Pronunciation Scoring for the Hearing-Impaired

Oytun Turk  Levent M. Arslan
BUMM, Electrical & Electronics Eng. Dept., Bogazici University, Istanbul, Turkey
SESTEK Inc., R&D Dept., Istanbul, Turkey
oytun@sestek.com.tr  arslanle@boun.edu.tr

Abstract

Automatic assessment of articulation and prosody is an important aid for speech therapy and language education. In this paper, we focus on speech therapy for hearing-impairment and propose methods for automatic articulation and prosody scoring. The pronunciation problems of the hearing-impaired are briefly discussed. Three methods are developed for automatic pronunciation scoring for the hearing-impaired. In the first method, Hidden Markov Model (HMM) based phoneme recognition and forced-alignment are employed for detecting articulation problems. The second method performs prosody scoring by operating at the syllable level and comparing the pitch and energy contours of a given utterance with a reference utterance. Finally, speech-to-text alignment information is used for duration scoring of a test utterance with a reference utterance. The proposed methods are evaluated on a database collected from native Turkish speakers. The articulation scoring method improves the correct recognition rate of confusable pairs of words by 5.6% yielding a correct recognition rate of 84.8%. The results also show that the prosody and duration scoring algorithms provide useful information to assess the match between the prosodic characteristics of two speakers.

1. Introduction

The improvements in speech processing technology have resulted in the development of computer-assisted methods for language education [1], [2], and speech therapy [3], [4]. In [3], the assessment of articulation problems in children is formulated while in [4] monitoring in speech therapy is addressed.

This study focuses on the development of automatic pronunciation scoring methods to:

• employ objective measures in the assessment of pronunciation problems and in performance monitoring,
• enable the design of software tools that provide visual feedback for the hearing-impaired,
• equip the standard user who is generally not a speech processing expert with the necessary speech processing techniques.

HMM based speech recognition [5] is an important tool for speech-to-text alignment which is required for acoustical analysis in pronunciation scoring. The proposed algorithm consists of three stages. In the first stage, an HMM based phoneme recognizer is employed for phoneme recognition and force-alignment of the input speech signal with the reference phonetic transcription using Viterbi decoding. The individual phoneme pronunciation errors are determined using log-likelihood scoring. In the second stage, prosodic characteristics of the input speech signal are compared to a reference recording which possesses correct pronunciation. First, the reference signal is force-aligned with the reference text. Syllables are determined and the time-aligned pitch and energy contours are extracted from the input and reference utterances. Normalized cross-correlation coefficient is used as a measure of prosodic similarity of the input utterance with the reference utterance. An overall prosodic match score is generated. In addition, the syllables which possess incorrect prosody are determined. In the final stage, the durational statistics of the input utterances are compared with those of reference utterances. The proposed methods are evaluated on a Turkish database collected from 29 speakers (15 male, 14 female).

Section 2 starts with a brief review on hearing-impairment and describes the most common articulatory and prosodic problems of hearing-impaired speakers. The pronunciation scoring algorithm is described in Section 3. The evaluations and results are given in Section 4. The paper is concluded with a discussion of the results in Section 5.

2. Hearing-Impairment

There are three types of hearing-impairment depending on the cause of the disorder. In conductive hearing-impairment, the cause is a disorder occurring in the outer ear or middle ear. Sensori-neural hearing-impairment results from a pathology in the inner ear or along the nerves connecting the inner ear with the brain stem. In psychological hearing-impairment, the problems can be attributed to emotional factors.

Articulation problems in the hearing-impaired can be classified as [6], [7]:

• Omission of sounds: “kitty” – “ki y”
• Substitution of sounds: “rabbit” – “wabbit”
• Addition of sounds: “sun” – “skun”
• Distortion of sounds

The most common articulation disorder is substitution in which a particular phoneme is replaced with another phoneme that is acoustically close to that phoneme. Substitution can be observed when particular phonemes occur at particular positions in the word. As an example, the patient may correctly articulate /p/ in the beginning of a word (i.e. “post”) but may face problems when the same phoneme occurs at the end of the word (i.e. “top”). In some cases, two phonemes can be correctly articulated individually. However, the patient may not be able to articulate the words that contain the two phonemes together. As an example, a patient who can correctly produce the phonemes /d/ and /t/ may not be able to say “drink”.

The articulation disorders in hearing-impairment are generally based on the problems in the production of consonants. The phonemes at the end of words are more problematic ([6], [8], [9], [10]). The problems in the production of consonants in hearing-impairment can be classified as follows:

• Substitution of voiced-unvoiced consonants (Example: /n/-/l/)
• Substitution of consonants that have identical place of articulation (Example: /b/ is substituted instead of /p/ which are both bilabials)
• Addition of nasality
• Omission of consonants in the beginning or at the end of the words

Prosody plays an important role in the intelligibility of speech. According to [8] and [9], the prosodic problems in hearing-impairment are:

• Insufficiency in f0 control
• Problems in breathing control when speaking
• Speaking in a slow manner: The rate of speech is generally lower in the hearing-impaired. Intelligibility is inversely proportional to the standard deviation of utterance duration.
• Incorrect use of stops: The number of stops is higher for a given sentence as compared to a normal speaker.
• Incorrect pitch: The pitch range is narrower and the pitch is higher or lower by 30–40 Hz on the average.
• Incorrect rhythm
• Incorrect stress
• Problems in voice quality

Although there exists significant correlation between speech intelligibility and the level of hearing impairment, it is possible to increase intelligibility in the patients with severe hearing loss with speech therapy.

In an extended study on the Turkish speaking hearing-impaired students described in [10], the author has shown that the most significant features that are correlated with speech intelligibility are utterance/stop durations and total number of stops. The author has found that pitch range and average pitch value have lower correlation with intelligibility. These features have medium level of correlation with the degree of hearing loss.

In this paper, we focus on substitution of consonants for articulation scoring. Log-likelihood scoring is employed for determining the consonant substitution errors. For prosody scoring, pitch and energy contours are compared both along the whole utterance and at the syllable level. Syllable level analysis enables accurate description of prosodic problems in a given utterance. Finally, analysis based on the utterance, word, phoneme, and stop durations provides additional information on pronunciation. The details of analysis and scoring methods are described in Section 3.

3. Pronunciation Scoring Algorithm

3.1. Articulation Scoring

In this study, we have focused on the substitution problems in the production of consonants. As described in Section 2, hearing-impairment may cause two types of substitution problems: Substitution of voiced-unvoiced consonants (i.e. /w/-/f/at) and substitution of consonants that have identical place of articulation (/p/-/b/at). Table 1 shows a classification of Turkish consonants according to their place of articulation in which hearing-impairment frequently causes substitution problems in phoneme pairs.

<table>
<thead>
<tr>
<th>Phoneme type (Place of articulation)</th>
<th>Phoneme pair</th>
<th>Example Turkish (English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labiodental</td>
<td>/f/-/v/</td>
<td>defa (times)-deva (cure)</td>
</tr>
<tr>
<td>Dental</td>
<td>/n/-/d/</td>
<td>katl (solid)-kadl (kadi)</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/s/-/z/</td>
<td>kas (muscle)-kaz (goose)</td>
</tr>
<tr>
<td>Alveo-palatal</td>
<td>/S/-/j/</td>
<td>be$\text{z}$ (five)-bej (beige)</td>
</tr>
<tr>
<td>/c/-/C/</td>
<td></td>
<td>$\text{C}$am (glass)-$\text{C}$am (pine)</td>
</tr>
<tr>
<td>Bilabial</td>
<td>/p/-/b/</td>
<td>put (idol)-but (thigh)</td>
</tr>
<tr>
<td>Velar</td>
<td>/k/-/g/</td>
<td>kar (snow)-gar (train station)</td>
</tr>
<tr>
<td>Liquid (/r/) and glide (/y/)</td>
<td>/r/-/y/</td>
<td>big (one)-bie$^{(*)}$</td>
</tr>
</tbody>
</table>

Table 1. Turkish phoneme pairs that frequently pose substitution problems in hearing-impairment. (*) This is a non-sense word.

In order to detect the substitution problems automatically with speech recognition techniques, special attention must be paid. This is due to the fact that substitution of a phoneme produces another word that is very similar to the original one in terms of
acoustics as the place of articulation of the phoneme pairs given in Table 1 are identical (except /r/ and /y/).

The flowchart of the articulation scoring algorithm for substitution of consonants is given in Figure 1. Scoring is based on the difference of log-likelihood scores between the phoneme recognition score and the forced-alignment score. Phoneme recognition is performed by using monophone HMMs trained with utterances from 205 native Turkish speakers. For forced-alignment, each recording is aligned with the corresponding phonetic transcription using the same monophone HMM set with the Viterbi algorithm. When the transcription matches the acoustical content of the speech signal, we expect to have the log-likelihood scores obtained in the phoneme recognition step to be close to the scores obtained in the forced-alignment step. The log-likelihood difference between the two steps is defined as:

$$\Delta L = L_p - L_f$$

where $L_p$ is the log-likelihood score obtained in phoneme recognition, $L_f$ is the log-likelihood score obtained in forced-alignment, and $\Delta L$ is the log-likelihood difference. We have used HTK for training the HMMs, phoneme recognition, and Viterbi decoding for forced-alignment. HTK was set to output log-likelihood scores in the phoneme level. Figure 2 shows the flowchart for computing the log-likelihood difference score.

![Flowchart of the articulation scoring algorithm](image)

Figure 2. Computation of the log-likelihood difference.

A step-by-step example for articulation scoring is as follows:

- Suppose that the patient utters “kas” (muscle) instead of “kaz” (goose) and we wish to determine the substitution error, i.e. /z/ is substituted for /s/.
- The distribution of the log-likelihood differences for /s/ and /z/ are pre-estimated in the training stage. We have used the mean and the variance of the log-likelihood differences for each phoneme.
- Two log-likelihood scores, $\Delta L_1$ and $\Delta L_2$, are computed with two alternative transcriptions (“kas” and “kaz”).
- The probability of the log-likelihood scores, $p_1$ and $p_2$, are computed. We assume normal distributions for the scores, therefore the mean and the variance values estimated from the training database for the corresponding phoneme is used for probability computation.
- Decision: Articulation problem in phoneme /s/ if $p_1 < p_2$, Correct articulation otherwise.

3.2. Prosody Scoring

The prosody scoring algorithm consists of two parts. In the first part, the pitch contour of a given utterance is compared with a reference utterance and the degree of match between the two pitch contours is estimated. In the second part, the energy contours are compared using a similar method.

In order to measure the similarity between two pitch contours, the normalized correlation coefficient was used. The following steps are employed for each pair of reference-subject utterance recording:

- Alignment and boundary detection: The subject’s speech signal is phonetically aligned at the state level with the reference utterance using an HMM based phonetic aligner. The syllable boundaries are determined automatically from the phonetic labels.
- Global scoring: The corresponding voiced segments of syllables of the two pitch contours are determined using the phonetic alignment information. For each segment pair, the shorter segment is linearly interpolated to match the length of the longer segment and all the segments within
an utterance are concatenated. We refer to these time-aligned, interpolated and concatenated segments as the time-aligned pitch contour profiles (TA-PCP). TA-PCP shows the corresponding f0 values in two utterances. Note that the TA-PCP vectors are of the same length for a pair of subject and reference utterance as a result of the interpolation step. An example is shown in Figure 3. We observe that the normalized correlation coefficient computed from the two TA-PCP’s, i.e. the two curves in Figure 3b is 0.845.

The energy contours of the subject and reference utterance are also compared using a similar method. For this purpose, the time-aligned energy contour profiles (TA-PEP) are computed and compared as follows:

- Global scoring: The corresponding syllables of the two energy contours are determined using the phonetic alignment information. For each pair of segments, the shorter segment is linearly interpolated to the length of the longer segment and the segments are concatenated. We refer to these time-aligned, interpolated and concatenated segments as the time-aligned energy contour profiles (TA-ECP). TA-ECP shows the corresponding energy values in two utterances. An example is shown in Figure 4.

- Segmental scoring: The segmental match between the subject and reference energy contours is determined by computing the normalized cross-correlation coefficient between corresponding

\[ r_e = \frac{\sum_{i=1}^{N} (e_1(i) - \mu_1)(e_2(i) - \mu_2)}{\sqrt{\sum_{i=1}^{N} (e_1(i) - \mu_1)^2} \sum_{i=1}^{N} (e_2(i) - \mu_2)^2} \]  

where \( e_1 \) is the TA-PEP for the subject’s utterance, \( e_2 \) is the TA-PEP for the reference utterance, \( \mu_1 \) is the mean of \( e_1 \), and \( \mu_2 \) is the mean of \( e_2 \).

- Segmental scoring: The segmental match between the subject and reference energy contours is determined by computing the normalized cross-correlation coefficient between corresponding

\[ \mu_1 = \frac{1}{N} \sum_{i=1}^{N} e_1(i) \quad \mu_2 = \frac{1}{N} \sum_{i=1}^{N} e_2(i) \]
subject and reference energy contour segments. Figure 4c shows an example for segmental scoring.

Figure 4. (a) Reference and subject energy contours to be compared corresponding to the utterance in Turkish “Kaza nedeniyle ulaşım azalı” from two different speakers. (b) Time-aligned energy contour profiles (TA-PEP) for the reference and subject energy contours showing the corresponding energy values in the two contours. (c) Segmental scores (i.e. the normalized correlation coefficient between corresponding energy contour segments) that show the match between the subject and reference energy contours segmentally.

3.3. Duration Scoring

As discussed in Section 2, the utterance duration is generally longer in the hearing-impaired with more stops. Therefore, it is important to compare durational characteristics in pronunciation scoring. For this purpose, we have used the phonetic labels obtained by forced-alignment with monophone HMMs. For each pair of subject and reference utterances, the sentence, word, phoneme, and silence durations are determined. The durations are normalized with the total utterance duration and the following scores are obtained:

- \( D_1 \): Ratio of subject and reference utterance durations excluding the silence periods in the beginning and at the end of the utterances
- \( D_2 \): Ratio of average word durations of the subject and reference utterances
- \( D_3 \): Ratio of average phoneme durations of the subject and reference utterances
- \( D_4 \): Ratio of average inter-word silence (stop) durations of the subject and reference utterances

Figure 5 illustrates an example of duration scoring. We observe that \( D_1 = 2.060 \), \( D_2 = 1.688 \), \( D_3 = 1.544 \), and \( D_4 = 33.333 \). Therefore, the subject utterance has longer average utterance duration \((D_1 > 1.0)\), longer word duration \((D_2 > 1.0)\), longer phoneme duration \((D_3 > 1.0)\), and more stops \((D_4 > 1.0)\) as compared to the reference utterance.

Figure 5. Example for duration scoring. \( D_1 = 2.060 \), \( D_2 = 1.688 \), \( D_3 = 1.544 \), and \( D_4 = 33.333 \) for a subject/reference utterance pair. Reference utterance spectrogram and labels (top), subject utterance spectrogram and labels (bottom).

4. Evaluations

4.1. Articulation Scoring

As our focus is mainly on substitution problems in this study, we prepared a list of confusable word pairs given in Table 3. We compared the performance of the proposed articulation scoring method in detecting the substitution errors with standard HMM based isolated word recognition.

The confusable word pairs given in Table 3 were collected from nine speakers (five male, four female). Each word was recorded five times in random order. An HMM with one state per phoneme is trained for each word. The distribution of the acoustical feature vectors was modelled by a Gaussian mixture with two components in each state. We have used MFCCs with energy, probability of voicing, and delta features as the acoustical feature vector.

In the first part of the test, isolated word recognition is employed between pairs of confusable pairs of words. The recognition performance is measured using K-fold cross validation. For each cross-validation step, the recordings of one speaker were reserved as the validation set and the HMMs were trained using the rest of the database. The recognition performance is measured as the average of the correct recognition rates of the nine cross-validation steps. The resulting correct recognition rate was 80.3% which is the average
recognition rate given a pair of confusable words in Table 3.

Table 3. List of confusable word pairs generated by substitution of the consonants given in Table 1.

<table>
<thead>
<tr>
<th>Phoneme Pairs</th>
<th>/ç/-/ç/</th>
<th>/p/-/b/</th>
<th>/l/-/d/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>çeket-çeket</td>
<td>pasta-basta</td>
<td>tabak-tabak</td>
</tr>
<tr>
<td>Middle</td>
<td>acı-acı</td>
<td>kapı-kabı</td>
<td>kât-kadi</td>
</tr>
<tr>
<td>End</td>
<td>avuç-avuç</td>
<td>dolap-dolab</td>
<td>yakut-yakud</td>
</tr>
</tbody>
</table>

Table 4. Results for articulation scoring.

4.2. Prosody and Duration Scoring

The prosody and duration scoring algorithms are evaluated using two databases. The first database consists of a subject/reference male speaker pairs. There are 15 sentence utterances. The subject speaker tried to mimic the reference prosody by listening to the reference recordings. In the second database, the subject and reference speakers read a paragraph from a novel. No effort was made for mimicking the prosody. Therefore, the subject and reference prosody was significantly different for this database. Table 5 shows the results of the prosody and duration scoring algorithms on these two databases. We observe that when the subject speaker mimics the reference speaker’s prosody, the prosody scores, r_p and r_e, and the duration scores are closer to unity.

Table 5. Results of the prosody and duration scoring algorithms on the two databases. Mean and standard deviation of each score is given. Note that the subject speaker tried to mimic the reference speaker’s prosody in Database 1.

5. Conclusion

In this study, three methods were proposed for articulation and prosody scoring. We have focused on the substitution of consonants which is a common articulation problem in the hearing-impaired. The articulation scoring algorithm employed log-likelihood scoring based on phoneme recognition and forced-alignment. A new method referred to as time-aligned pitch and energy contour profiles is developed which compares the pitch and energy contours of two utterances and estimates prosodic match scores. The proposed methods are evaluated on a database collected from native Turkish speakers. The articulation scoring method improves the correct recognition rate of confusable pairs of words by 5.6% yielding a correct recognition rate of 88.4%.

At the time of writing, a database collected from the hearing-impaired and annotated by speech therapists was not available. As future work, the methods will be tested on the recordings collected from the hearing-impaired. A software tool is under development which will provide feedback and enable the design of objective tests for the hearing-impaired.

6. References


