PHONETIC VARIABILITY IN SPEAKER RECOGNITION FOR FORENSIC PURPOSES

J. Ingram, R. Prandolini & S. Ong
University of Queensland
Queensland University of Technology

ABSTRACT - An experiment is reported on the impact of phonetic control in the selection of acoustic segments for formant trajectory based speaker identification under forensic conditions.

INTRODUCTION

Speaker identification systems for forensic purposes are usually obliged to operate with noisy signals, where only the most robust acoustic parameters are preserved, and only selected passages retain sufficient phonetic information for systematic comparisons with reference speech samples. One strategy commonly used, is to isolate phonetically analysable segments in the target material for text-dependent comparisons with phonetically comparable reference material. In this way, an attempt is made to control for various sources of speaker-independent acoustic variation and to isolate speaker-dependent sources of variation such as, individual differences in vocal tract size and shape (cross sectional area function), source signal characteristics, individual articulatory style, and speech variety (accent or dialect).

Traditional the aural-visual method has been used for forensic speaker identification. Koenig (1986) has shown that from 2000 spectrographic voice comparisons there were, 16% identified, 19% eliminated (speak was not in the set), 0.53% false eliminations, 0.31% false identifications and 65% of the tests were inconclusive. While the aural-visual method uses any utterance, this paper addresses the selection of the segments based on their linguistic content. This will hopefully yield a speaker identification rate which is superior to the aural-visual method. A particular sound in context may be variously realised phonetically by different individuals, even within the same dialect group. Nolan (1983) found individual differences in connected speech processes or coarticulation effects involving resonant sounds (/l/ and /r/) in post-vocalic position that could be exploited in speaker recognition.

THE EXPERIMENT

Aim and method

The aim of the present experiment was to obtain information on the speaker-discrimination potential of a range of phonetic targets in Aust. English, and information on the contribution of speaker-variable coarticulation effects or connected speech processes. A short, three sentence passage was constructed, containing a wide range of different vocalic nuclei in Australian English, some of which, because of context, were more likely, and others less likely, to be subjected to coarticulation effects, and some of which, such as nasal segments, would be expected to be minimally affected by context, but have a high component of speaker-dependent acoustic variability. The passage used, indicating points of acoustic segmentation, is shown in Figure 1.

Speakers and speech materials

Fifteen speakers read the test passage on two occasions, separated by a period of one month. Recordings were made in a quiet, but not noise-free environment, comparable to typical 'good quality' forensic recording conditions. The subjects were all male, native speakers of Australian English, between 25 and 45 years of age.

```
     voice identification system
1.   vɔɪs  ədɛnəfəkeɪʃn   sɪstəm
      1    2 3       4 5    6  7
      *    x *       * x    x  x

     per sonalin forma tionac ces via spo  kenlangua geout put
2.   pɜ  snlɪn   fəmeɪ ʃnæk   sɛs vaɪəspoʊ kənlæŋwɪ  ʤaʊt  pʊt
      1   2       3     4      5   6     7  8           9    0
      *   *       *     x                                    x

     please acknowlege whois here
3.   pliz     əknɒlɛʤ  huɪz  hɪə
      1        2        3     4
      *        *        *     *
```

Figure 1. The reading passage, indicating segmentation

Analysis

The recorded passages were digitised with 12 bit resolution at 10 kHz sampling rate. The speech samples were manually segmented at points of clear acoustic segmentability, using a waveform editor. A 12th order LPC analysis with 256 point Hamming window was performed on the signals, and formant trajectories for F1, F2, and F3 were extracted using the formant tracker of ILS.

Speakers were compared with one another and themselves across the two testing occasions on the basis of formant trajectories F_1, F_2 and F_3. A simple difference measure was used to compute a dissimilarity index for pairwise comparisons of formant trajectories. For the comparisons, segments were aligned at onset. No time normalisation of the signals was performed. A logarithmic transformation was applied to the formant frequencies to equalise the contribution of F_1, F_2 and F_3 to the dissimilarity measure. Zeros, caused by null formant estimates or temporal misalignment of trajectory offsets were ignored and the dissimilarity measure was normalised for the number of comparisons per segment.

A dissimilarity matrix was generated for each of the 21 segments in the passage. The main diagonal of each matrix represented the difference between the formant trajectories for a given speaker on a particular segment across the two recording sessions. The off-diagonal elements, were a measure of the dissimilarity between speaker_i on test occasion_1 and speaker_j on occasion_2. A row of elements measured the dissimilarity between one speaker on test occasion_1 and speaker_j on occasion_2. A column of elements measured the dissimilarity between one speaker on test occasion_2 and speaker_i on occasion_1.

The degree of success in discriminating between speakers may be assessed by comparing the values on the main diagonal of the dissimilarity matrix with the relevant off-diagonal elements. The dissimilarity score for a given speaker over the two test occasions should be smaller than any of the pairwise comparisons of that speaker with all other speakers on occasion_1 (the rows) or occasion_2 (the columns). See Table 2, the Combined Dissimilarity Matrix for 9 segments, and the Discussion below.

RESULTS

Individual segments

The success of individual segments in discriminating between speakers on the basis of formant trajectories is indicated in Table 1. Cell values indicate the number comparisons on a given row of the relevant dissimilarity matrix, where the main diagonal value was smaller than an off-diagonal element - in other words, where a speaker's formant trajectories across the two occasions were closer to each other than another speaker's formant trajectories.

The maximum score of 15 could be obtained only if all pairwise row comparisons could be made. Zero entries were obtained where the formant tracker yielded no formant frequency estimates for a particular segment and speaker and thus could produce no comparisons.

The column Means and counts of Zero entries in Table 1 provide an index of the speaker discrimination power of particular segments. Nine segments yielded formant trajectory comparisons for all speakers. These segments also posessed the highest Mean scores for speaker discrimination. They have been marked with a star (*) in Figure 1, together with segment 3.4, which yielded 1 zero but had a high Mean discrimination score. Segments with poor discriminating power are flagged with a cross (x) in Figure 1 and those with intermediate discriminating power are unmarked.

Stressed syllable nuclei provided much better speaker discrimination than unstressed nuclei. Stressed nuclei early in the utterance or in initial position or appeared to perform better than those close to the end of the utterance. The short utterance (3), which was spoken more slowly by most speakers, performed better than the long utterance (2). Segments with high speaker discrimination were broadly distributed over the Australian English vowel space. Segments with poor discrimination power were those with low acoustic energy levels, insufficient for reliable extraction of formant frequencies.

Subject No	Sentence.segment ->																				
	11	12	13	14	15	16	17	21	22	23	24	25	26	27	28	29	20	31	32	33	34
01	04	11	14	15	00	06	06	15	15	13	02	15	08	12	12	12	00	15	13	14	15
02	10	00	14	05	00	00	00	14	15	09	09	11	15	10	11	14	00	15	12	14	14
03	13	00	01	15	00	05	05	15	03	02	15	00	00	00	00	00	00	15	12	11	14
04	15	00	15	14	00	08	08	15	13	13	14	14	14	07	07	04	07	13	10	15	06
06	06	10	14	10	08	10	10	13	13	15	02	15	00	14	14	14	01	08	09	14	15
09	15	11	12	14	03	09	09	15	09	12	13	13	15	12	14	04	05	15	05	02	15
10	12	08	13	15	00	00	00	00	00	00	00	00	00	00	00	00	00	14	15	02	12
11	15	00	15	15	00	10	10	10	06	06	13	15	09	12	14	11	00	15	02	15	11
12	09	11	13	15	00	00	00	15	06	14	14	15	15	08	01	14	07	15	10	08	15
13	10	00	14	14	00	10	10	15	14	11	00	13	15	06	14	10	00	15	10	09	14
14	11	11	15	09	07	00	00	15	15	11	15	11	15	13	09	13	00	11	07	15	13
15	14	00	03	14	08	00	00	08	11	15	13	06	15	13	13	12	00	14	12	15	14
16	14	11	14	09	00	01	10	14	11	14	00	15	11	02	14	14	04	15	15	15	15
18	15	00	13	14	00	00	00	06	13	15	00	08	15	12	11	14	00	13	15	08	00
20	15	11	10	14	00	00	00	15	13	11	11	12	15	14	14	14	04	15	10	14	05
Means	12	06	12	13	00	04	05	12	10	11	08	11	11	09	10	10	02	14	10	11	12
Zeros	00	07	00	00	10	07	06	00	00	00	04	01	02	02	02	02	08	00	00	00	01

Table 1. Speaker discrimination scores

Combined Segments

The speaker discrimination power of the 9 segments which yielded pairwise comparisons of formant trajectories for all speakers is shown in Table 2. Stars (*) following rows and columns indicate speakers for whom the intra-subject dissimilarity scores (the main diagonal) are smaller than all of the row or column inter-subject dissimilarity scores (the off-diagonal elements).

In those cases where the main diagonal is not the row or column minimum value, its rank is indicated. One subject (s10) was discarded in the course of constructing the group dissimilarity matrix, because of an absence of formant measurements due to misreading of the target sentence. The combined segments yielded a high rate of speaker discrimination, with only 3 of 28 tests failing to absolutely discriminate a target speaker from all others.

The effect of phonetic targets

To estimate the effect of controlling for the phonetic identity of segments (text dependency), the above analysis using the same 9 segments was repeated, but with the phonetic identity of the segments randomised within speakers. In other words, we simulated the effect of text-independency. The resulting Combined Dissimilarity Matrix is shown in Table 3.

Speaker Number ->

1	2	3	4	6	9	11	12	13	14	15	16	18	20	
1.26	1.86	3.45	2.39	2.98	2.20	2.75	2.33	1.42	2.24	2.77	2.66	1.96	2.70	*
2.32	**1.73**	3.57	2.22	3.09	2.41	2.69	2.42	2.44	2.62	2.78	2.40	2.18	2.96	*
2.27	2.71	**3.05**	2.85	3.27	2.57	3.12	2.18	2.47	2.67	3.15	3.02	2.30	3.52	9
2.04	1.86	3.45	**1.22**	2.77	2.11	1.95	2.04	1.98	2.64	2.34	1.73	1.55	2.20	*
3.21	2.89	3.60	2.79	**2.11**	2.62	3.02	2.85	2.81	2.88	2.99	2.88	3.44	3.10	*
1.86	1.67	3.14	1.48	2.31	**1.40**	1.74	1.87	1.82	2.36	2.24	1.53	1.63	1.89	*
1.79	1.06	3.50	1.57	2.73	1.71	**1.47**	1.73	1.59	2.05	2.50	1.73	1.66	1.65	2
2.16	2.06	3.24	2.22	2.47	1.70	2.43	**1.56**	1.92	2.19	2.60	1.92	2.02	2.46	*
1.61	1.51	3.26	1.95	2.60	1.94	1.89	2.39	**1.34**	2.19	2.80	2.33	1.47	2.27	*
2.21	1.89	3.73	2.48	2.84	2.42	2.44	2.09	1.85	**1.48**	3.14	2.76	2.23	2.47	*
2.36	2.28	3.09	2.38	2.51	2.37	2.53	2.32	2.27	2.83	**1.60**	2.27	2.30	2.85	*
2.45	2.12	3.40	1.90	2.22	1.93	1.87	1.96	2.27	2.84	2.11	**1.05**	2.14	2.48	*
1.88	1.74	3.39	1.83	2.83	2.16	2.00	2.06	2.00	2.50	2.33	1.93	**1.32**	2.15	*
2.27	1.82	3.55	2.27	3.09	2.18	1.85	2.35	2.04	2.53	2.84	2.59	1.80	**1.41**	*
*	4	*	*	*	*	*	*	*	*	*	*	*	*	

Table 2. Combined Dissimilarity Matrix
(Phonetically controlled)

Speaker Number ->

1	2	3	4	6	9	11	12	13	14	15	16	18	20	
3.55	4.05	4.54	3.31	3.36	3.63	3.60	3.73	3.01	3.74	3.01	3.41	2.91	3.50	8
4.76	**4.73**	5.05	3.87	4.59	4.05	4.29	4.35	3.76	4.14	4.18	4.01	3.91	3.94	12
4.60	4.36	**5.46**	4.06	4.30	4.25	3.87	4.64	3.54	4.32	3.99	4.27	3.77	4.00	14
4.09	4.00	5.23	**3.38**	3.83	3.62	3.18	4.07	3.12	3.43	3.49	3.51	2.92	2.90	5
4.15	4.50	5.91	4.34	**3.82**	4.41	4.19	3.95	4.26	4.59	4.26	4.32	4.46	4.33	*
3.86	3.94	5.32	3.42	3.82	**3.54**	3.08	3.64	2.99	3.41	3.60	3.42	3.16	3.09	7
4.18	3.91	5.08	3.39	3.56	3.57	**3.13**	3.84	2.98	3.43	3.52	3.34	2.72	3.09	4
4.42	4.59	5.38	3.89	4.34	4.00	3.74	**4.12**	3.48	3.97	4.03	3.88	3.35	3.38	10
4.32	4.24	5.06	3.81	3.98	3.73	3.55	4.29	**3.20**	3.95	3.95	3.50	3.46	3.47	*
4.78	5.01	5.88	4.29	4.25	4.44	4.07	4.37	3.73	**3.88**	4.18	3.90	3.85	3.76	4
4.30	4.33	5.70	3.94	3.82	4.23	4.26	3.65	4.12	4.32	**3.78**	3.76	3.96	4.23	3
4.05	3.96	5.46	3.66	3.85	3.75	3.42	3.80	3.23	4.03	3.70	**3.61**	3.29	3.27	6
3.70	3.95	5.38	3.36	3.48	3.42	3.12	3.84	3.04	3.64	3.19	3.54	**2.78**	2.88	*
4.02	4.17	5.67	3.44	3.39	3.55	3.14	3.80	3.29	3.41	3.58	3.26	3.29	**3.02**	*
*	13	10	3	7	3	3	10	4	7	8	8	2	3	

Table 3. Combined Dissimilarity Matrix
(Phonetically Uncontrolled)

While speaker discrimination for the phonetically uncontrolled formant trajectory matching achieved results that were better than chance, the accuracy of discrimination was very much reduced. Only 5 of the 28 tests achieved absolute speaker discrimination.

CONCLUSIONS

The experiment indicated that phonetic control of the segments analysed yielded high speaker discrimination rates for small amounts of speech data in comparison with what is required for text independent speaker identification.

This finding is of practical significance for forensic applications where the amount of recorded speech material is often small and of marginal quality.

The ILS formant tracker frequently failed to yield formant estimates for acoustically marginal speech segments. However, for those segments where formant trajectories were calculated, the parameters appear to yield robust results. Clearly, further refinement of the parameter extraction methods is indicated.

Further investigation of subcomponents of the phonetic variability which clearly contributed to the speaker discrimination, such as that which is attributable to coarticulation effects versus the phonemic identity of the segments themselves is currently underway.

REFERENCES

Koenig, B. E. (1986) Spectrographic voice identification: a forensic study. J.Acous. Soc. Am., 79(6), pp2088-2090, June, 1986.

Nolan, F. J. (1983) The phonetic basis of speaker recognition Cambridge (UK): Cambridge University Press.