AURAL IDENTIFICATION OF UNFAMILIAR VOICES BY COMPARISON

M. Falcone(*), F. Ferrero(+), A. Paoloni(*), P. Cosi(+)
(*) Fondazione Ugo Bordoni, Roma, Italy
(+) Centro di Studio per le Ricerche di Fonetica (CNR), Padova, Italy

ABSTRACT

The performance of human listeners is the final objective of the automated speech input systems. In speaker recognition the need of a human-reference to assess automated method is a common procedure. We address the problem of identify unfamiliar voices, assuming that the listeners use a limited amount of speech to create the reference template. Our aim is to define and assess standard procedures to evaluate listener's capability in speaker recognition.

INTRODUCTION

The human speaker recognition capability is based on two main characteristics of the speech signal: '*acoustic-phonetic*' matching, '*prosodic*' matching. In this work we are interested in the first item, so this paper is concerned with the following experimental situation: a subject listen to a pair of *short* utterances and then he/she had to decide if the listened utterances belong to the same speaker or no. This is a common situation in the experimental evaluation of listener's ability to perform some speaker verification task [1][2]. Unfortunately in the past these tests were mainly intended to provide a basis for comparison with performance of automated systems. So the test design result substantially different time to time, and performance analyses do not allow cross comparison among the several experiments.

THE TEST DESIGN: A PROPOSAL

The test consists of listening to a pair of the same word, spoken by same/different speaker and then to give a judge on the speaker identity. The stimuli we use are: monosyllabic, trisyllabic, polysyllabic (more than five, less than eight syllables).

The response is gauged to a fixed number of choices. We fixed he following constrains to the listening test material:

- the same speech segment is never presented twice to the same listener;
- the amount of speech signal presented is the same for all speakers used in the test;
- the number of same-speaker pairs is the same of different-speaker pairs;
- the frequency distribution of the used words is uniform.

If **NS** is the number of the speakers available, **NR** is the number of repetitions of the stimulus (for each speakers), and **NC** is the numbers of pairs to be presented to the listener, the previous conditions set the following rules:

$$NR = 4 * (NS - 1)$$
$$NC = 2 * NS * (NS - 1).$$

So you have the following possible solutions:

Table 1. A list of possible values to be used in order to have a "balanced" test

NS	NR	NC
2	4	4
3	8	12
4	12	24
☞ ***5***	***16***	***40***
6	20	60

and so forth.

The duration of the listening test should be about 30 minutes long. Considering that we want to use three different words, the pointed solution of NS=5 results a good choice, as the number of pairs for listening session is 3*40=120, i.e. we consider, on the average, a total duration of 15s for each pair presentation plus user response. The definition of standard procedures for listening test in speaker recognition is a very important point.

As speaker verification and identification technology finally seem to have reached a mature degree, we expect a renewed and greater interest on these topics. The test design we propose may be a good starting point.

INSTRUMENTAL SET-UP

The listening tests have been executed in Rome and in Padua. The used hardware was the same in both laboratories, i.e.: a personal computer equipped with an audio OROS AU2X board, a CD reader and a colour VGA monitor. In addition an external amplifier and a monitor headphone AKG mod.K141 complete the required hardware to run the experiment. As hearing level is a crucial point in any listening test, special attention has been devoted to the calibration of the whole audio instrumentation chain.

Calibration

This problem may be split in two parts: the digital 'calibration' of the speech files; the analogue 'calibration' of the electrical chain from the line out of the audio board to the output of the headphone. The numerical normalisation of the speech signal is executed on line during the restitution of the speech file. The normalisation factor has been computed in order to amplify to a fixed dB value the frame (26ms) of maximum energy of the given stimulus. So the frame-peak energy is the same for all the stimuli. To calibrate the electrical equipment a reference 1kHz sinusoidal tone is used (see CCITT G711 recommendation). A MCL (Most Comfortable Level) strategy has been used. A measure of the mean speech levels after calibration, stated a value about 80dBA, that is, according to the measures reported in literature, a *reasonable* calibration level.

EXPERIMENT DESCRIPTION

The experiment is described by a test control file that contains information on the utterances to be played and the relative normalisation factors, as well the number of pairs to be presented and the filename where the results are saved. We build four tests: each test consists of 120 pairs.

Table 2. The list of the utterances used in the four tests executed in the experiment

	1sill	3sill	polisill
Test1	si	cancella	unoecinquedi
Test2	no	corretto	aeffebiotto
Test3	tre	indietro	novesetteduetre
Test4	sei	esegui	richiestadiaccesso

The speech material is part of the SIVA database [3], and it is real telephonic quality signal. We only use five speakers in this experiment; the same for all four tests. They belong to the same regional area of the South of Italy. The listening sessions have been executed in Rome (central Italy), and Padua, (north Italy) where listeners have not acquaintance with the southern speaker behaviours. Listening session have been executed in a *silent* room. Listeners do not perform any training, they only receive a page of written description of the test and relative instructions. Each laboratory contributed with 5 sessions per test, for a total of 20 tests. In summary we have responses on 4800 pairs' presentation. The subject can not listen more than once a pair. In fact after the pair presentation a menu describing the following four choices:

1.voices are certainly different
2.voices are probably different
3.voices are probably the same
4.voices are certainly the same

is displayed, and after the subject's selection the relative choice is highlighted on the monitor and the other choices are cancelled, then a confirmation is requested before the next pair will be submitted, otherwise the main menu is displayed again for a new selection. So corrections are possible, but the subject is not allowed to listen more than once the same pair.

ANALYSIS OF THE RESULTS

The obtained results have been analysed separately for the two groups, and then compared. Our goal is to measure the "human" performance in comparing speech samples in relation to the duration of the utterance (fig.1, fig.4); to analyse the listener variability in performing the identification task (fig.2, fig.5); and last to trace a '*relative operating characteristic*' (ROC) of the 'human' system in solving the given task. Direct measures of the obtained performances are the '*false acceptance error rate*', (FA) and the '*false rejection error rate*' (FR). The first is also referred as error TYPE I° and it is the probability that utterances of two different speakers

are assigned to the same person, it is the most severe error as it measures the probability that an impostor get in your system. The second is also referred as error TYPE II° and it is the probability that utterances of the same speaker are assigned to two different speakers, it is a less severe error as it measures the probability that your system do not let you get in, although you have the authorisation. These are 'crude' measures that only reflect the YES/NO decision taken. A more interesting measure is the ROC (fig3, fig.6) [4]. This is a standard XY dispersion plot where on the X axes there is the ***probability of listener deciding same when samples are, in fact, by different speaker*** (error), while in the Y axes there is the ***probability of listener deciding the same when samples are, in fact, by the same speaker*** (correct). If you have a total (positive plus negative) N rating scale the result is a set of (N-1) points on the graph. The fitting of these points, plus the origin point (0;0) and the infinite point (1;1), gives you the ROC curve. Roughly speaking, we may say that curves approximating the diagonal line from (0;0) to (1;1) describe more difficult tasks. If a real (automated) system measures a distance between two samples, it will be possible to set several thresholds and design a *real* ROC; in case of listening test this is *simulated* using a rated scale. Unquestionably a ROC curve gives a more detailed information than FA and FR values, but the standard procedure, well described in [4], considers the rating scale a linear discriminative scale.

Rome group result

The results obtained from the FUB group confirm the well-known fact that the performances in identifying speaker do not vary meaningfully after the 1.2 second duration. We see from fig.1 that errors decrease about 20% if we move from monosyllabic words to trisyllabic words, but only 5% from trisyllabic to polysyllabic words. This goes against our intuitive belief, but it is a well experimental accepted fact. From fig.3 we realise that the listener population has a great variability and that some subject also has a *strange* behaviour. For example one subject has a total error that is greater than 50% (a random generator works better!) and another has a 0% FA error and a 32% FR error (he/she is a good guardian!). Finally we had to compare the fig.1 and fig.3. They both report FUB listener group performance in relation to the used word, but in the first case we only use the binary Y/N information, while in the second we also utilise the degree of confidence the listener express utilising the two rates scale.

Padua group result

As expected the CNR results follow the same trend of the Rome listener group. We only find two light differences. First the overall error (FA+FR) is just a little bit greater, and this may be because northern listener may have less familiarity than central listener with the used database. Second the FA/FR ratio is smaller. This fact may not be explained easily. Also for CNR listener group we have some subject with strange behaviours. For example we have, again, a listener that has a 0% FA error, and another with a total error greater than 50%. The ROC curves clearly set that the speaker identification using monosyllabic word is really a hard task, while it makes no particular differences recognising people using trisyllabic words or polysyllabic words.

CONCLUSION

We execute a round-robin experiment for the evaluation of human capability in speaker recognition, when pairs of short utterances are submitted to the listener. Particular attention has been devoted to the calibration and balancing of the test itself, to avoid drift effects. The obtained results show high consistency among the two groups, and clearly set that: listeners belonging to a regional area farther (in a phonetically sense) from the one of the speaker to be recognised, have, on the average, a worse performance of few percents; performance response in relation to word length shows a threshold effect situated between monosyllabic and trisillabic words (other works report a value around 1.2s) for pairs of words or short utterances. The results are promising and although more efforts are necessary before a final solution will be reached, the possibility of using standard listening test as a reference in speaker recognition seems a good choice.

REFERENCES

[1] Stevens, K.N., et alii, "Speaker Authentication and Identification", *J.A.S.A.*,vol.44,n.6,pp.1596-1607

[2] Federico, A. et alii, (1989) "Comparison between automatic methods and human listeners in speaker recognition task", *proceedings EUROSPEECH 89*, pp.279-282

[3] Falcone, M., Contino U., (1995) "Acoustic characterisation of Speech Databases: An Example for the Speaker Verification", *these proceedings*

[4] Kecker, M., (1971), *Speaker Recognition: An Interpretive Survey of the Literature*, AHSA Monographs n.16

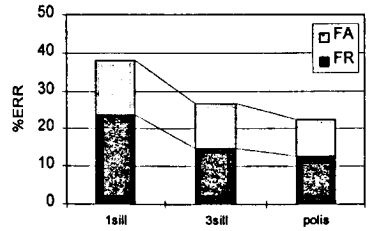

Figure 1. False Acceptance (FA) and False Rejection (FR) in relation to the utterance length. FUB listener group.

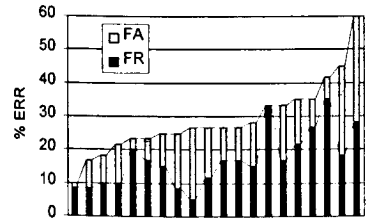

Figure 2. False Acceptance (FA) and False Rejection (FR) for each single listener. FUB listener group.

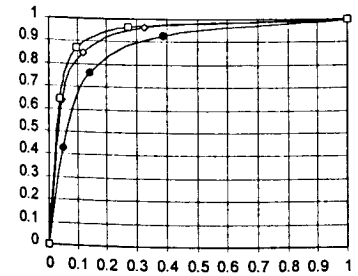

Figure 3. ROCs for monosyllable (•), trisyllable(+), and polysyllable (□) utterances. FUB listener group.

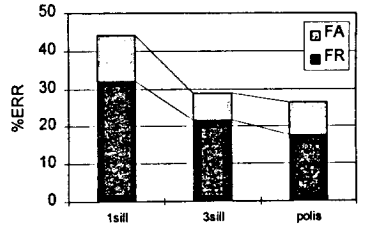

Figure 4. False Acceptance (FA) and False Rejection (FR) in relation to the utterance length. CNR listener group.

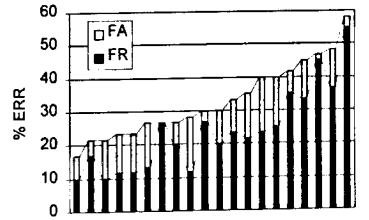

Figure 5. False Acceptance (FA) and False Rejection (FR) for each single listener. CNR listener group.

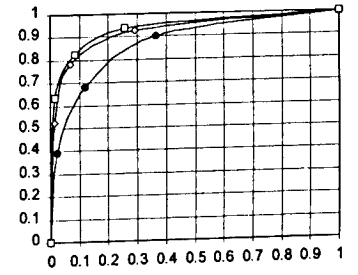

Figure 6. ROCs for monosyllable (•), trisyllable(+), and polysyllable (□) utterances. CNR listener group.