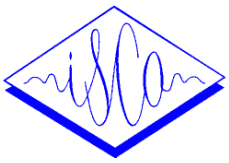

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Phoneme Performance in Speaker Recognition

J.P. Eatock & J.S.D. Mason
Department of Electrical & Electronic Engineering, University College Swansea, UK

10.21437/ICSLP.1992-377

1 Abstract

It is recognised that some phonemes are more useful for speaker recognition than others. For example, it is commonly accepted that unvoiced phonemes are less useful than voiced ones. This paper overviews published work investigating the speaker-discriminating properties of phonemes. Following this, a preliminary experiment is described, in which the performances of nasals and unvoiced fricatives are compared, using an on-line transputer-based speaker recognition system.

2 Introduction

While there is some evidence to suggest that certain phonemes exhibit greater speaker discriminating properties than others, there would seem to be surprisingly little work in this area especially in recent times. Indeed Heuvel [1] comments that "although a great deal of research on speaker recognition has been carried out over the past decades, the phonetic description of speech segments showing varying degrees of speaker-specific information has not received much attention as yet."

Clearly, with a greater understanding of which phonemes are the most reliable for use in the SR task, 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 phoneme classifier might be used to automatically identify phonemes and enable phoneme-dependent weightings to be applied in the recognition stage [2, 3]. We have described an equivalent approach in our previous work [4, 5] but this was based on spectral clustering rather than phoneme classification.

The aim of this research, therefore, is to investigate which phonemes provide the best SR performance. The approach adopted is to gather statistics from an on-line voice entry system, connected to the doors of the speech research laboratories at University College Swansea. Although the system is used by only a relatively small population (10 speakers), it is used regularly (almost daily and occasionally several times in one day) by each person.

3 Previous Work on Phoneme Assessment

Previous work in this field may be summarised as follows:

- **Höfker** [6] presents a rank-ordering of twenty-four German phonemes, indicating their relative performances in the speaker recognition task, for both a Mahalanobis and a Euclidean minimum-distance classifier, and suggests that on this basis code sentences can be selected for speaker recognition. His database comprises ten versions of each phoneme from twelve speakers. Höfker concludes that "there are significant differences between the phonemes with regards their information content concerning the speaker."

These rankings show that the nasals (i.e. /n/, /m/ and /ŋ/) provide the best speaker recognition performance. Höfker explains that this is "because they are mainly influenced by the nasal cavity, which can hardly be varied by the speaker". Of the consonants considered, as might have been anticipated, the unvoiced fricatives /s/ and /f/ perform the least well, whereas the voiced fricative /z/ is the next best after the nasals. The liquids /l/ and /r/ and the glide /j/ all appear in the top half of the overall phoneme rankings. According to this table the best vowels are /æ/, /i/ and /I/ and the worst are /o/, /u/ and /a/. As the phonemes used in Höfker's study are spoken in isolation, the results are perhaps not directly applicable to continuous speech, but nevertheless provide a good indication as to which parts of speech contain the most speaker-specific information.

- **Kashyap** [7] comments that "in the literature, one finds suggestions that some phonemes, like nasals or some vowels, are the best phonemes for speaker recognition" and presents the results of his own experiments to determine which phonemes are the most useful.

His initial approach is to take two phoneme sets, which differ by only one phoneme member, and compare their corresponding recognition performances. Using this method Kashyap determines that the phonemes /s/, /t/ and /b/ are less useful for speaker recognition than the vowels and nasals. Simple phonetic theory can perhaps account for this result. The fricative /s/ and the stops /t/ and /b/ are obstruents, for which the primary source

of excitation is a vocal tract constriction and it might therefore be expected that they perform less well than the nasals and vowels, for which the primary source of excitation is the glottis (sonorants). Furthermore the stops /t/ and /b/ are transitional phonemes whereas the nasals and vowels are produced under relatively steady state conditions.

In order to investigate which, of a set of nasals and vowels, is the most useful in the speaker recognition task, an alternative approach is used. This involves comparing "the different phoneme-speaker pairs by a suitable distance function which measures the discrepancy between two phonemes spoken by two different speakers". The phoneme set comprises of the vowels /i/, /I/, /ɛ/, /æ/, /u/ and the nasals /m/ and /n/. The results presented indicate that the vowel /i/ is the worst in the set, whereas teh vowels /I/ and /ɛ/ are the best, even outperforming the nasals. These results contrast with those of Höfker [6] who finds the nasals to provide the very best performance.

• **Broeders** [8] presents the results of a study to investigate the extent to which some specific consonants can function as cues in speaker recognition. The consonants /x/, /r/ and /s/ are selected because of their high inter-speaker variability in Dutch. The consonant /p/ is selected as the reference phoneme because it is reported to exhibit a low inter-speaker variability in Dutch. Sixteen speakers utter twenty-four monosyllabic words of the form CVC (consonant-vowel-consonant). Each of the consonants (/x/, /r/, /s/ and /p/) occur in each C slot, with the vowels /o/ and /a/ occupying the V slot, in 50% of the cases each. Speaker recognition results relate to five listeners.,

The consonants /x/, /r/ and /s/ are found on average to perform better than /p/, although Broeders points out that this trend is not significant at the 5% level. However "regardless of whether this response was right or wrong, listeners were significantly more confident of their scores for words containing /x/, /r/ and to a lesser extent /s/ than words containing /p/".

• **Glenn** [9] states that a problem with using the speech power spectrum in speaker recognition is that "since the articulators (tongue, lips, teeth) are in almost constant motion during normal speech, the configuration of the vocal cavities, hence the power spectrum of acoustic radiation is constantly changing". He points out however that one class of voiced speech sounds is produced with the vocal cavities and the articulators held fixed. This class is the nasal consonants. The frequency with which the nasals naturally occur in spoken American English is approximately 11% and this is highlighted as a further benefit. Based on the assumption that phonemes are produced at a rate of ten per second, Glenn suggests that a nasal consonant can be expected on average every second.

Experiments, using a long-time averaging technique, with a thirty speaker database of manually extracted examples of the phoneme /n/ give a recognition score of 93%. Glenn concludes that "the power spectrum of acoustic radiation produced during nasal phonation provides a strong clue to speaker identity". It is unfortunate that recognition scores are not presented for other phonemes.

• **Su** [10] suggests the coarticulation between a nasal and a following vowel as providing important speaker-specific information. When a vowel follows a nasal "the tongue anticipates the following vowel segment and moves to the vowel position during the nasal phonation. The extent of this coarticulation depends on whether the tongue takes a specific position to produce the nasal consonant. In English the tongue has no distinctive function for /m/, but has important functions for /n/ and /ŋ/. For /m/ before a vowel we would expect more coarticulation than for /n/ and /ŋ/ before a vowel."

Su measures the extent of the coarticulation present between consonants and following vowels. The nasals /n/ and /m/ are used with the front vowels /u/, /o/ and /a/ and the back vowels /i/, /e/ and /æ/. Su demonstrates that the coarticulation between /n/ and a vowel is approximately a third of that between /m/ and a vowel. It is also shown that the coarticulated nasal spectra, particularly of /m/, exhibit "strongly idiosyncratic characteristics". Su concludes that the "coarticulation was found to give more reliable clues than the nasal spectrum alone."

• **Wolf** [11] reports on a study investigating the selection of parameters that are closely related to voice characteristics. The most useful parameters are found in:

- fundamental frequency,
- features of nasal and vowel spectra,
- estimates of the glottal source spectrum slope,
- word duration, and
- voice onset time.

• **Sambur** [12] reports on a study undertaken to determine a set of acoustic features in the speech signal that are effective for the identification of a speaker. In all, ninety-two features are examined using a probability of error approach. The speech data used consists of CVC (consonant-vowel-consonant) utterances containing the vowels /æ/, /I/, /i/, /u/ and the consonants /n/, /m/, /s/ and /ʃ/. The features found to be the most useful are:

- the second resonance in /n/ (around 1000Hz),
- the third/fourth resonance in /m/ (around 1700 to 2000 Hz),
- the second, third and fourth formant frequencies in vowels, and
- the average fundamental frequency of a speaker.

Sambur points out that "the ordering is established in accordance with the measurements of a given group of speakers; the speech characteristics of another group may result in a different ordering of features." However the results afford "a general idea of what features are important in recognising an unknown speaker" Wolf's fundamental frequency claims are down rated somewhat by Sambur because of their variability from one recording session to another.

• **Goldstein** [13] examines formant trajectories in order to determine efficient parameters for speaker recognition. He points out that in American English the diphthongs and r-coloured sounds exhibit large dialect variations. Experimental results show that "features derived from these sounds, which presumably depend more upon

speaker habits than on vocal-tract anatomy, in fact show large individual differences." The features found to be most efficient in recognising speakers are:

- the minimum second formant in /ɑr/ (party),
- the maximum first formant in /ɑr/,
- the maximum second formant values in /o/ (load), /ɔI/ (boy), and
- the minimum third formant in /ɜ˞/ (bird).

The reason for features of /ar/ being so useful may be due to coarticulation effects between the /a/ and /r/. "One feature that was particularly uncorrelated with any of the others was the minimum second formant for /ar/. This feature gives an indication of how strongly a speaker's /a/ is affected by the adjacent /r/ and may also depend on the way he shapes his tongue blade for the retroflex /r/"

• **Paul** [14] presents the results of a study leading to the design of a speaker recognition system called SASIS (Semi-Automatic Speaker Identification System). The recognition approach adopted is to extract segments of speech from two separate speech utterances and compute a statistical measure indicating whether the two utterances are from the same or different speakers. The database used is of a considerable size, consisting of over 35,000 'phonetic-event tokens' recorded from over 250 speakers. To measure similarity within each phonetic category, a weighted Euclidean distance measure is used. These individual distances are combined to form an overall measure between the two utterances using a 'desensitized Fisher discriminant',[14].

Speaker identification is carried out on the basis of thirteen phonemes only. "Vowels were selected in nonnasalised, stressed positions since they have been shown to have good discriminating ability and are easily labeled. Fricatives and stop releases were not selected because they show less intraspeaker variability and are difficult to analyze. Diphthongs and glides were not selected due to the lack of steady-state intervals for analysis." The phonemes are compared and ranked in terms of their speaker discriminating properties. Paul concludes that front vowels, high vowels and nasals appear to possess the greater speaker discriminating properties.

• **Nolan** [15] presents a variety of acoustic and phonological evidence, suggesting the phonemes /r/ and /l/ as exhibiting fairly high inter-speaker variability. The selection of these phonemes for use in speaker recognition is further justified by their fairly high natural occurrence in spoken English. Further criteria are presented as indicating their suitability and these include their robustness in transmission.

Nolan concludes from experimentation that "spectral information from initial allophones of /l/ and /r/ has been shown to yield moderate identification rates on material recorded by a homogeneous group of untrained speakers.","It is likely then that, though of lower value than nasals, /l/ and /r/ are worth incorporating in a speaker identification scheme making use of segmental information."

The phoneme /l/ is found to show a greater extent of coarticulation than /r/ and "this underlay the markedly better performance of /l/ compared to /r/ when identification was based on coarticulatory distance between each speaker's consonants in front and back vowel environments." However Nolan points out that "such a method scarcely lends itself to practical identification, because of the large and highly sophisticated test, as well as reference corpus that it presupposes".

4 Preliminary Experiment

4.1 The System

The speaker recognition system used for this experiment is a transputer-based prototype designed for room-access control, which has been developed within the Speech Research Group at University College Swansea. A description of the hardware and early experimental results are presented in [16].

A VQ codebook based approach to speaker recognition, comprising both identification and verification, is adopted. In the identification stage the test features are compared with all of the codebooks corresponding to the set of ten users and the one yielding the minimum sum-of-minimum-distances, determines the identity of the speaker. In the verification stage this minimum distance is compared with the person-dependent threshold corresponding to this codebook, in order to determine whether the door should be unlocked or not.

Clearly there is a disadvantage to this approach, because in the verification stage an imposter's features are always compared with the codebook that yields the smallest distortion. Hence, compared with a verification system in which a user must provide a claimed identity, the probability of a false acceptance occurring is increased. However, this approach was adopted for its convenience of use: it is completely 'hands-free' and within 2 seconds of approaching the microphone (buried in the door frame), the door lock is activated. The system is observed to perform extremely well.

Some interesting features of the system are listed as follows:

1. the speaker models are adaptive. On every occasion that access is approved the relevant model is updated. Model performance improves as a function of usage and this encourages people to use the system regularly.
2. the thresholds are speaker-specific and dynamic, within an empirically set band. Every time access is approved the relevant threshold is modified. If the distortion measure lies well below the threshold then the threshold is reduced. On the other hand if the distortion value lies only just below the threshold then the threshold is increased marginally.
3. a codebook 'hit count' threshold is included in the decision process in order to give a degree of text-dependency to the system. The 'hit count' is simply the total number of different model centroids that are utilised in the pattern matching. For the correct password the average 'hit count' is fairly high,

whereas for noise (e.g. blowing into the microphone might give a low distortion) or the wrong text, the 'hit count' is very low.

4 the system is continuously monitoring the environment and makes a decision whenever the smoothed energy profile remains high for 1 to 1.5 seconds. Any other activity duration is rejected.

4.2 Approach

The literature survey suggests that nasals and vowels provide good performance and that unvoiced fricatives, such as /s/, provide poor performance. Based upon this, two passwords were devised, with the aim of demonstrating that nasals are indeed more useful than unvoiced fricatives for speaker recognition. The first password, 'meaning', was chosen to contain vowels and nasals only. The second password, 'sea-fish', was chosen to contain the same vowel sounds as the first password, but in combination with unvoiced fricatives, rather than nasals. Ten speakers used the system regularly over a period of months.

4.3 Results

The results are somewhat surprising. None of the users found 'meaning' to be a more satisfactory password than 'sea-fish'. In fact, the majority of users declared a clear preference for the password 'sea-fish'. Performance measures, recorded automatically by the system, support this finding, although the difference between the measures for the two passwords is small.

5 Conclusions

The initial phase has demonstrated the following:

- the system exhibits high user acceptability. It is very convenient, and therefore people are prepared to use it.
- the adaptation algorithm works well making the system more reliable, giving positive feedback to its use.
- the approach provides an excellent way of gathering statistics in a meaningful environment, with added flexibility over a static database.
- the unvoiced fricatives provide surprisingly good speaker recognition performance.

References

[1] H. Heuvel, B. Cranen, and T. Rietveld. Inter- and intra-speaker variability in dutch speech segments: towards an Analysis Framework. *Proceedings of the Tutorial and Research Workshop on Speaker Characterisation in Speech Technology, Edinburgh, June 1990.*

[2] M. Savic and S. Gupta. Variable Parameter Speaker Verification System Based on Hidden Markov Modeling. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1990.

[3] M. Savic and J. Sorensen. Phoneme Based Speaker Verification. *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1992.

[4] J. Eatock and J. S. Mason. Speaker-dependent speech classification in speaker recognition. *Proc. ESCA Tutorial and Research workshop on Speaker Characterization in Speech Technology,*, pages 94–97, June 1990.

[5] J. Eatock and J. S. Mason. Automatically focusing on good discriminating speech segments in speaker Recognition. *Proc ICSLP-90, Japan, Vol. 1*, pages 133–136, November 1990.

[6] U. Hofker. AUROS - automatic recognition of speakers by computers: phoneme ordering for Speaker Recognition. *Proc. 9th International Congress on Acoustics, Madrid*, pages 506–507, 1977.

[7] R. Kashyap. Speaker recognition from an unknown utterance and speaker-speech interaction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, No.6, December 1976.

[8] A. Broeders and A. Rietveld. Segmental marking as a cue in auditory voice identification of telephone speech'. *European Conference on Speech Communication and Technology*, European Conference on Speech Communication and Technology, Paris, 1989.

[9] J. Glenn and J. Kleiner. Speaker identification based on nasal phonation. *The Journal of the Acoustical Society of America*, Vol.43, No.2, 1968.

[10] L.-S. Su, K. Li, and K. Fu. Identification of speakers by use of nasal coarticulation. *Journal of the Acoustical Society of America*, Vol.56, No.6, December 1974.

[11] J. J. Wolf. Efficient acoustic parameters for speaker recognition. *J. Acoust. Soc. Am., Vol. 51*, pages pp.2044–2055, 1972.

[12] M. R. Sambur. Selection of acoustic features for speaker identification. *IEEE Trans. ASSP-23*, pages pp.176–182, 1975.

[13] U. Goldstein. Speaker-identifying features based on formant tracks. *1975.*

[14] J. Paul, A. Rabinowitz, J. Riganati, and J. Richardson. Development of analytical methods for a semi-automatic speaker identification system. *Proceedings of the 1975 Carnahan Conference on Crime Countermeasures, Lexington University of Kentucky.*

[15] F. Nolan. The phonetic bases of speaker recognition. *Ph.D. Thesis, Cambridge University*, 1983.

[16] J. Oglesby. Neural models for speaker recognition . *Ph.D. Thesis, University College Swansea*, 1991.

6 Acknowledgement

This work was supported partially by BT Labs.