Perceptual Differentiation Modeling Explains Phoneme Mispronunciation by Non-Native Speakers

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

One of the difficulties in second language (L2) learning is the weakness in discriminating between acoustic diversity within an L2 phoneme category and between different categories. In this paper, we describe a general method to quantitatively measure the perceptual difference between a group of native and individual non-native speakers. Normally, this task includes subjective listening tests and/or a thorough linguistic study. We instead use a totally automated method based on a psycho-acoustic auditory model. For a certain phoneme class, we measure the similarity of the Euclidean space spanned by the power spectrum of a native speech signal and the Euclidean space spanned by the auditory model output. We do the same for a non-native speech signal. Comparing the two similarity measurements, we find problematic phonemes for a given speaker. To validate our method, we apply it to different groups of non-native speakers of various first language (L1) backgrounds. Our results are verified by the theoretical findings in literature obtained from linguistic studies.

Index Terms—second language learning, auditory model, distortion measure, perceptual differentiation ratio, phoneme.

1. INTRODUCTION

Learning how to speak a L2 without a foreign accent is often an arduous task, especially for learners from a different language family background. Computer-assisted pronunciation training (CAPT) programs aim at helping the student master the pronunciation of L2 phonemes. These programs have been used extensively to help learners master the pronunciation. However, these methods may not correspond to the perceptual distances of human listeners. Therefore, the methods may detect or accept deviations in the pronunciation differently than native speakers. To address this issue, we propose to use a model of the human auditory system, which has become available with the development of sophisticated models of the auditory periphery, e.g., [5, 6].

Recently, a feature-selection method based on human perception only, called auditory model-based feature selection, was presented for speech recognition [7, 8]. The method uses knowledge of the human auditory periphery to perform robust selection of feature subsets. Let the complete feature set characterize certain (locally) audible components of the speech signal. Then, for a given subset cardinality, an auditory model is used to select the feature subset that best captures the most audible of these signal components.

In this paper, we modify the above method to be used in detection of problematic phonemes for foreign speakers. The underlying idea is the observation that L2 speakers have a less acute perception of the acoustic features that define an L2 phoneme, and may therefore mispronounce it. The mispronunciation may either be that the speaker unsuccessfully produces the L2 phoneme with large variations between attempts (low precision) or consistently replaces it by another phoneme (low accuracy). Our method focuses on using the human auditory periphery to explain the first case (low precision) by giving an estimate of the L2 speaker’s perceptual affinity for every phoneme. We measure the similarity of the Euclidean geometry of the speech signal’s power spectrum and the human auditory representation of it for each phoneme class, both for a group of native speakers and for individual non-native speakers. We then compare these two measures and find, quantitatively, the phonemes for which the L2 speakers have considerable perceptual, and hence pronunciation, difficulties with the precision.

This paper is organized as follows. Sec. 2 discusses a similarity measure for the perceptual and speech frequency domains and presents an auditory model. Sec. 3 applies the method to phoneme mispronunciation of L2 speakers. Sec. 4 describes the data used for our experiments. Sec. 5 presents the results and compares them with theoretical findings and Sec. 6 provides conclusions.

2. MEASURING THE PERCEPTUAL RELEVANCE

The human auditory periphery system has a major role in the perception of sounds in general, and speech in particular. Recent experiments [9] reveal the importance of auditory periphery in auditory scene analysis [10], which is the process to organize a sound into meaningful, perceptual components.

We aim to measure the Euclidean similarities between the perceptual domain and the speech frequency domain. We assume that the mapping from the speech to the perceptual domain is a good...
approximation of distance preserving. For this, we define a measure that calculates the degree of dissimilarity between the perceptual domain and the speech frequency domain.

2.1. Geometric space dissimilarity

We are solely interested in highlighting the geometric similarities between the speech data and the output of a psycho-acoustic auditory model. In doing so, we focus on small distances between sounds. For this task, we adapt the dissimilarity measure introduced in [7] to model. In doing so, we focus on small distances between sounds.

A distortion measure in the speech frequency domain is a mapping of two signals: \( \Phi : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+ \), where \( \mathbb{R}^+ \) are the non-negative reals. Let \( \mathbf{x}_i \in \mathbb{R}^N \) be the \( N \)-dimensional speech signal vector of frame \( i \in \mathbb{Z} \) and \( \hat{x}_{i,j} \) be the \( j \)th perturbation of \( \mathbf{x}_i \). In our study we always consider the power spectrum of the speech signal, hence vector \( \mathbf{x}_i \) is the periodogram. A Euclidean norm-based measure is then

\[
\Phi(\mathbf{x}_i, \hat{x}_{i,j}) = \| \mathbf{x}_i - \hat{x}_{i,j} \|^2.
\]

Accordingly, a distortion measure in the auditory model (perceptual) domain is defined as: \( \Upsilon : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}^+ \). We consider the perceptual domain signals \( y(\mathbf{x}_i) \) and \( y(\hat{x}_{i,j}) \), where \( \mathbf{y} : \mathbb{R}^N \rightarrow \mathbb{R}^N \) is a mapping of \( \mathbf{x}_i \) and \( \hat{x}_{i,j} \), respectively, to the \( M \)-dimensional perceptual domain. Then

\[
\Upsilon(\mathbf{x}_i, \hat{x}_{i,j}) = \| y(\mathbf{x}_i) - y(\hat{x}_{i,j}) \|^2.
\]

We calculate the geometric dissimilarity between the perceptual-domain distances and the speech frequency-domain distances using the following measure of dissimilarity

\[
A = \frac{1}{f} \sum_{i \in I} \frac{1}{C_{f,j}} \sum_{j \in J_f} [\Upsilon(\hat{x}_{i,j}, \hat{x}_{i,j}) - \lambda \Phi(\mathbf{x}_i, \hat{x}_{i,j})]^2,
\]

where \( \lambda = \frac{\sum_{i \in I} \sum_{j \in J_f} \Upsilon(\hat{x}_{i,j}, \hat{x}_{i,j}) \Phi(\mathbf{x}_i, \hat{x}_{i,j})}{\sum_{i \in I} \sum_{j \in J_f} \Phi(\mathbf{x}_i, \hat{x}_{i,j})^2} \) is an optimal scaling factor [7] to dissolve undesirable mismatches between the two domains. Indexes \( i \in I \) and \( j \in J_f \) represent a finite frame sequence and a finite set of acoustic perturbations, respectively.

The computation of the output of an auditory model is most cases a difficult task. Therefore, we approximate the perceptual distortion measure \( \Upsilon(\mathbf{x}_i, \hat{x}_{i,j}) \) by a simpler quadratic measure by considering small distances. Combining perturbation analysis techniques and the sensitivity matrix [11, 12], we compute Eq. (3). Consider \( \Upsilon(\mathbf{x}_i, \hat{x}_{i,j}) \) to be known. We assume that \( \Upsilon(\mathbf{x}_i, \mathbf{x}_i) = 0 \) and that this forms a minimum. We furthermore assume that \( \Upsilon(\mathbf{x}_i, \hat{x}_{i,j}) \) is differentiable in \( \hat{x}_{i,j} \). Then, for sufficiently small perturbations \( \hat{x}_{i,j} \approx x_i \), we can make the approximation

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

where \( \mathbf{D}_\Upsilon(x) = \frac{\partial^2 \Upsilon(\mathbf{x}, \hat{x})}{\partial \hat{x}_i \partial \hat{x}_j} \bigg|_{\hat{x}_i = x_i, \hat{x}_j = x_j} \) is the sensitivity matrix.

This matrix is found using an auditory model.

2.2. Auditory model

The auditory model used is the van de Par psycho-acoustic model [5] that considers two sound signals, the masker and the maskee, to be presented at the same time. It consists of a filter, approximated by the inverse of the threshold of hearing in quiet, representing the outer and middle ear followed by a gammatone filterbank that models the basilar membrane in the inner ear. In each channel \( f \), the ratio of the distortion \( x - \hat{x} \) to masker \( x \) is estimated, where \( x \) denotes the magnitude spectrum of speech. In the end, all these ratios are combined, to simulate the spectral integration property of the human auditory system. The complete model is described by

\[
\Upsilon(x, \hat{x}) = C_a L_e \sum_{g \in G} \frac{1}{N} \sum_{i} \mathcal{F}(f)|x(f) - \hat{x}(f)|^2 + C_u,
\]

where \( \mathcal{F}(f) = |h_{om}(f)(\gamma_f(\hat{x}(f)))^2 \). \( h_{om} \) is the outer and middle ear transfer function and \( \gamma_f \) is the \( g \)th gammatone filter. Constants \( C_a \) and \( C_u \) are calibrated based on measurement data, \( L_e \) is the effective duration of the segment according to the temporal integration time of the human auditory system, the integer \( g \) labels the gammatone filter and finally, \( G \) is the set of gammatone filters considered.

The sensitivity matrix in the speech frequency domain is obtained by the van de Par model as a diagonal matrix with the diagonal element for row and column \( f \) given by

\[
D_{\Upsilon,f,f}(x) \approx 2 C_a L_e \sum_{i} \frac{1}{N} \sum_{j} \mathcal{F}(f)|x(f)\hat{x}(f)|^2 + C_u.
\]

3. NATIVE AND NON-NATIVE PERCEPTUAL AFFINITY

Our work is focused on studying the difference in the perceptual affinity of native and non-native speakers for different phonemes. The acoustic signal varies greatly between L1 speakers for the same phoneme, due to individual differences in age, sex, vocal tract length etc. Native speakers can nevertheless relatively easily process native speech and distinguish the acoustic details of their L1 phonemes. L2 speakers, on the contrary, may have a less distinct sound perception of the acoustic characteristics of the L2 and they are therefore inclined to mispronounce some of the L2 phonemes. They can either not produce them with precision, i.e., within the range of variability of native speakers, or they replace them with other phonemes, e.g., similar ones in their own language. Targeting the first case, we perform the sensitivity analysis described in Sec. 2.1 and expect that it will indicate phonemes constituting a perceptual challenge for L2 speakers, as manifested in their speech production.

To evaluate this perceptual differentiation between native and non-native listeners, we computed the van de Par sensitivity matrix \( D_{p,L}(x) \) for each speech segment \( p \), phoneme category \( p \) and language group of speakers \( L \). A set of 100 vectors \( \mathbf{x}_{i,j} \) was computed by adding 30 dB SNR i.i.d. Gaussian noise to \( x_i \). Next, the perceptual and speech distortion measures \( \Upsilon_p(x, \hat{x}_{i,j}) \) and \( \Phi_p(x, \hat{x}_{i,j}) \), respectively were calculated as well as the dissimilarity measure \( A_p^L \). Finally, the perceptual differentiation ratio \( R_p^L \) was computed for every phoneme and L1 background as

\[
R_p^L = \frac{A_p^L}{A_p^nat},
\]

where \( A_p^nat \) is the dissimilarity measure of a phoneme \( p \) for the native speaker group. That is, \( R_p^L \) quantifies to what extent the non-native dissimilarity differs from the native, with \( R_p^L \) increasing with how problematic the phoneme is.

4. DATA

A speech corpus, sampled at 16 kHz, was recorded to be used in our experiments. The corpus was designed for L2 learners of Swedish.
as a part of the Ville [13] program. Ville is an embodied conversational agent that acts as a virtual language tutor for Swedish. 37 (23 male and 14 female) speakers of different language backgrounds (c.f. Table 1) took the Ville 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. To find the perceptual similarity of natively uttered phonemes, 11 (9 males and 2 females) Swedish speakers were also recorded each.

All 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., deletions and insertions), the accompanying text file was corrected to match the actual content. A phone-level transcription was automatically created from the speech signal and the text file using an HMM-based aligner [14].

Finally, we considered the native speech files which were separated into phonemes according to the phone-level transcription files. We did the same for the non-native speech of each L1 group. The speech signal was first pre-emphasized and the output was windowed by a Hamming window of 25 ms with an overlap of 10 ms. A discrete Fourier transform (DFT) of 512 was applied to the windowed frame to compute the signal’s power spectrum.

### RESULTS AND DISCUSSION

In this section we present our experimental results and discuss them in comparison with a linguistic study of problematic Swedish phonemes for different L1 groups [15]. At this point, we want to emphasize the scope of our study. Our work should not be considered as a comprehensive linguistic study. Our attention is to consider the sound signal of the uttered phonemes and find perceptual differences among native and non-native speakers. Hence, any potential grammatical mistake or syntactical error caused by the complexity of a word is beyond the focus of this study. We are solely interested in how low precision at the segmental level for an L2 speaker is connected to sound perception.

Fig. 1(a-b) illustrates the value of $R^L_p$ for each L2 speaker group for six different (a) vowels, and (b) consonants. The two figures highlight difficulties in Swedish phonemes for each L2 group, such as: 1) English speakers have problems with the distinction between rounded /ɒ/ and unrounded /ʌ/. 2) Chinese speakers treat /l/ and /ł/ as allophones of /l̩/. 3) German speakers are prone to make voicing errors for /v/ and /l̩/, and /l̩/ is not shown in Fig. 1(b), since its $R^L_p$ value was much larger for some languages ($R^L_p$ up to 14,000) than the $R^L_p$ of the phonemes shown. The unexpectedly low $R^L_p$ for some phonemes in Fig. 1 can partly be explained by the context of the test (repeating after a native speaker) and partly by the fact that low accuracy is not captured by $R^L_p$ (i.e., the Swedish phoneme was consistently replaced by the L1 counterpart). Table 2 shows the results for each group of L2 speakers. The results of our perceptual-based method (to the left) are, in general, in agreement with previous linguistic observations [15] (to the right). Our experimental results are ordered, beginning with the phonemes that are the most challenging for each foreign speaker group (or the individual speakers in this group). The corresponding theoretical list to the right is not ordered, since no quantitative measure was used in [15]. The phonemes in bold are the ones determined to be most problematic ($R^L_p$ greater than one or through linguistic observation, respectively).

From the table, it can be seen that in most cases the experimental and linguistic results are in agreement. The few divergences in the results are reported in parentheses.

We believe that the deviations in the results can largely be explained by methodological differences. The aim in [15] was to identify problematic phonemes based on linguistic theory and subjective observations, whereas we only focus on the objective acoustic precision of the phone. This signifies, e.g., that phonemes were highlighted in [15] because they give rise to ephenesis or elision or because they are problematic in specific contexts (e.g., that French speakers mispronounce the Swedish /hp/ when it precedes a stressed vowel). Our method does not take the context, the phone position within a word, coarticulation, or the substitution of an L1 phoneme into account. On the other hand, our method identifies problematic phonemes for individual speakers, whereas [15] aimed at identifying common problems for speakers of a certain L1. Since the scope we aim for is to automatically diagnose which phonemes that should be targeted in CAPT practise for individual L2 learners, it is an inherent quality of our perceptual-based method that it identifies the problematic phonemes for the tested speaker(s), rather than general problems associated with the L1. The L1 is an important factor, but individual differences, such as fluency in Swedish or social background may have an even higher influence on the speaker’s accent.
6. CONCLUSIONS AND FUTURE WORK

We presented a method to automatically and quantitatively measure the perceptual differences between native and non-native speakers in perceiving the target phonemes, based on a spectral auditory model. The current method estimates native perception of native speech and non-native perception of non-native speech. This makes it possible to study how the mispronunciation occurs. Our future work is instead targeted on objectively identifying the most mispronounced phonemes, using a native-only perception model. We then envision integrating our method in a CAPT program to tailor the practice for individual L2 learners.

7. ACKNOWLEDGMENTS

The authors wish to thank Preben Wik for providing the speech data, and Giampietro Salvi for valuable comments on the manuscript. This work is supported by the Swedish Research Council project 80449001 Computer-Animated Language TEAchers (CALATEA).

8. REFERENCES


