A QUANTITATIVE ASSESSMENT OF THE RELATIVE SPEAKER DISCRIMINATING PROPERTIES OF PHONEMES

J. P. Eatock † **& J. S. Mason**

University College Swansea, Singleton Park, Swansea, UK.

ABSTRACT

The aim of the study described in this paper is to provide a thorough and quantitative assessment of the relative speaker discriminating properties of phonemes. A VQ codebook based approach to speaker modeling is used in conjunction with a phonetically hand-labelled database to produce phoneme rankings based on speaker verification scores. In broad groupings the nasals and vowels are found to provide the best speaker recognition performance, followed by the fricatives, affricates and approximants, with the stops providing the worst performance of all. A comparison at the individual phoneme level produces a more detailed ranking and of particular interest is the surprisingly good performance of the unvoiced fricative /s/. The ranking of phonemes is found to be largely unaffected by changes in experimental parameters such as the model size, the feature type and the speaker population.

1 INTRODUCTION

It is generally recognised that some parts of speech are more useful for speaker recognition (SR) than others. Clearly with a greater understanding of which phonemes are the most reliable for use in SR, it should be possible to improve recognition performance. One way in which this might be achieved is to select passwords (or key phrases) containing a high proportion of 'good' phonemes. Alternatively, a front-end classifier could be used to automatically identify useful phonemes, enabling appropriate biasing to be applied in the recognition stage. We have described an equivalent approach in [1,2], but this was based on spectral clustering rather than phoneme classification.

2 PREVIOUS WORK

Previous work in this field is reviewed at length in [3] and can briefly be summarised as follows:

Hofker [4] presents a rank ordering of 24 isolated German phonemes, which indicates the nasals as providing the best SR performance, with the voiced fricative /z/ and the liquid /r/ also performing fairly well. The unvoiced fricatives /f/ and /s/ perform the least well. Kashyap [5] finds the phonemes /s/, /t/ and /b/ to perform less well than vowels and nasals. Broeders [6] shows /x/, /r/ and /s/ to perform on average better than /p/. Nolan [7] finds the liquids /r/ and /l/ to provide 'moderate' performance and points out that they are less useful than the nasals. Glenn [8] strongly promotes the nasals as providing good performance in SR, as does Su [9], who finds nasal coarticulation to be particularly useful. Both Sambur [10] and Wolf [11] find that the best parameters for use in SR are to be found in nasals and vowels. Goldstein [12] shows that phonemes with 'free variants' such as /r/ have a high speaker-discriminating content. Paul [13] finds that front vowels, high vowels and nasals provide the best performance.

3 PURPOSE OF STUDY

The above review reveals the lack of a thorough and quantitative assessment of the speaker-discriminating properties of phonemes. Each of the referenced studies examines only a relatively small subsection of phonemes. This study provides what is thought to be the first detailed assessment of phoneme performance, using a complete set of phoneme labels.

4 DATABASE

The database used here is of telephone quality, sampled at 8 kHz and comprises 125 speakers each uttering 6 sentences from a pool of 201 sentence-texts. The whole database is annotated by hand using a set of 75 phonetic labels.

5 EXPERIMENTAL APPROACH

To enable a comparison of the speaker-discriminating properties of the phonemes, a vector quantisation (VQ) ap-

† Now at Ensigma Ltd, Turing House, Station Road, Chepstow, Gwent, UK.

0-7803-1775-0/94 $3.00 © 1994 IEEE

proach is adopted. The database is split into test and training portions and 12th-order cepstral features extracted. Speaker-dependent phoneme models are constructed for every speaker deemed to provide sufficient data for both training and testing.

In the assessment phase a single phoneme is examined at a time. Length-normalised test tokens are used to remove possible biasing effects due to variations in phoneme lengths. For example it would clearly be wrong to compare the approximant /j/, which has an average length of 61 ms, with the vowel /ae/, which has an average length of 157 ms. The test tokens for the phoneme under examination are applied across all of the corresponding models and the overall performance recorded. Preliminary experiments [3] indicate that the speaker verification equal-error-rate (EER) is largely unaffected by variations in the speaker population. As the speaker population varies significantly in both size and make-up, across the phonemes, the EER coupled with the corresponding 95% confidence band, is deemed a suitable measure for this study.

6 PHONEME GROUP RANKING

The EER and corresponding 95% confidence interval, for each of the phoneme groups is presented in the bar graph of Figure 1. The numbers of test tokens and speakers used in this experiment are indicated in Table 1. In these broad groupings, the nasals and vowels provide the best performance, followed by the fricatives, affricates and approximants, with the stops providing the worst performance of all.

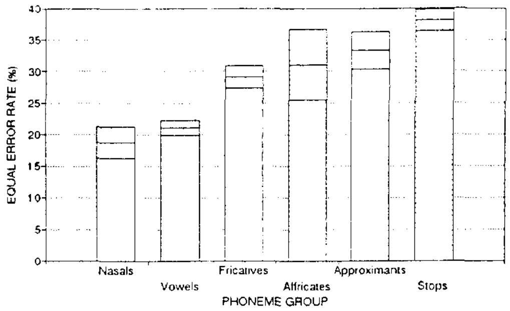

Figure 1: Phoneme Group Comparison: the bar chart compares the equal error rates for the phoneme groups and indicates the 95% confidence intervals.

PHONEME GROUP	NUMBER OF TEST TOKENS	NUMBER OF SPEAKERS	EER (%)	95% CONF. INT. (±%)
Nasals	926	125	18.8	2.5
Vowels	4713	125	21.1	1.2
Fricatives	2250	125	29.2	1.9
Affricates	257	71	31.1	5.7
Approximants	937	125	33.3	3.0
Stops	3038	125	38.1	1.7

Table 1: Phoneme Group Ranking: The groups are ranked from top to bottom according to increasing equal error rate (EER).

7 PHONEME RANKING

A more detailed ranking showing the EERs corresponding to the 35 phonemes, found to provide the smallest confidence intervals, is presented in Table 2. These are ordered from top to bottom according to performance, with the nasals and vowels near the top and the stops at the bottom. Of particular interest is the unvoiced fricative /s/. Although overall we find the fricatives to perform significantly less well than the vowels, we find /s/ to provide comparable performance to the vowels. The unvoiced fricative / / also exhibits surprisingly strong speaker discriminating properties.

8 FURTHER EXAMINATION OF /S/

It was hypothesised that the unexpectedly high performance observed for /s/ could be due to end-effects or segmentation errors. Indeed some occurrences of /s/ were observed to include portions of voiced speech at one end or the other. However, experiments in which portions of /s/ were discarded, show this hypothesis to be false and suggest that the central portions of /s/ are perhaps more useful than the end portions.

9 FURTHER EXPERIMENTS

Figure 2 shows a plot of the EER for female speakers versus the EER for male speakers, corresponding to the set of 35 phonemes found to provide the smallest confidence intervals. Clearly there is a strong correlation between the two sets of results. Interestingly the results for the female speakers are worst than those for the male speakers.

The graph of Figure 3 shows a plot of the EER obtained using fft-derived mel-cepstra features versus the EER obtained for lpc-derived cepstral features. There is a good corelation between the two sets of results, although it is noticeable that some of the vowels, in particular, have interchanged positions.

The phoneme rankings were found to be largely invariant to variations in parameters such as the model size and the volume of training data used. Also in general the steady state portions of speech were found to exhibit greater speaker-discriminating properties than the transitional parts, for the features used here.

IPA SYMBOL	EXAMPLE WORD	TRUE TALKER TESTS	EER (%)	95% CONFIDENCE LIMITS (±%)
ŋ	sing	371	19.7	4.0
ɑ	party SBS	1113	21.1	2.4
i	bead	1401	21.1	2.1
ɪ	kid	1027	22.5	2.6
ʌ	butter SBS	467	22.9	3.8
n	new	2175	23.0	1.8
m	month	1145	23.2	2.4
ei	gate WAL	531	23.7	3.6
u	soon	300	24.3	4.9
ɛ	pen SBS	1041	24.6	2.6
ɒ	one NB	969	24.7	2.7
ɜ	bear NB-LIV	319	25.7	4.8
s	straw	3289	26.0	1.5
aɪ	price NB	559	26.1	3.6
ʃ	ship	290	26.6	5.1
ae	sighed R-SCOT	381	27.6	4.5
ə	butter NB	1400	27.6	2.3
iə	clear SBS	277	27.8	5.3
l	lay R-IRISH	289	28.4	5.2
f	follow	1091	28.4	2.7
ɑi	price LON	727	28.8	3.3
a	trap WAL	1704	29.0	2.2
ɔ	warm SBS	992	29.2	2.8
z	easy	1087	30.2	2.7
ʤ	judge	260	31.5	5.6
ɹ	real	1048	33.5	2.9
ʧ	church	412	35.2	4.6
w	wear	662	35.2	3.6
h	hello	213	36.2	6.5
d	dream	1274	37.0	2.7
t	mast	3886	38.6	1.5
b	bathe	1956	40.9	2.2
p	pot	1360	41.4	2.6
k	mask	2246	43.6	2.1
ʔ	butter	480	45.6	4.5

Table 2: Phoneme Ranking: The phonemes are ranked from top to bottom according to increasing equal error rate (EER). This table includes only those phonemes which provide confidence limits below ± 7%.

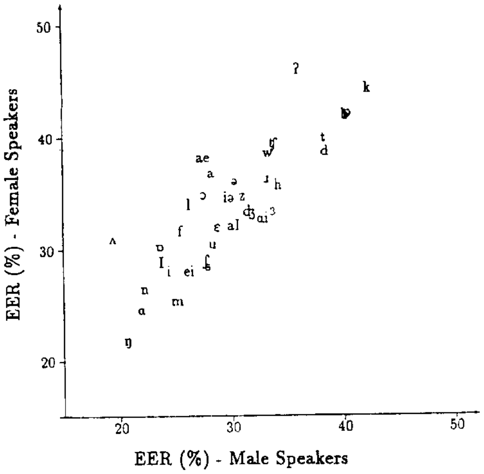

Figure 2: Plot of equal error rate (EER) for male speakers versus EER for female speakers.

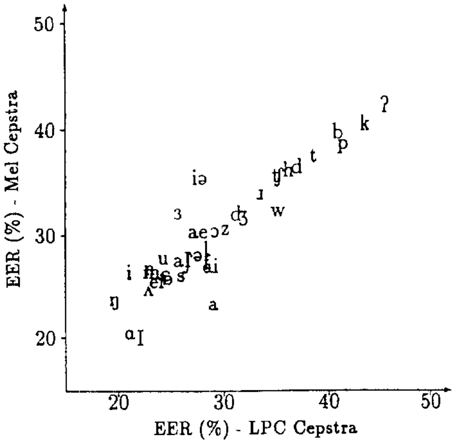

Figure 3: Plot of equal error rate (EER) obtained using fft-derived mel-cepstra features versus EER obtained for lpc-derived cepstral features.

10 DISCUSSION

The results presented here corroborate the results of a recent Dutch study by Heuvel [14], who finds nasals and vowels to outperform fricatives, which in turn outperform the plosives and /r/. He presents the following ranking for the consonants examined:

/m/ > /n/ > /s/ > /t,k,r/ > /d/ > /p/

Here we find:

/n, m/ > /s/ > /r/ > /d,t,p,k/

In accordance with the findings here, Heuvel reports that (i) the vowel steady states appear to contain more speaker information than the transitions and (ii) male and female speakers give almost identical rankings.

The surprising result that /s/ exhibits strong speaker-discriminating properties is clearly in conflict with Hofker's finding that /s/ performs poorly. This may be explained by the fact that whereas Hofker uses isolated phonemes, here the phonemes are extracted from continuous speech. In support of this, Soli [15] reports the presence of second formant peaks in the spectra of /s/, due to anticipatory vowel coarticulation effects.

11 CONCLUSIONS

In general groupings, the nasals and vowels are found to provide the best performance, followed by the fricatives, affricates and approximants, with the stops providing the worst performance. A comparison at the individual phoneme level, produces a more detailed ranking, and of particular interest is the surprisingly good performance of the unvoiced fricative /s/.

12 ACKNOWLEDGEMENT

This work has been supported by BT.

13 BIBLIOGRAPHY

[1] J. Eatock & J. Mason, ESCA, Edinburgh, 1990.
[2] J. Eatock & J. Mason, Proc. ICSLP, Japan, 1990.
[3] J. Eatock, Ph.D Thesis, U.C.Swansea, UK, 1992.
[4] U. Hofker, Proc. 9th Int. Cong. Acoust., Madrid.
[5] R. Kashyap, IEEE Trans. ASSP, 1976.
[6] A. Broeders, Eurospeech, Paris, 1989.
[7] F. Nolan, Ph.D. Thesis, Cambridge University, 1983.
[8] J. Glenn, JASA, Vol.43, No.2, 1968.
[9] L.-S. Su, JASA, Vol.56, No.6, 1974.
[10] J. Wolf, JASA, Vol.51, 1972.
[11] M. Sambur, IEEE Trans. ASSP-23, 1975.
[12] U. Goldstein, 1975.
[13] J. Paul, Proc. Carnahan Conf., Kentucky, 1975.
[14] H. van den Heuvel, ICSLP, Canada, 1992.
[15] S. Soli, JASA, Vol.70, No.4, Oct. 1981.