Language-universal speech audiometry with automated scoring

Bart Vaerenberg1,2, Louis ten Bosch3, Wojtek Kowalczyk4, Martine Coene1,3, Herwig De Smet6, Paul J Govaerts1

1 The Eargroup, Antwerp-Deurne, Belgium
2 Laboratory of Biomedical Physics, University of Antwerp, Belgium
3 CLS/CLST, Radboud University Nijmegen, The Netherlands
4 LIACS, Leiden University, The Netherlands
5 Vrije Universiteit Amsterdam, The Netherlands
6 Karel de Grote Hogeschool, Antwerp, Belgium

vaerenberg@eargroup.net, l.tenbosch@let.ru.nl, wojtek@liacs.nl, m.m.r.coene@vu.nl, herwig.desmet@kdg.be, dr.govaerts@eargroup.net

Abstract

The clinical assessment of speech discrimination by professional audiologists is resource intensive. Yet discrepancies in language or dialect between the test subject and the audiologist may cause a significant bias in the test result. To address these issues, a speech audiometric test (SAT) has been designed to be language/dialect independent and to allow automated scoring by means of an MFCC-based Dynamic Time Warping alignment measure. A Pearson correlation of 0.83 was found between the automatic scores and human phoneme scoring. Normative data were obtained and compared to conventional SATs which revealed differences in speech reception thresholds within 2 dB.

Index Terms: audiometry, speech perception, automatic assessment, cochlear implants

1. Introduction

Since the introduction of cochlear implants (CI) and similar therapies, assessment of speech perception has become more and more important in the clinical practice [1]. Speech audiometric results are interesting because they relate closely to the patient's hearing performance in daily life. As such speech audiometry is routinely performed for the selection of CI candidates, for the evaluation of outcome in CI recipients [2] and even to steer the programming (i.e., patient-dependent optimization) of CI speech processors [3]. From a patient's initial intake onwards and throughout the long term follow-up, speech audiometry is performed repeatedly, resulting in a substantial load on the clinic in terms of time and resources [4].

In conventional speech audiometry tests (SAT), words or sentences are presented acoustically to the subject at predefined intensities. The subject is instructed to repeat what he/she has heard and a trained professional (audiologist) scores the subject's responses. The scoring of responses depends on the stimulus material used and can either be a phoneme score (e.g., in CVC monosyllable tests) in which errors on the phoneme level are recorded, a word score in which each utterance is judged as entirely correct or incorrect, or a sentence score in which the subject's correct repetitions of a number of keywords in a sentence is counted. To minimize test-retest variability it is important that enough stimuli are presented in a single test run. Typically, 20 to 50 items are used per presentation level [5].

In addition, to obtain reliable and consistent results, it is essential that these speech perception tests are well designed in terms of stimulus quality, difficulty across lists, and output level calibration [6]. However, in many languages such standardized speech stimuli do not exist or there is no normative data available for them. Another drawback is that stimuli need to be representative for the subject's language in terms of vocabulary and phonemic content and that these linguistic variables show great variation across languages and dialects. As such, the scoring of responses may result in a significant error if the audiologist and the patient do not have the same native language or dialect [7].

In this study, we address these issues by developing a new type of test for speech understanding in quiet and in noise. The fundamentals of this test consist of the construction of a personal, yet language representative speech test for each individual subject based on his/her own lexicon and an automated, language and dialect independent scoring of responses, allowing subjects to perform the test on their own. During an initial session, words from a subject's daily readings are presented visually for him/her to pronounce. The subject's utterances are recorded and subsets of these words are presented acoustically, for the subject to repeat, during later test sessions. At that time, a scoring mechanism compares the original, visually prompted utterance to the repeated, acoustically prompted one. The former is assumed to be a phoneme score (e.g., in CVC monosyllable tests) in which errors on the phoneme level are recorded, a word score in which each utterance is judged as entirely correct or incorrect, or a sentence score in which the subject's correct repetitions of a number of keywords in a sentence is counted. To minimize
fourth section handles the normative data and in section 5 our results are presented, which are discussed further in section 6.

2. Assessment of pronunciation differences

2.1. General approach

There are different methods to assess the difference between two pronunciations. One way is to use methods based on Automatic Speech Recognition (ASR) [8]. Using ASR the difference between utterances can be measured in terms of different results from e.g., a free phone loop decoding, or by specific techniques used to measure differences between pronunciations from a learner and a teacher in Computer Assisted Language Learning [9]. The necessity of trained acoustic models, however, is a drawback for the application of the ASR method in a clinical setting, in particular in cases of under-resourced languages. It was therefore opted to use another way to assess differences between utterances: by using a language ignorant Dynamic Time Warping (DTW) approach [10]. Unlike ASR, DTW does not require substantial speech data for training; instead a few parameters must be chosen. The disadvantage of DTW is that it accumulates all acoustic deviations between the two utterances along the found best alignment path, irrespective whether these differences would make sense phonemically according to a listener. However, the use of DTW before ASR approaches came into fashion shows that the DTW alignment score can be a useful measure for distinguishing two pronunciations. This idea is further exploited in the descriptions below.

2.2. Application

This particular application required a system for comparing 2 different pronunciations of a short word using a language-ignorant approach that does not involve individual training. Mel-Frequency Cepstral Coefficients (MFCC [11]) have been used for extracting perceptually relevant features from the short-term speech spectrum. Combined with DTW, the resulting MFCC DTW distance measure meets the defined requirements and was chosen to be used in this study.

As a first step, MFCC frames are calculated using a frame analysis window width of 25 ms and frame shift of 10 ms. The MFCC frames are augmented by their first and second order temporal derivatives (delta, delta-delta), and an utterance-based Cepstral Mean and Variance Normalization (CMVN) is applied to minimize the between-speaker differences and thereby to optimize the generalization of the speaker-independent DTW settings. The DTW then operates upon pairs of sequences consisting of these augmented MFCC vectors and the Euclidean distance is used to compute local scores. No additional costs are attributed to frame insertion and deletions. The total score of the best alignment path is normalized by dividing this score by the number of traversal steps making up that path and serves as the eventual DTW alignment score (henceforth ‘DTW distance’).

The applicability of the DTW distance for this particular application was validated by considering 4 different sub databases. These databases contain pairs of words pronounced by the same speaker: 1) 300 pairs of a same word (SAME, e.g., cat-cat); 2) 25 pairs of minimally different words (MINIMAL, e.g. lon-lom); 3) all pairs of typically different words (TYPICAL, e.g., cat-goes), chosen from a set of 300 short words and 4) 25 pairs of maximally different words (MAXIMAL, e.g., ‘put’-‘sil’). Each of these 4 data sets was collected from 47 speakers.

3. Mapping DTW to human scores

3.1. General model

For the application to be usable in a clinical setting MFCC DTW distances had to be transformed to a psychometric scale ranging from 0 to 100, resembling the shape and range of the human scoring in a conventional SAT. This mapping is non-linear and it has been derived from experimental data.

3.2. Model tuning data

To record the tuning data used to appropriately map DTW scores, a software application was developed to record, present and score speech utterances. Subjects were recruited from both normal hearing and hearing impaired populations of 4 different languages/regions (Dutch, Flemish, German and Portuguese, as listed in Table 1). For each subject a speech intelligibility rating (SIR) [12] was assessed and only subjects with SIR1 (completely intelligible in conversation) were included in the experiment. During an initial session (SES1), subjects were asked to select a text source from their daily readings, for example a book or an online newspaper. From that text a set of 300 short words (3 to 5 characters) was extracted by the system. These words were presented visually on a computer monitor and pronounced by the subject. In addition, each subject pronounced 300 words extracted from a conventional SAT test for his/her native language (NVA [13] [14] for Dutch/Flemish, Freiburger Einsilber [15] for German and Crianças Dissilabica [16] for Portuguese, as listed in Table 1). All utterances were recorded at a 16 kHz sampling rate in a quiet office room environment.

<table>
<thead>
<tr>
<th>Language</th>
<th>Subjects</th>
<th>Conventional SAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch (Belgium)</td>
<td>206</td>
<td>NVA Flemish</td>
</tr>
<tr>
<td>Dutch (Netherlands)</td>
<td>43</td>
<td>NVA Dutch</td>
</tr>
<tr>
<td>German</td>
<td>101</td>
<td>Freiburger Einsilber</td>
</tr>
<tr>
<td>Portuguese</td>
<td>21</td>
<td>Crianças Dissilabica</td>
</tr>
<tr>
<td>Total</td>
<td>371</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Number of subjects and conventional SATs used for each language/region.

After a minimum interval of 2 weeks subjects returned for a second session (SES2), during which subsets of the recordings of both the conventional SAT words and the subject’s own 300 short words were presented acoustically, in 2 different listening conditions. A speech shaped noise source was adjusted in intensity to create listening conditions that were expected to result in both easy (> 50% correct) and difficult (< 50% correct) speech understanding for each particular subject. At each of the 2 listening conditions the subject was asked to repeat the words that were presented and an audiologist performed a phoneme score on the responses, resulting in 4 human scores (own words in difficult conditions, own words in easy conditions, conventional SAT words in difficult conditions and conventional SAT words in easy conditions). Each of those scores was the result of 24 presentations (20 presentations for German). At the same time, the system also calculated the DTW distance for each pair of
recordings, resulting in the respective 4 machine scores, defined as the average DTW distance across that set of utterance pairs. Correlation coefficients were calculated between the human and the DTW session scores for all languages separately and for all languages together.

4. Normative data

Conventional SATs often come with normative data (normal curves). These data consist of the average speech discrimination scores of normal hearing listeners at different intensities. To obtain comparable normative data for the DTW based SAT, a new data set (distinct from the tuning data set) has been obtained as follows.

Ten normal hearing Dutch speaking listeners were asked to pronounce and record 2 sets of words. The first set (OWN) contained 300 short words extracted from their daily readings, the second set (SAT) contained 300 words extracted from a conventional SAT's word lists (Brugse Monosyllable CVC Speech Lists [17]). The recordings were obtained after visual presentation, following the same procedure for the initial session (SES1) as described above. After a minimum of 3 days, speech perception was assessed in these subjects, using both sets of stimuli. The OWN words were scored by DTW; the SAT words were scored both by DTW and by a professional audiologist (HUMAN). The stimuli used and the scoring performed are summarized in Table 2.

<table>
<thead>
<tr>
<th>Test</th>
<th>Words</th>
<th>Scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>OWN</td>
<td>Daily Readings</td>
<td>DTW</td>
</tr>
<tr>
<td>SAT</td>
<td>Brugse CVC</td>
<td>DTW &amp; HUMAN</td>
</tr>
</tbody>
</table>

Table 2: Speech perception tests performed in 10 normal hearing subjects.

Stimuli were presented monaurally under headphones (TDH39) in a clinical sound treated room. The desired output level was obtained for each individual stimulus by endpointing and RMS-equalizing the signal. The initial presentation level of 40 dB SPL was increased by 5 dB until a score of 90% or higher was obtained and then decreased from 40 dB SPL in steps of 5 dB until a 0% score was obtained.

5. Results

5.1. DTW distance measure

In our DTW calculation, the input feature vectors consist of the 12 MFCC coefficients, which are augmented by their first and second order temporal derivatives of the MFCC vectors. In combination with the log(E) coefficient, this amounts to $3 \times (12 + 1) = 39$ parameters. In Figure 1, the cumulative distributions of DTW scores are presented for the four test sets (SAME, MINIMAL, TYPICAL and MAXIMAL).

The performance of the DTW distance was measured in terms of equal error rate (EER) of each class when compared to the SAME class, as shown in Table 3.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAME - TYP</td>
<td>5.7</td>
</tr>
<tr>
<td>SAME - MIN</td>
<td>40.2</td>
</tr>
<tr>
<td>SAME - MAX</td>
<td>2.5</td>
</tr>
</tbody>
</table>

Table 3: The ability of the DTW distance to separate classes in terms of Equal Error Rate (in %) between the SAME class and the other 3 classes.

5.2. Correlation between DTW and human scoring

The average DTW distance, of a set of 24 (20 for German) presented words and their average phoneme score as judged by a professional audiologist are depicted in Figure 2. The scoring for Flemish (the Dutch dialect spoken in Belgium) was performed by 2 different audiologists, each of which produced half of the data set. Other languages/regions were scored by a single audiologist each. The correlations found between human scoring and DTW distance are listed in Table 4. When languages were pooled, an overall correlation of -0.83 was found.

<table>
<thead>
<tr>
<th>Language</th>
<th>Subjects</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch (Belgium)</td>
<td>206</td>
<td>-0.84</td>
</tr>
<tr>
<td>Dutch (Netherlands)</td>
<td>43</td>
<td>-0.79</td>
</tr>
</tbody>
</table>
Table 4. Correlations between human phoneme scoring and MFCC DTW distance.

<table>
<thead>
<tr>
<th>Language</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>101</td>
</tr>
<tr>
<td>Portuguese</td>
<td>21</td>
</tr>
<tr>
<td>All</td>
<td>371</td>
</tr>
</tbody>
</table>

Table 4. Correlations between human phoneme scoring and MFCC DTW distance.

5.3. Mapping DTW distances to human scores

The scatter plots presented in Figure 2 show that the MFCC DTW distance is negatively correlated with the human score. Moreover, both measures take values in different ranges and the relation between them is non-linear. Therefore our transformation of the DTW distance into the human score involves two steps. First, we linearly transformed the MFCC DTW distance $d$ into the DTW score $x$ which is positively correlated with the human score and takes values in the range $[0, 100]$. Next, we assumed that a functional relation between the DTW score $x$ and the human score $y$ can be modeled with help of a generalized logistic function:

$$y(x) = A + \frac{K-A}{(1+Qe^{-B(x-M)})^1/v}$$

where the lower asymptote ($A$) is set to 0, the upper asymptote ($K$) is set to 100, $v$ is set to 1. Additionally, to compensate for a bias in the data (over- and under-representation of high and low scores, resp.) we added to the data set 1000 virtual points with DTW score 25 and human score 0 and 100 points with DTW score 75 and human score 100. Then values of the remaining parameters ($B$, $Q$ and $M$) were determined using the Nelder-Mead simplex algorithm [18], to minimize the Root Mean Square Error (RMSE) over the data set specified. The optimal parameters for the model given by formula (1) are: $B = 0.1034$; $Q = 1.2721$; $M = 44.4934$. Figure 3 shows the optimal fit using this model plotted on top of the individual data points.

5.4. Normative data

Normative data obtained in 10 hearing listeners are depicted in Figure 4 as the median scores at the presented intensities. The median 50% speech reception threshold (SRT) for both the OWN words and the SAT words when scored by DTW was shown to be 21 dB SPL. When scored by a human the median SRT for the SAT words showed to be 19.5 dB SPL. The published normative data for the Brugse SAT specifies a 50% SRT of 20 dB SPL for monaural normal hearing listeners. It is remarkable to see that, even at higher stimulus levels, median DTW scores do not reach more than 85%.

6. Discussion

The strong correlations found between human and DTW scores indicate that an automated scoring mechanism may be suitable to assess speech perception deficits. The correlation in the Dutch (Netherlands) is slightly lower, and the correlation in the Portuguese is markedly lower, than in the other languages/regions. The authors have no explanation for this observation. It may be attributed to the audiologist's day to day variance in phoneme error judgment. Another reason could be the variance in background noise when the initial recordings have been obtained in a different location than where the actual test session took place.

When comparing the median 50% SRTs in normal hearing listeners, it is clear that the use of short words extracted from the subject’s daily reading is equivalent to using conventional SAT words, like the Brugse CVC words. Not only the SRTs resulting from the use of both types of stimuli are the same, but also below and above the 50% score point, results from both tests are very similar.

A clinically irrelevant 1.5 dB upward shift is observed in the median SRT when human scoring is replaced by DTW scoring. This makes us believe that the results obtained by DTW scoring may be comparable to conventional speech audiometric test results and therefore clinically usable. The ceiling effect observed around 80% when using DTW scoring, may be attributed to the fact that in obtaining these normative data, the recordings took place in an office room, while the actual speech perception was assessed in a clinical sound treated room. The difference in ambient noise between the two rooms may have introduced a floor effect in the DTW distance, which presents itself as a ceiling effect in the speech perception scores, however further investigation is needed to confirm this hypothesis.

7. Acknowledgements

The reported study was conducted within a Seventh Framework Programme Project FP7 OPTI-FOX 262266, supported by the European Union.

B. Vaerenberg received a PhD grant for this work from the IWT (Agency for Innovation by Science and Technology, Baekeland-mandaat IWT090287).
8. References


