Variation in Long-Term Fundamental Frequency: Measurements From Vocalic Segments in Twins' Speech

Deborah Loakes
School of Languages and Linguistics
The University of Melbourne
Australia
dloakes@unimelb.edu.au

Abstract

This paper is a preliminary analysis of suprasegmental aspects of twins' speech. The focus of analysis is variation in long-term fundamental frequency, and reference measurements are presented for read and spontaneous speech, across non-contemporaneous samples. Measurements are discussed across the group of speakers, between twin pairs, and within individual speakers. Results show that speakers tend to fall within a specific F0 range, and that twins tend to have a more similar mean long-term F0 than what has previously been reported for unrelated pairs of speakers. Results also show that while there are between-speaker differences, within-speaker variation is also present; some speakers have considerable variation, while others have almost none. Certain methodological concerns are also addressed.

1. Introduction and Background

1.1. Long-term F0

Acoustic analysis of speech samples involves comparison of long- and short-term features. Short-term features are based on analysis of individual segments, such as formant patterns of a particular vowel or coarticulatory phenomena between specific segments, while long-term features are calculated over larger stretches of speech (Nolan 1983, Rose 2003). The focus of this paper, fundamental frequency sampled at various points throughout a speech sample, is categorised as a long-term feature.

Fundamental frequency (hereafter F0) reflects the rate of vibration of the vocal folds, and to this end plays a part in reflecting speaker-specific behaviour. In addition, F0 is easy to measure and readily available even in short samples of samples of speech, and for these reasons, it is often measured in forensic speaker identification case-work and experiments (see the discussion in Rose 2002:244-248). However, F0 is a feature which is also subject to a relatively large amount of variation (see section 1.4), and this is the focus of the current paper.

1.2. Existing F0 Measurements

Research on long-term F0 has resulted in a number of published values that can be used as reference data, especially for forensic speaker identification tasks. Values for both male and female speakers, across a number of different languages and in both read and spontaneous speech, are summarised in Tranmüller and Eriksson (1995). Mean F0 values for read speech produced by male and female speakers, which are used as a reference for forensic speaker identification in Europe, are reported in Künzel (1989). Lindh (2006) presents long-term F0 values taken from short (up to two minutes) samples of spontaneous speech for 109 young male speakers of Swedish, Kinoshita (1998) reports long-term F0 measurements in non-contemporaneous samples of both read and spontaneous speech for a male speaker of Japanese, and Kinoshita (2005) reports both individual speaker and group measurements for male speakers of Japanese. For Australian English, Rose (2003) reports individual long-term F0 measurements for non-contemporaneous read speech produced by six male speakers.

Most studies report average F0 values of a group of speakers (ignoring individual measurements), and most groups of speakers have a relatively wide age range. For example, the study analysing participants with the narrowest age range is Lindh (2006), whose speakers were aged between 20 and 30. Age range is expected to affect group measurements of long-term F0 in the sense that, for male speakers, F0 decreases with age between puberty and approximately 35 years of age (Tranmüller and Eriksson 1995).

1.3. Variation in F0

While F0 can be useful for measuring between-speaker variation, it is also varies within speakers. This is caused by factors such as intoxication, emotion and disguise, or even the time of day (Braun 1995; see Zetterholm 2003 for further discussion of disguise). In terms of normal variation, without

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

extraneous factors such as disguise or intoxication, Rose (2003) reports mean and modal long-term F0 vary around 6-7 Hz within-speakers, and 11-12 Hz between-speakers of Australian English.

1.4. Individual Identifying Power of F0

Long-term F0 has been found to have some speaker-characterising value. As discussed above, Rose (2003) has shown that between-speaker variation is greater than within-speaker variation for a group of Australian English speakers. In terms of discriminating speakers (using a likelihood ratio approach), in a separate study Rose (2002) analysed Cantonese speech samples and found long-term F0 somewhat useful. Kinoshita (2005) analysed Japanese speakers, and found a very small discriminating potential.

These results point to the fact that long-term F0, in combination with other parameters, can *assist* in determining speaker identity. More usefully, long-term F0 measurements can show whether a speaker falls outside the "average" with an unnaturally high or low long-term F0. Average F0 values can also lead to eliminating suspects, if, for example, values are distributed around a particularly high mean in a forensic speech sample, and particularly low mean F0 in a suspect speech sample.

In an exploration of the speaker-specificity of long-term F0, Braun (1995) concludes with a number of methodological points that are relevant to the present study. She notes that analyses should include read and spontaneous speech, and that rather than analysing only average values, consideration should also be given to the range of variation in the data.

1.5. Twins' Speech

In the present study, twins' speech is analysed because where short-term differences are concerned, twins have been shown to be harder to discriminate, and to have fewer overall differences, than unrelated pairs of speakers (see Loakes 2006). In this study, measuring F0 in twins' speech will give an impression of the degree of variation that can be expected for a long-term parameter between pairs of speakers who are as similar as possible.

1.6. Aim

This paper is a preliminary investigation into suprasegmental aspects of twins' speech. More specifically, the aim is to contribute to knowledge on variation in long-term F0 by:

1. providing long-term F0 measurements from samples of speech produced by young adult male speakers of Australian English (four twin pairs),
2. assessing the amount of within- and between-speaker variation in long-term F0, across different types of non-contemporaneous speech samples (read and spontaneous).

2. Method

2.1. Data and Participants

The data consists of speech produced by eight male speakers of Australian English (three identical twin pairs and one non-identical twin pair), aged between 18 and 20. The speakers are all university students enrolled in different courses, although each speaker shared the same education as his twin until the end of high school (to 17 years of age). The speakers in this investigation are TbY & TfY, PF & CF and LG & RG (identical twin pairs), and RH & ZH (non-identical twin pairs).

Each speaker took part in two Labovian-style interviews with the author, separated by a period of around six months. Approximately eight minutes of spontaneous conversational speech from each interview is analysed in the present study and these samples, consisting of monologic speech, are the *spontaneous speech* samples. Before each of these interviews, the speakers read a wordlist consisting of focus items in /hVd/ context, as well as a number of foils, and these recordings make up the *read speech*. For each of the eight speakers, four samples of speech were analysed (two spontaneous, two read).

The data was originally collected for a project exploring individual variation in the speech patterns of identical and non-identical twin pairs (Loakes 2006). The speakers were chosen from a wider corpus on the basis of auditory similarity, which was a requirement of the original project.

2.2. Recording and Analysis

The participants were recorded in the phonetics laboratory at the University of Melbourne, between August 2002 and March 2003. The recordings were made on 120 minute Sony Digital Audio Tapes using a Sony ECM-999 PR electret condenser stereo microphone, positioned approximately 80mm from the participant's mouth (the same distance in each case) and set to 120°, using a studio quality rack mount Tascam DA-30 DAT recorder.

The data were digitised with *ESPS/Xwaves+*, and consonant and vowel segments were labelled manually with *The EMU Speech Database System*, version 1.4.1 (cf. Cassidy and Harrington, 2001). F0 can be measured in a number of different ways, and in this study it was extracted from the midpoint of all labelled vowel tokens using *EMU* version 1.8 and *R* version 1.7.1.

Outliers, where the pitch was doubled or halved (common for schwa, and in some especially creaky segments) were excluded from analysis. Outliers due to intonational variation were included in the analysis.

It should be noted while the spontaneous speech samples were all approximately eight minutes in duration, the number of tokens sampled varied greatly (according to both speech rate and the number and type of lexical items used by the participants). The number of vowel tokens sampled in the read speech was consistent, between 27 and 32, depending

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

on the pronunciation of the 21 lexical items (see Loakes 2006 for specific details about the participant tasks). In addition, it should be noted that the read speech samples were longer than the recommended minimum of 15-20 seconds for a "normal" (unemotional) communicative style (Braun 1995). The entire duration of the read speech samples in this study ranged between 17.1 seconds to 31.4 seconds.

In this study, mean, median, mode and standard deviation of long-term F0 were recorded for each sample. Most researchers report mean and standard deviation F0 measurements; however Lindh (2006) notes that the median should also be included when reporting results, and Rose (2002, 2003) discusses the importance of modal F0.

3. Results and Discussion

3.1. Population Measurements

To give a general overview of long-term F0 in the speech of the participants, F0 measurements for the population are provided first. The population includes the eight speakers (four twin pairs), and also includes non-contemporaneous samples from each speaker.

The mean, median, mode and standard deviation are presented in Table 1 below.

measurement	spontaneous	read
mean	105.2	109.2
median	103.1	107.4
mode	107	105
sd	16.4	12.6

Table 1: Mean, median, mode and standard deviation of F0 (Hz) across the population spontaneous and read speech samples.

First, it should be noted that the mean values are somewhat lower than the mean long-term F0 reported in many other studies based on English speaking European males. For example, the values here are lower than those summarised in Tranmüller and Eriksson (1995), apart from one study by Johns-Lewis (1986) which assessed English-speaking males' conversational speech (with an average 101 Hz). The lower overall mean values in this study were not expected, given that younger male speakers (between puberty and 35) are said to have a higher F0 (see section 1.3 above).

Comparing the values for read and spontaneous speech in this study, mean and median values are around 2 Hz higher in read speech compared with the spontaneous speech, while the mode is around 2 Hz lower. The standard deviation is greater in the spontaneous speech, which is not surprising given that greater excursions in pitch are expected in natural speech (see e.g. values reported in Braun 1995, Kinoshita 1998). A t-test shows significant differences between the means of the read and spontaneous speech ($p < 0.01$ using a two-tailed paired t-test).

The values presented in Table 1 could be used as a starting point for a reference sample of long-term F0 in the speech of twin pairs, and could also inform as to the general long-term F0 of young male speakers of Australian English within a narrow age range (18-20). In addition, the results in Table 1 also give an idea of the degree of difference between read and spontaneous speech across a group of speakers.

While the results discussed above give a picture of long-term F0 across the population, it is also important to inspect the distribution of the data. This is shown, for the spontaneous speech samples, in Figure 1 below.

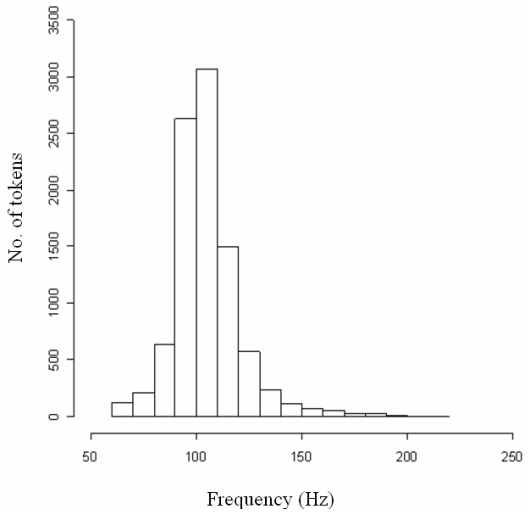

Figure 1: Distribution of F0 across the eight speakers' spontaneous speech samples.

First, it should be noted that the distribution shows positive skewing, which is typical for long-term F0 (see e.g. Lindh 2006, Rose 2002; 2003, Tranmüller & Eriksson 1995).

This figure shows that the majority of tokens sampled fall between 100-109 Hz, with a relatively large number of tokens falling between 90-99 Hz, and 110-119 Hz. In terms of the range of data, the lowest observed F0 across the population was 60.6 Hz, while the highest was 215.5 Hz. Values at the lower end of the distribution (between 60-79 Hz) were observed relatively frequently, while values at the higher end (over 150 Hz) occurred far less often. It is likely that these higher F0 values were sampled from especially emphatic speech and rising tunes.

Turning now to the read speech, distribution of F0 across the eight speakers is shown in Figure 2 below.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

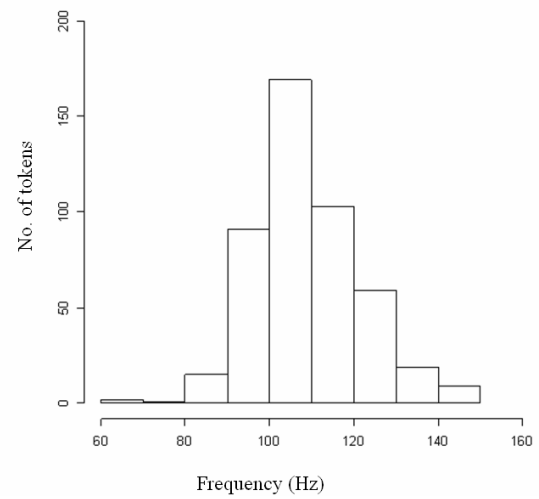

Figure 2: Distribution of F0 across the eight speakers' read speech samples.

This figure shows that in the read speech, the majority of tokens lay between 100-109 Hz, as for the spontaneous speech. The read speech also shows positive skewing.

A major difference between the read speech and the spontaneous speech is, aside from the number of tokens, the F0 range. The spontaneous speech distribution shows values up to 215.5 Hz, whereas the highest sampled F0 in the read speech was 149.3 Hz. These observations support the suggestion made above that the higher values in the spontaneous recordings are likely to have been sampled from emphatic speech and rising tunes which are absent from the read speech. Further, while the means are relatively similar, the difference in F0 distribution across the read and spontaneous speech suggests that read speech is an inadequate reflection of natural speaker behaviour.

3.2. Individual Measurements

The mean, median, modal and standard deviation for long-term F0 in each of the eight speaker's read and spontaneous speech samples are presented in Table 2. These measurements are discussed in sections 3.3 and 3.4.

	s 1	s2	r 1	r 2	av.
TbY mean	96.2	98.5	99.8	99.4	98.5
TbY median	93.2	100.5	95.9	99.5	97.3
TbY mode	91	97	94	93	93.7
TbY sd	19.7	16.7	10.4	7.4	
TfY mean	105.7	105.9	114	112.6	109.5
TfY median	104.2	101.6	111.5	108.3	106.4
TfY mode	98	100	110	110	104.5
TfY sd	12.8	18.2	10.5	11.2	
PF mean	105.1	107.9	106.2	114.7	108.5
PF median	103	106.5	106.5	112.8	107.2
PF mode	103	104	109	111	106.7
PF sd	13.4	12.8	7.4	6.3	
CF mean	100.28	102.35	96.8	104.5	100.9
CF median	98.3	100.2	96.7	105.7	100.2
CF mode	96	96	96	105	98.3
CF sd	13	16.4	8	12.7	
LG mean	113.1	99	125.7	113.2	112.7
LG median	110.5	102.7	125.7	111	112.5
LG mode	111	95	125	108	109.8
LG sd	15.9	15.8	6.8	9.9	
RG mean	120.7	114.8	129.5	119.8	121.2
RG median	119.6	110.5	127.3	116.4	118.5
RG mode	116	110	126	115	116.8
RG sd	14.2	19.5	8.4	9.5	
RH mean	101	102.2	94.9	104.5	100.7
RH median	96.4	99.5	94.2	105	98.8
RH mode	94	96	96	105	97.8
RH sd	15.7	16.1	7.5	5.9	
ZH mean	105.5	105.8	105.1	106.4	105.7
ZH median	103.6	103.5	104.6	106.1	104.5
ZH mode	97	100	103	106	101.5
ZH sd	16.2	11	5.2	4.3	

Table 2: Mean, median, mode and standard deviation of F0 (Hz) across the eight speakers' spontaneous (s) and read (r) speech samples. Average values (av.) are also shown.

3.3. Within-speaker variation

In general, results in Table 2 show that F0 values are relatively consistent for some speakers, and extremely variable for others. For example, across TbY's speech samples there is very little variation in mean long-term F0, however there is a relatively wide variation across his median and modal long-term F0. There is also very little variation across ZH's mean long-term F0, very little variation across his median F0 values, but a wide variation in modal F0. For a number of other speakers, mean F0 values vary.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

For some speakers, the variation in long-term F0 is seen in only one of the four speech samples. For example, three of PF's speech samples (s1, s2 and r1) fall within a narrow range, while the fourth (r1) is higher. A similar situation is seen for RH, for whom three of four speech samples (s1, s2, r2) were within a narrow range, while the fourth (r1) is lower. For other speakers, the variation occurs across a number of speech samples. For example, LG, who appears to have the greatest amount of variation in mean long-term F0, has two speech samples (s1 and r2) which have almost the same mean F0. This contrasts with a very low mean F0 for the second spontaneous speech recording, and a very high mean F0 for the first read speech sample. Here there is extreme variation for one speaker; only 0.1 Hz difference between one of the spontaneous and read speech samples, and a 26.7 Hz difference between the others. While this within-speaker variation exists in LG's speech, there is also a level of consistency. That is, LG has an overall higher long-term mean F0 in the first recording session compared with the second recording session.

While within-speaker variation across read and spontaneous speech is apparent amongst the speech samples, this variation is significant only for LG (p <0.01 using a two-tailed paired t-test).

3.3.1. A note on standard deviation

An important finding is that standard deviation measurements from read speech reported by Rose (2003) and Künzel (1989), and for spontaneous speech reported by Kinoshita (2005) and Lindh (2006), are consistently higher than the standard deviation values for speakers in this study. This is probably due to the fact that here, F0 was measured from the mid-point of vowels only. In contrast, Rose and Kinoshita sampled F0 at designated time steps throughout the analysis (and so included all voiced segments, not just vocalic tokens). Exact measures by Künzel (1989) and Lindh (2006) are not clear.

The method used to analyse F0 in this study, (i.e. from same-segment tokens) is thus a useful way of measuring a speaker's F0, because less within-speaker variation is observed.

3.3.2. Between-speaker variation across the population

Table 2 shows that F0 values for speakers tend to fall within a particular range. Considering the mean F0 across the population (105.2 in spontaneous speech and 109.2 in read speech) it can be seen that speakers tend to have average F0 values which are distributed either side of these group averages. For example, TbY's average F0 values are lowest across the corpus, and always fall under the group average. In contrast, RG's long-term F0, while variable, is consistently above the mean. The only exception to this is LG, for whom significant within-speaker differences were observed (see above).

3.3.3. Between-speaker variation across twin pairs

Assessing the average F0 values reported in Table 2, it can be seen that twin pairs do not necessarily have a similar long-term F0, although equally, they are not on average widely different. There is 11 Hz difference between TbY and TfY's mean long-term F0, 7.6 Hz between PF and CF, 8.5 Hz between LG and RG, and 5 Hz between the non-identical twins RH and ZH. The 11 Hz difference between PF and CF is standard for male speakers of Australian English (Rose 2003, and see section 1.4 above), whereas the difference between the other twin pairs is smaller. That is, three of the twin pairs have a closer mean F0 than the average reported between-speaker differences for unrelated similar-sounding speakers.

If speakers are listed in order of the lowest to highest observed mean F0 – *TbY, RH, CF, ZH, PF, TfY, LG, RG* – it can be seen that twin pairs do not necessarily have the closest mean F0 values. For example, TbY has a closer mean F0 to RH and CF than to his twin. Two-tailed paired t-tests comparing mean F0 values between twin pairs showed that differences are significant between TbY & TfY's read speech (p <0.02) and between PF & CF's read (p <0.03) and spontaneous (p <0.04) speech. No other significant differences were observed. However, it should also be noted that the average mean is not the only measurement that should be assessed: while RH and ZH have similar mean values, RH tends to have a lower median and modal F0. This supports suggestions that median (Lindh 2006) and modal F0 (Rose 2003) should also be incorporated into analysis of long-term F0. Mean values do not necessarily reflect important between-speaker variation amongst twin pairs.

4. Conclusions

4.1. Reflection on Research Aims

This paper has contributed to knowledge on variation in long-term F0, by reporting average measurements for both read and spontaneous speech, across non-contemporaneous speech samples provided by a group of eight speakers.

The aims of the paper were two-fold. The first aim was to provide long-term F0 measurements from samples of speech produced by young adult male speakers of Australian English, and these are reported in sections 3.1 and 3.2. The second aim was to assess the amount of within- and between-speaker variation in long-term F0, and this was addressed in section 3.3.

The study has shown that long-term F0 is not especially variable within-speakers, even across non-contemporaneous data sets which have been recorded in the same manner. However, this is not categorical; one speaker in this study (LG) had significant variation across his speech samples. Additionally, while somewhat variable between-speakers, F0 is not markedly different across the population of speakers (which agrees well with other studies discussed in 1.4 above). While this study is not suggesting that long-

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

term F0 can be used to discriminate twin speakers, measurements such as provided in the current study are still a useful guide for forensic speaker identification. The measurements show what can be expected in long-term F0 measurements for a given population, within a pair of similar-sounding (closely related) speakers, and also across one speaker in different communicative situations.

It should be noted that the twin pairs in this study have all been classed as similar-sounding speakers, and the range of values presented in Table 2 suggest that for most twin pairs, long-term F0 might well be a contributing factor as far as their auditory similarity is concerned.

4.2. Other implications of the analysis

In addressing the research aims, some other important findings arose from the analysis. Firstly, the difference in distribution across the read and spontaneous speech showed that read speech is an inadequate reflection of speaker behaviour. Read speech is simply not reflective of the range of intonational variation that occurs in conversational style speech. However, for forensic speaker identification purposes, comparison of read and spontaneous speech is not necessarily problematic; in this study significant differences were only found across one of the speakers' speech samples.

Additionally, an important methodological point arose from the analysis. It was found that when analysing long-term F0, some control for segment type is beneficial. In this study, only vocalic segments were analysed, and compared with other studies the results showed less within-speaker variation. As such, future forensic speaker identification experiments and case-work analysing long-term F0 should confine the analysis to same-segment tokens in order to better reflect speaker-characteristics.

4.3. Future Direction

The degree of F0 variation within- and between-speakers is a more complex matter than simply reporting average values. In order to further explore suprasegmental aspects of twins' speech, a more detailed analysis of the range of F0 variation within-speakers, and an analysis of intonational pitch events between twin pairs, is also planned.

5. Acknowledgements

Thanks to the Australian Twin Registry for assistance with data collection, and thanks to Associate Professor Janet Fletcher and Mary Stevens for discussion of ideas and results.

6. References

Braun, A. (1995) Fundamental frequency – How speaker-specific is it? In A. Braun and J.P. Köster (eds.) *Studies in Forensic Phonetics* 9-23, Trier: Wissenschaftlicher Verlag.

Cassidy, S. and J. Harrington (2001) Multi-level annotation in the EMU speech database management system *Speech Communication* 33, 61-77.

Johns-Lewis, C. (1986) Prosodic differentiation of discourse modes in C. Johns-Lewis (ed.) *Intonation in Discourse* Croom Helm: London, pp. 199-219.

Kinoshita, Y. (1998) Japanese forensic phonetics: Non-contemporaneous within-speaker variation in natural and read-out speech in R. Mannell and J. Robert-Ribes (eds.) *Proceedings of the 5th International Conference on Spoken Language Processing*, Australian Speech Science and Technology Association, Canberra, 145-148.

Kinoshita, Y. (2005) Does Lindley's LR estimation Formula Work for Speech Data? Investigation Using Long-term F0 *International Journal of Speech, Language and The Law* 12 (2), 235-254.

Künzel, H.J. (1989) 'How Well Does Average Fundamental Frequency Correlate with Speaker Height and Weight?' *Phonetica* 46, 117-125.

Lindh, J. (2006) 'Preliminary Descriptive F0-statistics for Young Male Speakers' *Lund University Working Papers* 52, 89-92.

Loakes, D. (2006) *A Forensic Phonetic Investigation into the Speech Patterns of Identical and Non-Identical Twins* PhD Thesis, School of Languages and Linguistics: The University of Melbourne.

Nolan, F. (1983) *The Phonetic Bases of Speaker Recognition* Cambridge: Cambridge University Press.

Rose (2002) *Forensic Speaker Identification* London: Taylor and Francis.

Rose (2003) 'The Technical Comparison of Forensic Voice Samples' In I.S. Freckleton and H. Selby (eds.) *Expert Evidence* Lawbook Co: North Ryde, Ch.99.

Tranmüller, H. and A. Eriksson (1995) *The Frequency Range of the Voice Fundamental in the Speech of Male and Female Adults* (unpublished manuscript) Retrieved August 8, 2006 from: http://www.ling.su.se/staff/hartmut/aktupub.htm

Zetterholm, E. (2003) 'The Same but Different – Three Impersonators Imitate the Same Target Voices' in. M.J. Sole, D. Recasens and J. Romero (eds.) *Proceedings of the 15th International Congress of Phonetic Sciences* Barcelona, Spain, 2205-2208.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review