Speaker Identification By Computer And Human Evaluated On The Spidre Corpus

Hassan EZZAIDI and Jean ROUAT
ERMETIS, DSA, Université du Québec à Chicoutimi, Chicoutimi, Québec, Canada, G7H 2B1.

1. INTRODUCTION

Although many experiments on clean speech report high identification rates for computer systems, results on noisy telephone speech with different handsets are usually too poor for practical identification tasks (noise, limited bandwidth, effect of the channel, telephone handsets variability)[1].

What would be the identification rate of humans in the same conditions? A reference is necessary in order to evaluate the performance of computer systems. The comparison between computer and human has been already made. For a review one can refer for example to the work by Doddingtion [2]. As the performance of human has been chown to be dependent of the speech nature, we propose to examine the effect of telephone handset variability for text-independent speaker identification of telephone speech. We report human and computer speaker identification with the SPIDRE database.

Section 2 describes the experimental conditions while section 3 and 4 are the results and discussion. Section 5 is the conclusion.

2. EXPERIMENTAL CONDITIONS

2.1 SPIDRE database

Closed set Speaker Identification experiments were performed on a SPIDRE subset of the Swichboard corpus with *matched* and *mismatched* telephone handset conditions. We refer to the *matched* conditions, when (for a same speaker) the training and testing sessions were collected from the same telephone handset. In the *mismatched* conditions, different handsets were used for training and testing sessions.

Based on the pitch frequency (F_0) distribution, we first ran a speaker identification experiment on the 45 speakers of the SPIDRE corpus. We then extracted the most confusable speakers to create a subset of 10 women speakers (Female speakers with similar F_0 distribution). Each speaker has 4 conversations originating from 3 different handsets. The sampling rate is 8 Khz.

2.2 Listening conditions

Sixty pairs of sentences were randomly chosen and played through a Sennheiser HD250 linearII headphone. Ten naïve listeners (one woman and nine men) were asked to tell if the speaker was the same for both sentences. The listeners could not use sex as discrimination criteria. For each sentence, five seconds of speech were played. The listeners are French speaking and most of them could not understand spoken American English. For each pair, the listener had to make four choices:1. certainly the same speaker; 2. probably the same speaker; 3. probably different speakers; 4. certainly different speakers.

2.3 Computer experiments

Speech analysis

The speech is first preemphazised (0.97), then, a sliding Hamming window with a length of 32 ms and a shift of 10 ms is positioned on the signal. Twelve cepstral Mel coefficients, twelve delta Mel cepstral coefficients (computed according to the regression weighting), one log power and one delta log power are then extracted by using a liftering of 22 . Cepstral mean normalization is also performed. The final dimension of the MFCC vectors is 26.

Identification

We use two clustering technics. The first recognizer is based on a nonparametric pattern recognizer. For now, we use the LVQ-SLP as proposed by J. He and al. [3]. Each speaker is characterized by one codebook. The codebook size is the same for all speakers. We performed experiments with codebook sizes of 128, 256 and 512. The second recognizer is based on a parametric estimation of the probability distributions of the MFCC. A Gaussian Mixture Model (GMM) is associated to each speaker (one model for each speaker). For a given speaker, the GMM is supposed to model the statistical distribution of the MFCC. We used models with 32 Gaussian mixtures. We also assumed a diagonal variance matrix for each mixture component and parameters were estimated via the E.M algorithm.

Training and testing

The impact of mismatched and matched conditions is evaluated. In matched condition, the same telephone handset is used for training and testing. One conversation is used for training and the second one for testing. With mismatched handsets, training is performed on 3 conversations (pronounced through 2 handsets) and testing is made on the 4^{th} Conversation coming from the 3^{rd} handset.

Recognition criterion

The tested conversations were divided into fixed block lengths of 10 ms. With the GMM, the log likelihood of each block is computed, whereas, the nearest neighbour algorithm is used for the LVQ-SLP. A speaker is recognized if, for the entire test conversation (all blocks) it has the minimal distance (LVQ-SLP) or the maximum-likelihood (GMM).

3. EXPERIMENTS AND RESULTS

3.1 Listening tests

Table 1 reports rates for intra-speaker (columns 2 and 3) and inter-speaker (column 4) identification. Listeners are reported in column 1. In the matched conditions, one finds an averaged rate of 81%. The variance is significant and is mainly due to listeners L1 and L6. For the mismatched conditions, the recognition rate falls of 11%, with a weaker variance. In the case of the inter-speaker identification it is not possible to verify if the same telephone handset can yield confusions between speakers (speakers declared to be same speaker instead of declaring different) as the database labeling does not include the description of the handset characteristics. It is observed that the identification rates are coherent with those of

columns 2 and 3.

Analysis of the tests clearly shows that the handset has a predominant influence on the perception of listeners. In many situations with the same speaker and two handsets for the two sentences, listeners identified the two conversations as coming from different talkers.

Listeners	Identical speaker in test		Different speakers in test
	In Matched condition	In Mismatched condition	
L1	60 %	68 %	67 %
L2	90 %	74 %	77 %
L3	90 %	72 %	75 %
L4	70 %	72 %	72 %
L5	90 %	72 %	75 %
L6	50 %	64 %	65 %
L7	90 %	71 %	92 %
L8	83 %	62 %	75 %
L9	100 %	72 %	62 %
L10	90 %	72 %	75 %
Mean σ	81 % 16	70 % 4	73.5 % 8.2

Table 1 : Averaged scores for 10 listeners . Last column, refers to conversation pairs involving two different speakers using unknown handsets.

3.2 Computer tests

Handsets condition	*Codebook size (LVQ-SLP)*			*32 GMM*
	512	*256*	*128*	
Matched	90 %	90 %	90 %	-
Mismatched	60 %	60 %	60 %	90 %

Table 2: Computer speaker recognition rates. *Matched* (One conversation in training, another in testing, identical handset); *Mismatched* (Three conversations in training, another in testing, different handset).

The LVQ-SLP recognizer yields an identification rate of 90% when talkers use the same telephone handset. With different handsets in training and testing (mismatched) the scores drop to 60%. It is interesting to note that the LVQ-SLP rates are independent of the codebook sizes.

The GMM recognizer outperforms the LVQ-SLP recognizer when mismatched handsets are used (90% in comparison to 60%). With the same handset we would expect better results.

4 DISCUSSION

It is observed that the confusion between speakers is mainly due to the strong telephone handset influence. Thus, the speaker acoustical characteristics are found to be largely degraded by the telephone handsets.

Except for L1 and L6, all listeners presented the same faculty to distinguish between the 10 women. In matched conditions, if one does not consider L1 and L6 in the statistics, the average identification rate can increase around 90%.

Interestingly, the difference in performance when changing from matched to mismatched condition is smaller for listeners than for computers.

Although the task presented to listeners and computers is not comparable – the listeners task is easier with a comparison of two speech segments and the computer has to carry out the classification between 10 speakers presented simultanously - it is observed that the mismatched conditions degrade the performance for human and computer.

It is possible to infer that the success of the computer is related to the efficiency of the models and to the quality of the parameters (reduction of the channel and handset effects) for the subset of 10 women.

5 CONCLUSION

Even if the task was easy in comparison to the identification of forty speakers, the relatively low performance of the listeners gives an idea of the complexity of the SPIDRE corpus.

The selection of ten females has been based on pitch frequency distribution. They have a similar distribution of pitch. We already found that based on the pitch, the task was tedious for computers when using exclusively cues derived from the pitch distribution [4]. Manipulation of the pitch frequency confuses listeners when identifying speakers [5]. This suggests that pitch frequency is in fact a fundamental cue which can not be fully exploited by listeners on our test set because of the pitch distribution similarity. Furthermore, Itoh and Saito [6] found that the spectrum envelope is more important in speaker identification than excitation. The MFCC recognizers rely mainly on the spectrum envelope and formants and are probably more accurate than listeners to identify speakers based on spectral characteristics only.

For a subset of 10 female speakers with high confusable pitch distribution the recognizer based on the MFCC and GMM outperforms the listeners.

ACKNOWLEDGEMENT

Mohammed Bahoura wrote the listening tests and performed the listening evaluations. Many thanks are due to the 10 listeners.

REFERENCES

[1] C.R Jankowski Jr., T.F. Quatieri, and D.A. Reynolds. Measuring fine structure in speech: Application to speaker In *IEEE-ICASSP*, pages 325-328, 1995.

[2] G. Doddington: Speaker recognition-identifying people by their voices In *Proc. IEEE* Vol 73, pp1651-1664.

[3] Jialong He, Li Liu, and Günther Palm. Speaker identification using hybrid LVQ-SLP networks. In Proc. *IEEE ICNN'95*, volume 4, pages 2051—2055, 1995.

[4] J. Rouat, H. Ezzaidi et M. Lapointe. Nouveaux algorithmes d'extraction en vue de caractériser le locuteur. *Technical report*, ERMETIS, Université du Québec à Chicoutimi, March 1999. Contrat W2213-9-2234/SL, rapport final, 67 pages.

[5] S. V. Bemis and S. W. Nunn. Acoustic features and human perception of speaker identity In Proc. AVIOS, pages 85-96, 1998.

[6] Itoh K. and Saito S. Effects of Acoustical Feature Parameters on Perceptual Speaker Identity *In Review of the Electrical Communications Laboratories*, vol. 35, N0.1.