ^{rbf500}	$\Theta_\ell^{rbf1000}$
Better performance in no. of language groups	5	***6***	5	5
Mismatches with theory (total)	20.2%	20.2%	***19.7%***	20.2%
Seriously problematic phonemes missed (total)	19.5%	***18.6%***	***18.6%***	***18.6%***
Mismatches in top 5 phonemes	***15***	16	18	18
Seriously problematic captured in top 5 phonemes	28	***29***	27	27

Table 5: Summary of findings for consonants.

equally better for American English, German, Russian and Arabic speakers in comparison to the Euclidean-based measure. Finally, Θ_ℓ and Θ_ℓ^{rbf250} are better for the Polish group and Θ_ℓ^{rbf250} and Θ_ℓ^{rbf500} for the Chinese speakers.

Table 5 summarizes the findings of the Table 4. The Euclidean-based measure has less mismatches with Bannert in the top five most problematic consonants as compared to the RBF kernel-based approaches. Θ_ℓ^{rbf250} is better in most language groups and can better capture the seriously problematic consonants both when the focus is on the five most problematic ones, but also in terms of the total number of the seriously problematic consonants (although in the latter case, all three RBF kernel-based approaches perform equally good). In addition, Θ_ℓ^{rbf500} has the least mismatches with the linguistic study.

The Swedish retroflex /ʂ/ is very problematic according to the reported results and likewise the unique "sje-sound" /ɧ/, a rounded velar fricative that does not exist in other languages. Moreover, many L2 speakers seem to have problems producing the velar nasal /ŋ/, which is commonly mispronounced as /ŋg/. Another difficult consonant is the fricative /ɕ/ that is also one of the most problematic sounds for second language speakers of Swedish.

In summary, RBF kernel generally seems to work better for consonants vis-à-vis vowels when compared with (Bannert, 1984) but also with the Euclidean distance measure. A small improvement in the percentage of the seriously problematic consonants is confirmed – accomplished by all three RBF-based measures – compared to the Euclidean measure. The figures remain better even in the case in which only the five most problematic consonants are taken into account. The results of the RBF kernel metrics are still in a better agreement with linguistic findings in comparison to the Euclidean-based one. On the other hand for the case of vowels, the two metrics perform nearly the same based on the criteria listed in Table 3. While for instance, Θ_ℓ is better considering the total number of mismatches with theory and, in addition, misses less seriously problematic vowels according to Bannert's study, Θ_ℓ^{rbf250} has a slightly better performance when concentrating on the five most problematic vowels and $\Theta_\ell^{rbf1000}$ is mainly preferable for more L2 groups compared to the rest of the metrics. Generally speaking, RBF-kernel may be considered to outperform to a small extent the Euclidean measure, though the two measures do not have major differences and they both seem to work well and achieve positive results as they regurarly agree with Bannert's linguistic survey. It is noted that the intention of the research described in this paper was to investigate alternative measures for the perceptually-motivated PED approach and carry out experiments to explore their behavior. Moreover, the deviations from the theoretical findings that both distance measures have can, for the most part, be explained by the nature of the two studies (theoretical linguistic *vs.* computational automatic) and the methodology that was followed. Bannert studied the pronunciation problem from a pure linguistic perspective, including lots of subjective observations and analysis. The computational methods do not consider many linguistic aspects, such as context and influence from preceding or succeeding phonemes. In addition, the PED methods aim at diagnosing mispronunciations made by the examined learners and are not designed to be used for identifying general problems related to the L1 of a group of speakers as Bannert's study was. It is noted that Bannert collected data from L2 speakers that were not influenced by repeating after a native speaker. The reason was that the study was aimed at making an inventory of mispronunciations for various groups of L2 students that would be used as a reference list for the teachers of Swedish as a second language. This may partly explain why some of the seriously problematic phonemes in Bannert's study were not diagnosed likewise with the nPAD approaches.

4 Conclusions

In this paper, a RBF kernel-based similarity measure was investigated as part of a pronunciation error detection algorithm previously presented in (Koniaris and Engwall, 2011; Koniaris et al., 2013) where a Euclidean distance measure was utilized. The idea was to investigate whether it can achieve good performance in relation to relevant linguistic literature and in comparison to the Euclidean similarity measure. The experiments show that good results can be obtained using this measure. In the future, it will be interesting to extend the idea by applying support vector machines in combination to the RBF kernel measure.

References

Bannert, R. (1984). Problems in learning Swedish pronunciation and in understanding foreign accent. *Folia Linguistica*, 18(1-2):193–222.

Braun, M. L., Buhmann, J. M., and Müller, K.-R. (2008). On relevant dimensions in kernel feature spaces. *J. Machine Learn. Research*, 9:1875–1908.

Delmonte, R. (2000). SLIM prosodic automatic tools for self-learning instruction. *Speech Communication*, 30(2-3):145–166.

Flege, J. E. (1995). *Second-language speech learning: theory, findings, and problems*. Strange, W. (Ed.), Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research. Timonium, MD: York Press Inc.

Franco, H., Neumeyer, L., Kim, Y., and Ronen, O. (1997). Automatic pronunciation scoring for language instruction. In *IEEE Int. Conf. Acoust., Speech, Sig. Proc., Munich, Germany*, pages 1471–1474.

Guion, S. G., Flege, J. E., Ahahane-Yamada, R., and Pruitt, J. C. (2000). An investigation of current models of second language speech perception: the case of japanese adults' perception of english consonants. *J. Acoust. Soc. Am.*, 107(5):2711–2724.

Koniaris, C. and Engwall, O. (2011). Phoneme level non-native pronunciation analysis by an auditory model-based native assessment scheme. In *Interspeech, Florence, Italy*, pages 1157–1160.

Koniaris, C., Salvi, G., and Engwall, O. (2013). On mispronunciation analysis of individual foreign speakers using auditory periphery models. *Speech Communication*, 55(5):691–706.

Neumeyer, L., Franco, H., Digalakis, V., and Weintraub, M. (2000). Automatic scoring of pronunciation quality. *Speech Communication*, 30:83–93.

Neumeyer, L., Franco, H., Weintraub, M., and Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. In *Int. Conf. Spoken Lang. Proc., Philadelphia, PA , USA*, pages 1457–1460.

Park, J. G. and Rhee, S. C. (2004). Development of the knowledge-based spoken english evaluation system and its application. In *ISCA Interspeech, Jeju Island, South Korea*, pages 1681–1684.

Piske, T., Flege, J., and MacKay, I. (2001). Factors affecting degree of foreign accent in an l2: a review. *J. Phonetics*, 29(2):191–215.

Raux, A. and Kawahara, T. (2002). Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning. In *Int. Conf. Spoken Lang. Proc., Denver, CO, USA*, pages 737–740.

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Fonetik*, pages 93–96.

Strik, H., Truong, K., de Wet, F., and Cucchiarini, C. (2009). Comparing different approaches for automatic pronunciation error detection. *Speech Communication*, 51(10):845–852.

Tepperman, J. and Narayanan, S. (2005). Hidden-articulator markov models for pronunciation evaluation. In *Proc. ASRU, San Juan, Puerto Rico*, pages 174–179.

Tepperman, J. and Narayanan, S. (2008). Using articulatory representations to detect segmental errors in nonnative pronunciation. *IEEE Tr. Audio, Speech, Lang. Proc.*, 16(1):8–22.

Truong, K. P., Neri, A., de Wet, F., Cucchiarini, C., and Strik, H. (2005). Automatic detection of frequent pronunciation errors made by L2-learners. In *ISCA Interspeech, Lisbon, Portugal*, pages 1345–1348.

van de Par, S., Kohlrausch, A., Charestan, G., and Heusdens, R. (2002). A new psychoacoustical masking model for audio coding applications. In *IEEE Int. Conf. on Acoust., Speech, Sig. Proc., Orlando, FL, USA*, volume 2, pages 1805–1808.

Wei, S., Hu, G., Hu, Y., and Wang, R.-H. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, 51(10):896–905.

Weigelt, L. F., Sadoff, S. J., and Miller, J. D. (1990). Plosive/fricative distinction: the voiceless case. *J. Acoust. Soc. Am.*, 87:2729–2737.

Wik, P. and Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech Communication*, 51(10):1024–1037.

Witt, S. M. and Young, S. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30:95–108.

Xu, S., Jiang, J., Chen, Z., and Xu, B. (2009). Automatic pronunciation error detection based on linguistic knowledge and pronunciation space. In *IEEE Int. Conf. Acoust. Speech Sig. Proc. (ICASSP), Taipei, Taiwan*, pages 4841–4844.

Yamashita, Y., Kato, K., and Nozawa, K. (2005). Automatic scoring for prosodic proficiency of english sentences spoken by japanese based on utterance comparison. *IECE Trans. Inform. Systems*, E88-D:496–501.