Do Humans and speaker verification system use the same information to differentiate voices?

Juliette Kahn 1,2, Solange Rossato2

1Laboratoire Informatique d'Avignon, Université d'Avignon, Avignon, France
2 Laboratoire Informatique de Grenoble, Université Stendhal, Grenoble, France
juliette.kahn@univ-avignon.fr, solange.rossato@imag.fr

Abstract
The aim of this paper is to analyze the pairwise comparisons of voices by a speaker verification system (ALIZE/Spk) and by human. A database of familial groups of 24 speakers was created. A single sentence was chosen for the perception test. The same sentence was used the test signal for the ALIZE/Spk trained on another part of the corpus. Results shows that the voice proximities within a familial group were well recovered in the speaker representation by ALIZE and much less returned in the representation from perception test.

1.Introduction
The courts have a growing interest in the field of forensic phonetics, more particularly in speaker recognition. Evaluating the human ability to recognize a voice or the performance of a speaker recognition system is a major issue for taking into consideration a witness or an expertise. Several studies, mainly carried out by psychologists, have investigated the human ability to recognize voices. The voice-based identification rates were very different according to the experimental context.

In 1937, McGhee [1] brought to light the effect of the length of time between the first listening test (one voice) and the second listening test (identification of the speaker among five). The identification tests took place one day, two days, two weeks, three weeks and five months later. After one day, the voice was identified with 83%, whereas it was identified with only 37% five months later.

The familiarity between speakers and listeners is also an important factor in speaker identification performance as demonstrated by the study by Hollien and Doherty [2] within a group of ten speakers. Listeners who knew the ten speakers correctly recognized them with 98%, while listeners who did not know the speakers recognized them with 40%.

Emotional expression is instrumental in speaker identification by human. Saslove and Yarmey [3] in 1980 conducted an experiment where 15 listeners had to identify familiar speakers from sentences spoken in either angry or neutral tones. All the listeners correctly identified speakers with the neutral sentences, while only four of them succeeded in identifying the speakers with the angry sentences.

Furthermore, the length of sentences had an effect on the human ability to recognize voices. Indeed, Blatchford and Foulkes [4] compared the human performances with short sentences like “Get him” and with long sentences like “Face down on the ground and hands behind your back now!”. The percentage of correct identification varied from 52% with the short sentences to 82% with the long sentences.

Some studies pointed out that the characteristics of the speaker relied on different language levels. Analyzing the segmental level, Amino and al in 2007 [5] had shown that phonetic features played a role in speaker identification: Nasal segments provided more information on speaker identity than oral segments (82% and 77%, respectively). Regarding the suprasegmental level, Van Dommelen [6] had shown, with the help of signal transformations, that pitch contour and pitch height modified the auditor choices during an identification task.

Furthermore, the National Institute of Standards and Technology (NIST) has been coordinating yearly evaluations of speaker recognition (SRE) technology since 1996 [7]. The main tasks were One-Speaker Detection with Limited Data and One-Speaker Detection with Extended Data where the session data were controlled (type of microphone and phone, speakers' language...). The task consisted in determining if the test segment was produced by the same speaker who produced the training segment. System performances were assessed using the Equal Error Rate (EER) or the minimum of the decision cost (minDCF) estimated on false rejection and false acceptance.

In NIST-SRE, speaker verification systems have reached excellent results [7]. For example, ALIZE/Spk, an open source toolkit for text independent speaker recognition based on the GMM/UBM approach, had obtained a 4.38% EER with male data in 2005 [8]. The same year, the system developed by the SRI [9], a phone-based polynomial SVM system, had got a 4.21% EER.

These performances fluctuated according to various factors. The size of the data used in training and testing are instrumental as demonstrated by [10] on NIST-SRE 05. Indeed EER varied from 8.67% to 15% when the length of the test signals had dropped by 2.5 min to 10s and EER reached 20% when the length of the training signals had dropped by 2.5min to 10s.

If most of the systems used MFCC, some of them had added segmental or suprasegmental information. The SRI’s system was based on phoneme classes and obtained different EER according to the phoneme models. The EER was 6.25% when using nasal and glide models, whereas the EER amounted to 8.41% with obstruct model [9]. The suprasegmental can also provide information. For example, the system described in [10] was a combination of different sub-systems which one used the prosody.

Several factors seem to have the same impact for both human ability and speaker verification system performance in voice speaker recognition: Familiarity (that can be related to the length of the training data for speaker verification system), length of the test sentences, segmental features, prosody. Do human and speaker verification systems use the same information from speech signal in speaker recognition?

The aim of this paper is to compare the speaker verification system and the human behavior with respect to the same data. The matter is not to evaluate the human ability in speaker identification nor the speaker verification system performance.

In NIST-SRE detection task, the evaluation is constituted by approximately 12500 tests in phone transmission where 2 sig-
nals are compared (one for the training and one for the test). Speaker verification systems provide a likelihood ratio (on which is based the decision) between training data and test data. Therefore, listeners had to do similar tasks: Attributing a proximity value to a couple of signals. Obviously, the number of tests could not be as high as 12,500 tests. 
The first part of this paper contains the description of the data used for the two experiments detailed just after. In the second part, the results are presented for the perception test and for the speaker verification system. The last section synthesizes the analyses and discusses the signal characteristics probably used in these comparison tasks.

## 2. Data and experiments

Two kinds of experiments have been realized to compare the answers of humans and speaker verification system. The first experiment was a perceptive test where listeners had to evaluate the “proximity” between two speech signals. The second used an speaker verification system, ALIZE/SpkDet which is an open source toolkit[8]. For each experiment, an easier representation has been proposed to visualize the proximity scores for the signal pairs.

In both tasks, the aim is to get a “proximity value between two voices” for a couple of signals; therefore the data was collected from speakers who were hypothesized to have “similar voices”: groups of French speakers with the same gender and with familial relationships.

### 2.1. Data

#### 2.1.1. Corpus

This corpus contained four parts. The first section (Part I) was an aloud reading of 94 words which were phonological alternations. The second one (Part II) consisted in a list of 29 sentences which favoured some specific features for the vowels. Some sentences were composed with a lot of nasal vowels like un bon vin blanc /ebóvəblɛ/ or back vowels (e.g. Poule aux choux rouges / puloʃrouʃ/). The vowel height was also taken into account: High vowels (e.g. Un archiduc russe stifle une musique /œnarkiˈdykuːsɨʃlyˈmyziktyks/), mid vowels (e.g. Gébé hébé té a essayé de fleter /ʒeɡebebeʃetɛʃəfletɛ/). One type of French vowels was concerned in each sentence, but all the sentences hold the French vocalic system. The third part (Part III) was an aloud reading of a text where some of the 94 words were used in context. Part I and III are compatible with the PFC protocol [11]. The fourth section (Part IV) was a free talk of about ten minutes with the interviewer, in order to have more spontaneous speech. This last part has not yet been exploited.

#### 2.1.2. Speakers

24 speakers (13 women and 11 men) had been recorded forming 10 blood relationship groups: 2 groups of sisters, 1 group of brothers, 3 groups of daughter/mother, 4 groups of son/father. Table 1 summarizes the relationships between speakers.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Relationship</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1/F2/F3</td>
<td>Mother/daughter/daughter</td>
<td>Great-daughter/great-daughter/great-mother</td>
</tr>
<tr>
<td>J1/J2</td>
<td>Sister/sister</td>
<td>Great-daughter/great-daughter/great-mother</td>
</tr>
<tr>
<td>M1/M2/M3</td>
<td>Father/son/son</td>
<td>Max/Min/Max</td>
</tr>
<tr>
<td>N1/N2</td>
<td>Father/son</td>
<td>Max/Min/Max</td>
</tr>
<tr>
<td>Q1/Q2</td>
<td>Father/son</td>
<td>Max/Min/Max</td>
</tr>
<tr>
<td>P1/P2</td>
<td>Father/son</td>
<td>Max/Min/Max</td>
</tr>
<tr>
<td>Q1/Q2</td>
<td>Brother/brother</td>
<td>Max/Min/Max</td>
</tr>
</tbody>
</table>

It is worth noting the important difference in age: The youngest was fourteenth while the oldest was eighty. All the speakers were recorded with the same cardioid microphone connected to a Marantz PMD 660. Speakers were recorded at their home, in a quiet room. The recording lasted roughly 15 minutes for each speaker. The whole corpus length was about 6 hours. The speech signals were sampled at 44.1 kHz.

### 2.2. Perception test

In this experiment, a single sentence from the second part of the corpus: Que fait la fée / kofɛlefe/ was selected. An interesting point is that this sentence could be produced [kofɛlefe] or [kofɛlefe], according to the speaker vocalic system. The 24 stimuli were this sentence pronounced by each speaker and normalized in intensity. Listeners heard 266 pairs of stimuli, exclusively from the same gender (11*10=110 male stimuli or 13*12=156 female stimuli). For each pair, listeners had to give a score between 1 (the voices are very similar) and 5 (the voices are very different). This scoring procedure presented the advantage of leaving some leeway for listeners in their answers. Listeners never heard twice the same stimulus. So, no score was given for intra-speaker comparison. The perception tests lasted about 40 minutes. At the end of the session, listeners explained what the indicators they used to meet the task were in their opinion.

### 2.3. Speaker recognition system

ALIZE is a free software based on the well-known UBM/GMM approach. It includes the latest speaker recognition developments such as latent factor analysis. Its performance has been demonstrated within the framework of the NIST’06 SRE evaluation campaign [6]. The use of ALIZE was divided in two steps, the model training and the testing.

#### 2.3.1. Training

In UBM/GMM approach, the models of each speaker are adapted from a Universal Background Model. This model is a GMM that represents the distribution of a large amount of speech data. The UBM models were those used by the LIA [12] in NIST’08 SRE evaluation campaign. During training, the mean values of UBM are adapted to model the distribution of a given speaker training signal. We adapted the model of each speaker with the words and text he or she read (Part I and III of the corpus). The training signals were similar in time to those used in NIST, which is about 2.5 minutes. As usually done, men and women were considered separately.

#### 2.3.2. Tests

First of all, the signal test was constituted by all the sentences produced by a speaker (Part II of the corpus). The length of these sentences was approximately 2.5 minutes. All the cross-speaker tests have been realized: 169 (13*13) tests for women and 122 (11*11) tests for men, in order to obtain a
score of each same-gender pair of speakers. These 291 tests, albeit small in number, were similar in time to the NIST tests. Nevertheless, a second set has been carried out with the sentence Que fait la fée ? to be strictly comparable to the perception test device. During this set, the test signal lengths were about 1 sec, which was much lower than the shortest test segments in the NIST evaluation [7].

For each set, a matrix of scores $s(i,j)$ for a signal test $X_i$ compared to the model $M_j$ was obtained:

$$s(i,j)=\log p\left(X_i|M_j\right)-\log p\left(X_i/\text{UBM}\right)$$

(1)

It is worth noting that men and women were considered separately.

### 2.4. Geometric representation of results

The both experiments previously described provided either a matrix of perceptive distance or a matrix of scores. These matrices were not an easy way to interpret the relations between the speakers. Therefore, we applied a bijective transformation to obtain a spatial configuration of the speakers.

#### 2.4.1. Perception test

In the perception test, listeners annotated the proximity between two voices. All the pairs were presented; therefore we had a value for stimulus $A$-stimulus $B$ and another value for stimulus $B$-stimulus $A$. Equivalent pairs were combined and averaged, to symmetrize the resulting matrix. To obtain a spatial configuration, we use the Multidimensional Scaling algorithm (MDS) [13]. Multidimensional scaling “is the problem of representing $n$ objects geometrically by $n$ points, so that the interpoint distances correspond in some sense to experimental dissimilarities between objects”. We used this technique to obtain a 2D spatial representation of the speakers.

#### 2.4.2. Speaker verification system

Collet et al. proposed a speaker verification system using anchor models [REF]. In this approach, a speaker was represented relatively to a set of well-trained GMM models. A PCA/LDA orthogonalization was applied. In our experiment, the anchor models were all the speaker models. PCA was applied and the 2D space represented by the two first axes of the PCA was used as the geometric representation of the speakers, stand for the maximum of variance.

### 3. Results

#### 3.1. Representation from ALIZE

##### 3.1.1. Testing all the sentences

First of all, the tests were performed with the Part II of the corpus for the 11 men and the 13 women separately. After PCA, the two first axes represented 85% of the variance. The geometric space obtained is presented in Figure 1 for the men. We found 3 groups of speakers. The first is composed by N1, N2. The second is composed by M1, M2, M3 plus Q2. The third is more heterogeneous and grouped O1, O2, P2, P1, Q1. This representation highlights the ALIZE mainly retained the familial relationships in its speaker voice characterization. Two remarks can be done: i) the two brothers Q1/Q2 voices were considered as very different; ii) the two groups O1/O2 and P1/P2 were very close together.

![Figure 1: Representation of the 11 men speakers from ALIZE scores obtained on part II.](image1)

##### 3.1.2. Testing “Que fait la fée ?”

The tests were performed with the single sentence Que fait la fée ? /œflaf]/ used in the perception test. Figure 2 shows the geometric representation obtained for males on that sentence. The two first axes explained 85% of variance. Familial groups are roughly similar to those obtained with the entire Part II of the corpus. Both spatial representations show the same structuring even if the data available were much shorter than the segments used in NIST-SRE (1 s vs. 2.5 min).

![Figure 2: Representation of the 11 men speakers from ALIZE scores obtained on “Que fait la fée ?.”](image2)

#### 3.2. Representation from perception test

29 listeners (19 females and 10 males) who did not know the speakers took part in this perceptive test. They were native speakers of French without known hearing impairment. A symmetric matrix was calculated by averaging the individual scores for the equivalent pairs of stimuli. A zero value was added in the diagonal of the matrix. The mean variance between listeners’ scores was 0.85 with values from 1 to 5, showing a quite substantial agreement among listeners. The MDS space is presented in Figure 3. The MDS stress and the distance correlation were 0.08 and 0.88 respectively. The 2D space was calculated for the women separately (MDS stress = 0.07 $R^2 = 0.82$). The proximities between the familial voices are much less obvious than what was observed for ALIZE (cf. Figure 2 for men). The choice of the recorded speakers
was based on their voice proximity, i.e. their voices could be confused on phone. Such dispersion in familial voices was unexpected. It might be due to the procedure used in that perception test where two voices were presented simultaneously, highlighting differences between voices rather than similarities.

4.Discussion and conclusion

With these both experiments, we obtained two representations of proximity between speakers related to the human perception and to a speaker verification system. Theses representations revealed different patterns within familial groups. In these pairwise comparisons of voices, humans seem to enhance differences. When listeners expressed what indicators they used to score the voice similarity, they pointed out the intonation, the age, the rhythm and the F0 values. Furthermore, we noticed that in the 2D representation from ALIZE, men produced [kəfelafə] on the left (M1 M2 M3 N1 N2 Q2) and [kəfelafə] on the right (O1 O2 P1 P2 Q1). This observation has been confirmed when analyzing the vocalic system of the speakers thanks to the Part I and II (cf. Table 2). It is worth noting that Q1, even if his phonologic system did not opposed /ɛl vs /el/, he read aloud the sentence [kəfelafə].

Table 2: Vocalic systems for men and women

<table>
<thead>
<tr>
<th>Vocalic system</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>/juɛəʊə/</td>
<td>N1 N2 M2 M3</td>
<td>G1 G2 G3 H2</td>
</tr>
<tr>
<td>/juɛəʊʊə/</td>
<td>M1 Q1 Q2</td>
<td>F3 I2 J1, J2</td>
</tr>
<tr>
<td>/juɛɛʊə/</td>
<td>P1 P2</td>
<td>F1 H3 H2</td>
</tr>
<tr>
<td>/juɛæʊə/</td>
<td>O1 O2</td>
<td>F2 H1</td>
</tr>
</tbody>
</table>

Could prosodic or phonological features be relevant for explaining the axis of the representations obtained? A preliminary experiment had been conducted. The F0 mean and range, the length of the vowels in the sentence, and the difference between the first, the second and the third formants in /ɛl/ and /el/ were measured. The correlations between these measures and the position of speakers along the axes of the representations were calculated. No significant correlation was found for the speaker representation obtained from ALIZE. Regarding the MDS space for men, the correlation between the acoustic parameters and Axe 1 was $R^2 = 0.95$, but not significant ($F = 8.45$, $p = 0.054$). Acoustic parameters related to F0, duration and formants have been studied separately. The main part of the correlation was due to the F0 ($R^2=0.58; F=5.49, p<0.05$). Regarding the women MDS space, Axe 1 was slightly correlated to the correlation ($R^2=0.44; F=3.94, p=0.055$) and Axe 2 with the F0 ($R^2=0.42; F=3.70, p=0.063$).

This preliminary study had several weaknesses: the acoustic parameters were basic and the number of speakers was only 11 men and 13 women. Nevertheless the main indicators expressed by the listeners at the end of the perception test had been confirmed.

In conclusion, this study has compared the proximities between speakers from several familial groups using both a speaker verification system and a perception test. The voice proximities within a familial group were well recovered in the speaker representation by ALIZE and much less returned in the MDS space from perception test.

5. Acknowledgements

The authors would like to thank Jean-François Bonastre and Laurent Besacier for their help.

6. References