Voice Quality Assessment by Means of Comparative Judgments of Speech Tokens

A. Kacha (*), F. Grenez (*), J. Schoentgen (**)

(*) Department Signals and Waves, Faculty of Applied Sciences, Université Libre de Bruxelles, Brussels, Belgium
(**) National Fund for Scientific Research, Belgium
akacha@ulb.ac.be, fgrenez@ulb.ac.be, jschoent@ulb.ac.be

Abstract
The presentation concerns a perceptual assessment of disordered speech, which rests on comparative judgments. Pairs of stimuli are presented randomly to listeners who are asked to label the most degraded sample. At the end of a listening session, speech signals are assigned overall scores on the base of individual comparative judgments. To contrast the perceptual assessment of voice disorders via comparative judgments with a conventional method, two groups of listeners have taken part in the experiment. A first group has been comprised of six naive listeners, i.e., listeners without any training in speech therapy or laryngology. A second group of listeners has been comprised of three speech therapists. They rated from 0 to 3 the grade G (of the GRBAS scale) of each sound sample by means of an ordinal scale. Results show that comparative judgments give rise to high intra-judge and inter-judge agreements for sustained vowel [a] as well as for sentences even though the listeners had no prior experience in rating disordered voices.

1. Introduction
Assessment of disordered voices is routinely based on listener perception of speech. For example, clinicians rate the degree of perceived overall abnormality, called grade, to monitor the voice of patients. Indeed, the so-called GRBAS scale foresees rating voice quality according to five factors, which are the grade, roughness, breathiness, asthenia and strain. Typically, voices are assigned a degree of perceived abnormality from 0 to 3 along one of these five scales. A drawback of perceptual ratings is intra and inter-judge variability [1, 2]. Experiments have shown that to obtain reproducible evaluations, listeners must have substantial experience in voice timbre rating.

In this presentation, we propose a perceptual rating of hoarseness that is based on comparative judgments of pairs of speech samples. The question that we would like to address is whether it is easier for a listener to compare a list of voice samples two by two, rather than rate each voice sample on an ordinal or visual analog scale.

In the framework of a listening session, pairs of stimuli are presented randomly to a listener who is asked to designate the most abnormal sample of the pair. The experience is repeated until all sample pairs have been presented. At the end of the session, samples are assigned scores on the base of all possible pair comparisons. Indeed, when a sample is judged to be the most abnormal at each comparison, it cumulates the highest score at the end of a listening session that comprises all possible pairs. The least abnormal voice, on the contrary, is assigned a small score after the session. Since the final score can range between 0 and \(N - 1\), where \(N\) is the number of speech samples to be compared, comparative judgment scoring results, in general, in a scale that is subdivided in a number of levels that is larger than the usual four to seven levels typical of ordinal scales.

Since what is requested is the comparison of two speech samples, the task is expected for naive speakers to be easier than the rating on a multi-level scale. The latter requests listeners to be experienced in the rating of voice timbre.

In this study, scores obtained on the base of multiple comparisons are confronted to scores obtained by rating voices on an ordinal scale between 0 and 3. The former are assigned by a group of six naive listeners and the latter by a group of three speech therapists. Each naive listener has devoted separate listening sessions to vowel [a] and four sentences produced by 22 speakers. Results show that intra-naive-listener and inter-naive-listener scores are highly correlated for all types of stimuli.

The remainder of the presentation is organized as follows. Section 2 presents multi-comparison scoring as well as the data used in the experiment. Section 3 gives the assessment results obtained by means of comparative judgments of speech stimuli and compares them to those obtained by speech therapists performing conventional ratings. The results are discussed in section 4.
2. Methods

2.1. Corpora

Speech data comprise sustained vowels [a], including onsets and offsets, and four French sentences produced by 22 healthy or dysphonic speakers (10 male and 12 female speakers). The corpus comprises 20 adults (from 20 to 79 years), one boy aged 14 and one girl aged 10. Five speakers are normophonic, the other are dysphonic.

The sentences are the following: “Le garde a endigué l’abbé”, “Bob m’avait guidé vers les digues”, “Une poule a picoré ton cake” and “Ta tante a appâté une carpe”. They are referred to as S1, S2, S3 and S4, respectively. The sentences have the same grammatical structure, the same number of syllables and roughly the same number of resonants and plosives. Sentences S1 and S2 are voiced by default, whereas S3 and S4 include voiced and unvoiced segments.

Speech signals have been recorded at a sampling frequency of 48 kHz. The recordings were made in an isolated booth by means of a digital audio tape recorder (Sony TCD D8) and a head-mounted microphone (AKG C41WL) at the laryngology department of a university hospital in Brussels, Belgium. The recordings have been transferred from the DAT recorder to computer hard disk via a digital-to-digital interface. Silent intervals before and after each recording were removed.

2.2. Voice quality assessment by comparative judgments

The experiment explores the ability of naive listeners to compare two stimuli in terms of grade, i.e., perceived overall degree of deviance of the voice. The aim is to hierarchize a set of recordings from the least to the most anomalous via comparative judgments of all possible sample pairs within the set. During a listening session, sample pairs are presented to the listener, who is asked to label the sample with the highest perceived abnormality. The listener has also the possibility to label both samples of a pair as equally (ab)normal.

1. The list of all possible different pairs of speech samples is formed on the base of a set of recordings. When \( N \) denotes the number of recordings, the total number of sample pairs is \( N(N-1)/2 \).

2. All scores are initialized to zero.

3. A randomly selected pair of speech samples is presented to the listener, who is asked to label the sample with the highest perceived abnormality. The listener has also the possibility to label both samples of a pair as equally (ab)normal.

4. The total score of the sample labeled as the most deviant is increased by one. If both samples of the pair are judged to be equal, the total score of both samples is increased by 0.5.

5. Steps 3 and 4 are repeated until all possible pairs that belong to a same session have been presented.

The hierarchization (stimuli presentation and score computation) is handled by a computer program. The listener communicates his choice by activating buttons.

Sound samples have been presented via a digital-to-analog audio interface (Digidesign Mbox) and dynamic stereo headphones (Sony MDR-7506). Loudness has been fixed at a comfortable level by the listener. Listening sessions have been held in a quiet room. Each session has taken about one hour. At half-time, listeners have taken a rest of about five minutes.

2.3. Listeners

To confront the perceptual assessment of voice disorders via comparative judgments with a conventional method, two groups of listeners have taken part in the experiment. A first group has been comprised of six naive listeners (one female, five males), i.e. listeners without any training in speech therapy or laryngology. All reported normal hearing. Their ages ranged from 24 to 57. They were asked to assess voice samples by means of comparative judgments of the overall degree of deviance of the voice. One listening session was devoted to a set of 22 stimuli. The total number of “naive” sessions has therefore been equal to 6 x 5 = 30. The same experiment has been repeated by three listeners after a period of a day at least to gauge intra-judge reliability. The total number of “naive” retest sessions has therefore been equal to 3 x 5 = 15.

A second group of listeners has been comprised of three speech therapists. They rated from 0 to 3 the grade G (of the GRBAS scale) of each sound sample by means of an ordinal scale. Each clinician rated vowel [a] as well as sentences S1 and S3 spoken by the 22 speakers. The total number of “clinical” sessions has therefore been equal to 3 x 3 = 9.

3. Results

The aim is to summarize the raw scores and to analyze the agreement within and between judges. Also, voice quality scores obtained via the six naive listeners are compared to the grade scores assigned by three clinicians.

The quartiles of the scores obtained via comparative judgments by naive listeners are given in Table 1.
3.1. Inter-judge agreement of naive listeners

Concordance between naive judges has been expressed by means of Pearson’s product moment correlation ($\rho_p$) between listener scores. The scores of the four sentences were arranged for a given judge into a single series by stacking the sentence scores. Correlation coefficients have been calculated by means of the score series of each listener. The values for sustained vowel [a] and sentences S1 to S4 are given below and above the diagonal of Table 2 respectively. While testing for the statistical significance of the Pearson product moment correlations, the method of Bonferroni has been used to account for multiple comparisons [3]. The number of independent measures involved in the test has been set equal to 22. The null hypothesis ($\rho_p = 0$) has been rejected for all table entries (one-tailed test, $\rho_{\text{crit}} = 0.56$, $p < 0.05$).

<table>
<thead>
<tr>
<th></th>
<th>J1</th>
<th>J2</th>
<th>J3</th>
<th>J4</th>
<th>J5</th>
<th>J6</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a]</td>
<td>0.96</td>
<td>0.98</td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1 to S4</td>
<td>0.92</td>
<td>0.95</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Pearson’s product moment correlation values between scores obtained via comparative judgments during test and retest sessions of three naive listeners.

3.2. Intra-judge agreement of naive listeners

Intra-judge agreement has been examined by calculating Pearson’s product moment correlation between the scores obtained by means of the test and retest sessions of three naive judges. For each judge, the scores of the four sentences have been stacked into a single series. The results are listed in Table 3. To account for multiple comparisons, Bonferroni’s method has been used. The number of independent realizations involved in the test has been set equal to 22. The null hypothesis ($\rho_p = 0$) has been rejected for all table entries (one-tailed test, $\rho_{\text{crit}} = 0.46$, $p < 0.05$).

<table>
<thead>
<tr>
<th></th>
<th>J1</th>
<th>J2</th>
<th>J3</th>
</tr>
</thead>
<tbody>
<tr>
<td>[a]</td>
<td>0.96</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td>S1 to S4</td>
<td>0.90</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Table 2: Pearson’s product moment correlation values between scores obtained via comparative judgments by six naive listeners for sustained vowel [a] (below the diagonal) and sentences S1 to S4 (above the diagonal).

Figure 1 shows, as an example, the scattergram of the scores obtained by judges J1 and J2 for sentence S2.

Figure 1: Scores obtained by judges J1 and J2 for sentence S2 ($\rho_p = 0.94$).
3.3. Comparison of scores obtained via comparative judgments by naive listeners and grade scores assigned by speech therapists

Scores obtained by means of comparative judgments have been compared to grade scores assigned by speech therapists. Vowel [a] and sentences S1 and S3 have been rated clinically. Sentence scores have been stacked to form a single series. Pearson’s product moment correlation values between therapist and naive listener scores are given in Table 4 and 5 for sustained vowels [a] as well as sentences S1 and S3 respectively. Bonferroni’s method has been applied to account for multiple comparisons (one-tailed test, \( r_{\text{crit}} = 0.57, \ p < 0.05 \)). The number of independent realizations involved in the statistical test has been set equal to 22.

In Table 4, one sees that for vowel [a], therapist ratings are statistically significantly correlated with scores built up from comparative judgments by naive listeners. For sentences S1 and S3, only the ratings of therapist C2 are statistically significantly correlated with the scores obtained by comparative judgments (Table 5).

<table>
<thead>
<tr>
<th>J1</th>
<th>J2</th>
<th>J3</th>
<th>J4</th>
<th>J5</th>
<th>J6</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.78</td>
<td>0.61</td>
<td>0.79</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td>C2</td>
<td>0.81</td>
<td>0.62</td>
<td>0.74</td>
<td>0.81</td>
<td>0.83</td>
</tr>
<tr>
<td>C3</td>
<td>0.84</td>
<td>0.82</td>
<td>0.79</td>
<td>0.79</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Table 4: Pearson’s product moment correlation values between therapists’ ratings and scores obtained by comparative judgments of vowel [a].

<table>
<thead>
<tr>
<th>J1</th>
<th>J2</th>
<th>J3</th>
<th>J4</th>
<th>J5</th>
<th>J6</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.61</td>
<td>0.53</td>
<td>0.55</td>
<td>0.49</td>
<td>0.44</td>
</tr>
<tr>
<td>C2</td>
<td>0.81</td>
<td>0.75</td>
<td>0.65</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>C3</td>
<td>0.58</td>
<td>0.58</td>
<td>0.43</td>
<td>0.60</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 5: Pearson’s product moment correlation values between therapists’ ratings and scores obtained by comparative judgments of sentences S1 and S3.

Inter-therapist agreement has also been examined. Results for sustained vowel [a] and sentences S1 and S3 are listed in Table 6. Ratings for sentences S1 and S3 have been stacked to form a single series. Bonferroni’s method has been used to account for multiple comparisons (one-tailed test, \( r_{\text{crit}} = 0.46, \ p < 0.05 \)). The number of independent realizations involved in the statistical test has been set equal to 22. Correlations between grades assigned to sentences S1 and S3 are displayed above the diagonal and correlation between grades assigned to vowel [a] are displayed below the diagonal. It is seen that the correlation between sentence grades assigned by therapist C1 and therapist C3 is not statistically significant.

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.57</td>
<td>0.69</td>
<td>0.41</td>
</tr>
<tr>
<td>C2</td>
<td>0.76</td>
<td>—</td>
<td>0.69</td>
</tr>
<tr>
<td>C3</td>
<td>0.69</td>
<td>0.67</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 6: Pearson’s product moment correlation between grades assigned by three speech therapists to sustained vowel [a] (below the diagonal) and sentences S1 and S3 (above the diagonal).

4. Discussion

Correlation analyses show that inter-naive-listener and intra-naive-listener agreement is high for both sustained vowel [a] and sentences S1 to S4. Correlations between naive listener scores and therapist’ ratings are statistically significant for vowel [a]. Correlations were high (> 0.74), except for the moderate correlation between naive listener J2 and therapists C1 and C2. Results are different for the sentences, however. Only therapist C2’s ratings correlate statistically significantly in their totality with the scores obtained via comparative judgments. This is expected because the inter-clinician agreement has been poor for sentences S1 and S3 (Table 6). As a consequence, therapist C1 and C3’s ratings must correlate poorly with the scores obtained via comparative judgments, which have been shown to cohere strongly.

5. Conclusion

In this presentation, we have explored a method for disordered voice assessment based on comparative judgments of speech stimuli. Results show that comparative judgments give rise to high intra-judge and inter-judge agreements for sustained vowel [a] as well as for sentences, even though the listeners had no training in speech therapy or laryngology.

6. References