EVALUATION OF SECOND LANGUAGE LEARNERS’ PRONUNCIATION USING HIDDEN MARKOV MODELS

Simo M.A. Godijn\textsuperscript{1} and Guus de Krom\textsuperscript{2}
\textsuperscript{1} Forensic Science Laboratory, Rijswijk
\textsuperscript{2} Computer and Humanities Department / Utrecht Institute of Linguistics-OTS
University of Utrecht, Trans 10, 3512 JK Utrecht, the Netherlands
Tel: + 31 30 2536-59, Fax: + 31 30 2536000, E-mail: Guus.deKrom@let.ruu.nl

ABSTRACT

In this study, Hidden Markov Models (HMMs) were used to evaluate pronunciation. Native and non-native speakers were asked to pronounce ten Dutch words. Each word was subsequently evaluated by an expert listener. Her main task was to decide whether a word was spoken by a native or a non-native speaker. For each word type, two versions of prototype HMMs were defined: one to be trained on tokens produced by a single native speaker, and another to be trained on tokens produced by a group of native speakers. For testing the different types of HMM, forced recognition was performed using native and non-native judged tokens. We expected that recognition with multi-speaker HMMs would allow a more effective discrimination between native and non-native tokens than recognition with single-speaker models. A comparison of Equal Error Rates partly confirmed this hypothesis.

INTRODUCTION

This paper describes an experiment in which Hidden Markov Models (HMMs) were applied for the evaluation of pronunciation in second language learning. Most studies on the automatic evaluation of pronunciation have dealt with the development and evaluation of aids for the speech and hearing impaired. These systems are often based on template-matching techniques. Although these systems have proven useful [1], they have a conceptual weakness, namely that the acoustic templates against which speech fragments are compared are usually based on utterances produced by just one speaker. A disadvantage of using speaker-dependent templates is that any acoustic deviation from a given template increases the pronunciation error score, even if the deviations have their origin in factors like age, gender, and speaking style, which are speaker related, rather than language specific. Systems that employ speaker-dependent templates are therefore prone to false rejection errors. An alternative could be the use of templates based on a speaker’s current “best utterance” [1]. However, to determine whether an utterance is the “best so far”, a supervisor must be present, which may be appropriate in certain settings, but is less suited for second language teaching purposes. Thus, instead of modelling the acoustic properties of utterances produced by a single native speaker, which results in templates that are more narrowly defined than desired, one could try to model the acoustic variance found in several realisations of a given word or utterance, as produced by a variety of native speakers. HMMs can be used for this purpose. Some researchers have performed experiments in which HMMs were used for the automatic evaluation of pronunciation in second language learning (although template-based methods are more common here too, [2]). In one experiment, HMMs were used to distinguish between minimal pairs [3], a task basically similar to an ordinary word recognition task. However, pronunciation errors may involve other types of error than segmental ones that lead to minimal pair differences. In this study, we performed an experiment in which HMMs were evaluated with respect to their ability to discriminate between native and non-native pronunciation. The HMMs were defined to model the acoustic variance of a composite group of native speakers, rather than just one native speaker (multi-speaker HMMs). To allow comparison with speaker-dependent methods for the evaluation of pronunciation [2], we also used HMMs based on the speech of a single native speaker (single-speaker HMMs).

Because the multi-speaker HMMs were trained to model the variance found in a group of speakers, we expected that these would yield lower false rejection rates than single-speaker HMMs. As for false acceptances: provided the multi-speaker HMMs are trained on a sufficient variety of word tokens, we expected that the single- and multi-speaker HMMs would perform equally well. Combining these expectations, we hypothesised that multi-speaker HMMs would perform better as regards the discrimination between native and non-native speakers than the single-speaker HMMs.

EXPERIMENTS

Recordings: material, speakers and procedure

Recordings were made of 53 native Dutch speakers (27 males, 26 females) and 49 non-native speakers (18 males, 31 females), varying in age between 17 and 60
years. None of the native speakers had an obvious regional accent. The non-native speakers had various language backgrounds and differed widely in their fluency in Dutch. Each speaker was asked to read a list of ten monosyllabic Dutch words that are found to be mispronounced regularly by foreign students: man /mɑn/, deur /ˈdœr/, echt /ˈɛxt/, rook /ˈrʊk/, nuijs /ˈnuiʃ/, maat /ˈmaːt/, kijk /ˈkiːk/, kwiekt /ˈkwiekt/, kind /ˈkɪnd/, and pot /ˈpɔt/. One male and one female native speaker were chosen as “norm speakers”. They were not selected on any specific criterion, although care was taken not to choose speakers likely to be judged as very idiosyncratic. These norm speakers each read the list ten times. Recordings were made in a sound-treated booth, with a condenser microphone on digital audio tape. Speakers were instructed to read the list with short breaks in between the words. To reduce possible begin- and end-of-list effects, the lists began and ended with two words that were not further used in the experiment. The recorded word tokens were downsampled to 9.8 kHz and stored on disk.

I. Perceptual evaluation of pronunciation

Listening test; tasks and procedures
The recorded words were perceptually evaluated by a professional teacher of Dutch as a foreign language. The listener’s primary task was to judge each word token as coming from a native or a non-native speaker, by clicking a corresponding button on a computer screen. She was allowed to listen to a specific word as often as desired, but once a judgement was made, it could not be recalled. In a first session, she judged 295 stimuli. In a second session those 295 stimuli were presented to her again, together with the remaining 955 stimuli, totalling 1250 stimuli (of which 750 were uttered by native speakers and 500 by non-native speakers). Within the second session, 20 stimuli were repeated once. The judgements on the stimuli that were presented twice were used to judge consistency. Prior to the first session, the listener was asked to do a test session involving ten words, in order to get used to the task.

Listening test; results
Some 6% of the words spoken by native speakers were judged non-native. Conversely, 12% of the words spoken by non-native speakers were judged native. All but one of the 200 word tokens produced by the two norm speakers were judged native. Native speakers did not always sound native; especially in the pronunciation of the word rook they were judged non-native a few times. A total of 295 stimulus words were judged in the first as well as in the second session. Of those, 22 (7.5%) were judged inconsistently as regards nativeness. Within the second session, twenty stimuli were offered to the listener twice. One of those was judged differently with regard to nativeness the second time. The listener was thus generally quite consistent in her judgements. We therefore felt confident to use her judgements as a yardstick measure for the HMM-based evaluation part of the experiment.

II. HMM-based evaluation of pronunciation

Speech file parameterisation
The recorded speech signals were converted to parameter files using the HCode tool of the Hidden Markov Toolkit [4]. Signal frames (duration 25.6 ms, shift interval 10 ms) were Hamming windowed, pre-emphasised, and converted into MFCC-format. The resulting parameter files contained data vectors with 26 coefficients: 12 MFCCs, one log energy coefficient, and their first-order derivatives.

HMM prototype definition
Three prototype whole-word HMMs were defined for each word: one multi-speaker HMM, trained with tokens produced by multiple speakers, and two single-speaker HMMs, trained with tokens produced by a single male and female speaker, respectively. The HMMs had three states per phoneme and diagonal covariance in the transition matrices. One-state skips were allowed. The multi-speaker HMMs had two mixtures, to model possible differences in the distributions of acoustic parameters for males and females.

HMM initialisation and training
Forty native-judged tokens (20 males and 20 females) were used to train the multi-speaker HMMs. The tokens were chosen randomly from the native material, and included a few words produced by fluent non-native speakers. Tokens that had been judged inconsistently with regard to nativeness were not used for training. The single-speaker HMMs were trained using tokens produced by the two norm speakers. The prototype HMMs were initialised using the Forward-Backward procedure, and further improved using Baum-Welch re-estimation [4, 5]. All models converged within 100 iterations.

HMM recognition
The speech material that had not been used for training was used for testing: it consisted of 129 native and 492 non-native tokens. For recognition, the Viterbi-algorithm was used [4, 5]. For every speech frame, the probability of the frame being generated by every state was calculated. The maximum of those probabilities defines what state is being used for the speech frame. The mean of those maxima over all the speech frames is output as the recognition value (MPPF, max. probability
per frame). As the HMMs were trained on native material, the tokens produced by native speakers were expected to yield higher MPPF values than tokens produced by non-native speakers. Thus, one could define an MPPF threshold, such that tokens with an MPPF exceeding the threshold are labelled native and tokens with a lower value as non-native. Ideally, there should be no overlap between the MPPF values for native and non-native tokens, in which case a (range of) MPPF value(s) can be defined which effectively separates native and non-native tokens. In practice, however, MPPF values for native and non-native tokens will overlap to a certain degree, which means that the choice for a given threshold results in a combination of false acceptance and false rejection errors (false acceptance: accepting a non-native pronunciation as native; false rejection: rejecting a native pronunciation as non-native). The problem is how to define an optimal threshold, as there is a trade-off between the two types of error. The question of which error rates are acceptable depends largely on the purpose for which the system is used: a beginning second language learner should not be discouraged by a system that hardly accepts anything he or she says. In such a case, the threshold should be biased towards false acceptance, i.e. at a low MPPF value. On the other hand, if a student already has a very good pronunciation, but strives for perfection, he or she should not be misled by a threshold biased to false acceptance. False rejection is in that case a more tolerable type of error. In the present study, we chose to consider a false rejection error equally bad as a false acceptance error. The native / non-native MPPF threshold was therefore defined by calculating the Equal Error Rate (EER), the rate at which the number of false rejections equalled the number of false acceptances. Figure 1 gives false rejection and false acceptance curves for the word kijk for both single- and multi-speaker HMMs.

**Results**

To determine the relative performance of the multi- and single-speaker HMMs, a number of tests were performed. In a first test, recognition was performed using all available HMMs (i.e. the one multi-speaker HMM and the two single-speaker HMMs for a total of 10 word types, 30 models in total). This test can be compared to a standard isolated word recognition experiment. Given the small number of different words, it was expected that recognition would be fairly good. Eighty-eight of the tokens (14 %) were nevertheless misrecognised. However, given the fact that 87 of these 88 confusions were tokens spoken by non-native speakers, and that all these tokens were judged non-native by the listener, we concluded that severe mispronunciation was the most probable cause of trouble, rather than inappropriate training of the HMMs. In a second test, three HMMs were used for recognition: the two single-speaker and the one multi-speaker HMM of the actual word to be evaluated (a forced recognition task). For all tokens, except those produced by the two norm speakers, MPPF values were higher for multi-speaker models than for single-speaker models. Thus, in a direct competition, the tokens would be recognised by the multi-speaker models, rather than the single-speaker models. Note that this does not necessarily mean that the single-speaker HMMs are inferior to the multi-speaker HMMs as regards evaluation of pronunciation, as it is the degree of overlap in the distributions of the MPPF values for native and non-native tokens which determines the false rejection and false acceptance error rates, rather than the actual
Table I: EER for multi-speaker HMMs and single-speaker HMMs for ten Dutch words.

<table>
<thead>
<tr>
<th>word type</th>
<th>man</th>
<th>deur</th>
<th>echt</th>
<th>rook</th>
<th>muis</th>
<th>maat</th>
<th>kijk</th>
<th>kwiek</th>
<th>kind</th>
<th>pot</th>
</tr>
</thead>
<tbody>
<tr>
<td>multi-speaker HMM</td>
<td>23</td>
<td>29</td>
<td>34</td>
<td>37</td>
<td>33</td>
<td>37</td>
<td>36</td>
<td>18</td>
<td>43</td>
<td>31</td>
</tr>
<tr>
<td>single-speaker HMM</td>
<td>30</td>
<td>29</td>
<td>39</td>
<td>25</td>
<td>55</td>
<td>48</td>
<td>50</td>
<td>27</td>
<td>41</td>
<td>19</td>
</tr>
</tbody>
</table>

magnitude of the obtained MPPFs. Performance of the single- and multi-speaker HMMs was compared by examining the EERs obtained with these two types of models. Our hypothesis was that multi-speaker HMMs generate lower false rejection rates than single-speaker models because multi-speaker HMMs are trained to incorporate the acoustic variability found in a group of speakers. To yield a low EER, low false rejection rates should not be traded off by high false acceptance rates. MODELS with a low EER were considered superior to models with a high EER. Results are given in Table I for each word type.

According to Table I, multi-speaker HMMs had lower EERs for six word types (man, echt, muis, maat, kijk and kwiek). Conversely, for two words -rook and pot-, EER was lower for the single-speaker HMMs than for the corresponding multi-speaker models. For two other words -deur and kind- the results were inconclusive.

In evaluating the results of this experiment, a number of shortcomings should be mentioned. First of all, the listener’s task was to judge if a given word token was spoken by a native or a non-native speaker. After the listening sessions, she remarked that some of the tokens that she judged as native could have been spoken by natives with a strong regional Dutch dialect. There were, however, no native speakers with a severe regional dialect among the subjects. Some tokens produced by non-native speakers might therefore have been erroneously assigned to the native part of the corpus. In retrospect, the listener should perhaps have been asked to judge if the tokens that she thought native were actually pronounced in the kind of non-regional manner that would be recommended to second language learners of Dutch. Had this been done, a possible degradation of the native HMMs could have been reduced. Second, we would like to emphasise that the HMMs were based on speech parameters that are conventionally used for speech recognition purposes. It might be that the distinction between native and non-native pronunciations requires different parameters, or at least a different degree of accuracy. Duration, for instance, is only partly modelled in HMMs by the allowance of state-skips in the transition matrix, but could be modelled more explicitly. Intonation is not modelled in this study: Yet, intonation is a very important cue for non-nativeness, one that was certainly used by the listener during the evaluation of the speech samples. For future investigation it is therefore recommended to model F0-information in some way.

CONCLUSION

The results of this experiment indicate that Hidden Markov Models can be used with some success to distinguish between native and non-native pronunciations of a word. We hypothesised that multi-speaker HMMs would be more effective in discriminating between native and non-native pronunciations than single-speaker models. The results were in agreement with this hypothesis for six out of the ten words that were investigated. For two words, single-speaker and multi-speaker EERs were approximately equal, whereas another two words got lower EERs on single-speaker models. The lower EERs for multi-speaker models could be attributed to a decrease in false rejection rates which was not traded off by an increase in false acceptances. We conclude that there is some evidence that an explicit modelling of the acoustic variability of native speakers may improve the accuracy of automated evaluation of second language pronunciation.

REFERENCES