Articulation Rate as a Speaker Discriminant in British English

Erica Gold

Department of Linguistics and Modern Languages, University of Huddersfield, United Kingdom
e.gold@hud.ac.uk

Abstract

Identifying speech parameters that have both a low level of intra-speaker variability and a high level of inter-speaker variability is key when discriminating between individuals in forensic speaker comparison cases. A substantial amount of research in the field of forensic phonetics has been devoted to identifying highly discriminant speaker parameters. To this end, the vast majority of the existing literature has focused solely on vowels and constants. However, the discriminant power of speaking tempo has yet to be examined, despite its broad use in practice and it having been recognized.

This paper examines, for the first time, the discriminant power of articulation rate (AR) in British English. Approximately 3000 local ARs were measured in this study for 100 Southern Standard British English male speakers. In order to assess the evidential value of AR, likelihood ratios were calculated. The results suggest that AR performs well for same speaker comparisons. However, for different speaker comparisons, the system is performing just worse than chance. Overall, it appears that AR may not be the best speaker discriminant, although it is important to still consider AR in forensic speaker comparisons as there may be some individuals for which AR is highly idiosyncratic.

Index Terms: articulation rate, speaking tempo, forensic speaker comparison, forensic phonetics, likelihood ratios

1. Introduction

In forensic speaker comparison (FSC) casework an expert analyzes a range of phonetic and linguistic variables (e.g. vowels, consonants, lexical choices) in order to compare speech in the criminal and suspect recordings. An expert's role is to provide the trier(s) of fact with an opinion regarding the probability of obtaining the speech evidence (the similarities/differences between the criminal and suspect samples) under the hypothesis that the samples came from the same person, versus the probability of obtaining the evidence (the typicality of the analysed speech parameters) under the hypothesis that two different speakers produced the criminal and suspect samples.

In an ideal world, FSC casework would be simple insofar as an expert would only need to analyze a single phonetic parameter in order to arrive at a final conclusion. If a single parameter was to constitute an entire analysis, that said parameter would have to be so idiosyncratic in individuals' speech that no two people in the entire world shared the same realization. Unfortunately, this is not the case, and rather experts advocate for the consideration of many parameters in conjunction with one another in order to arrive at a final conclusion [1]. As a result, forensic phoneticians have devoted a lot of research into looking for good speaker discriminants to use in combination with other phonetic and linguistic parameters. However, the vast majority of speaker discriminant research is focused on segmental parameters [2-7].

Although there is no previous literature that has quantified the discriminant power of speaking tempo, 93% of forensic experts reported analyzing speech tempo in FSCs [1]. Furthermore, 20% of experts found speech tempo to be most useful in their own casework for discriminating speakers, ranking it as the third most helpful parameter overall out of all possible parameters used for analysis in casework. Therefore, it is important to examine the discriminant power of speaking tempo in order to understand the ability speaking tempo may have in differentiating between individuals. This will also allow forensic phoneticians to properly evaluate the strength of evidence that may be associated with speaking tempo. In addition to the forensic phonetics community, those in the automatic speaker recognition (ASR) community may find the results of interest as ASR systems do not traditionally capture information related to a speaker's tempo [8]. Those working at the interface between forensic phonetics and ASR may indeed consider speaking tempo in addition to automatic results.

2. Background

In phonetics, speech tempo is typically captured through one of two ways: speaking rate or articulation rate. Speaking rate (SR) measures the rate of speech over an entire speaking-turn. It includes all speech material, both linguistic and non-linguistic, in addition to silent pauses that are contained across the overall speaking-turn [9]. Articulation rate (AR) is the rate at which a given utterance is produced. Articulation rate of speech material therefore excludes silent pauses given the definition of an utterance, which "begins and ends with silence" [9]. The difference between the two measures is that speaking rate captures disfluencies and filled/unfilled pauses in the calculation, whereas articulation rate is intended to present a rate independent of disfluencies and unfilled pauses. Within the field of forensic speech science the majority of experts find articulation rate more helpful in FSC casework than speaking rate [Gold and French, Künzel].

Künzel [10] examined AR, SR, and various pausing parameters in German. He retested claims that inter-speaker variability was lower in AR as opposed to SR in order to provide conclusions regarding the values of the different speech tempo measures. Künzel was able to confirm prior results in the literature and establish that intra-speaker variability is much smaller in AR than it is in SR. For the experiment, five males' and five females' speech was analyzed for both read and spontaneous speech, and SR was found to be higher in read speech than in the spontaneous speech that was collected. This is largely due to the fact that speakers use far fewer hesitation pauses in read speech than in spontaneous, resulting in a higher SR. AR, on the other hand, did not provide significant differences between read and spontaneous speech, and AR for individual speakers had coefficients of variance that were smaller than they were with SR. To further evaluate the possible discriminating power of SR and AR, Künzel looked at cumulative distributions of both intra- and inter-speaker differences. According to equal error rates calculated, AR was found to have more speaker-discriminating power than SR.

Following Künzel's [10] study that found AR to be a better discriminator than SR, further investigations have gone on to examine AR in more detail. Jessen [11] analyzed the AR of 100 male speakers of German. AR was measured for both spontaneous and read speech in all individuals. It was found that, unlike Künzel [10] the mean AR was significantly higher in read than in spontaneous speech. In order to calculate ARs, Jessen was the first to implement a new methodology in which "memory stretches" were utilized as opposed to "interpause stretches" and "intonation phrases" [12] which are the typical methodologies employed in previous studies. Jessen describes the methodology behind "memory stretches" as "the phonetic expert [going] through the speech signal and [selecting] portions of fluent speech containing a number of syllables that can easily be retained in short-term memory." After listening several times the expert then counts the number of syllables that he/she is able to recall from memory to be included in this portion of speech [11]. This innovative method for identifying speech intervals was reported to save time in the analysis, while also providing reliable figures.

Cao and Wang [13] followed the methodology of Jessen [11] and examined the ARs for 101 male Chinese speakers. All of the analyzed recordings included spontaneous speech and were made over the telephone. They investigated inter- and intra-speaker variation of AR, and found both the global ARs (GAR) and averages of local ARs (LARmean) to be fairly normally distributed. The mean global articulation rate (GAR) was 6.58 syll/sec and the mean of the local articulation rates (LARmean) was 6.66 syll/sec. They also reported that the range of AR for a given speaker is relatively small and stable. Although the previous AR literature has not investigated the discriminant power of speaking tempo, they have established AR as a potentially stable and valuable parameter to consider in FSC cases.

3. Methodology

3.1. Data

The data for the current study are from the Dynamic Variability in Speech (DyViS) Database [14]. The speech data consists of 100 male speakers of Southern Standard British English (SSBE), aged 18 to 25. All participants are native English speakers, university educated, and reported no speech or hearing impairments. Each participant was recorded across four different speaking tasks, however, the data used in the current study is taken only from the second task. Task 2 involves each participant speaking to an accomplice (Research Assistant) about a fictional crime they were involved in. The conversation is meant to allow the participant and the accomplice to corroborate their stories for the police. The 100 recordings ranged in length from 15 to 25 minutes.

3.2. Measuring articulation rate

The general methodology employed in this study follows very closely that of Jessen [11]. In measuring AR a number of decisions related to the methodology have to be made [10,12]. Jessen [11] explains that the first concern in measuring AR is the "kind of linguistic unit on the basis of which AR is counted." As noted in Gold and French [1] the majority of forensic phoneticians use the syllable as a unit of measure, rather than sound segments or words, in turn, producing AR rates in syllables per second as opposed to words per second (or minute). As a native speaker of a language, one has a fairly reliable intuition about the number of syllables that appear in a specific segment of speech. In terms of analysis, this avoids the need to rely on the energy peaks alone for each syllable that appears in the acoustic signal, since that would be a much less reliable method. For these reasons, syllables in this study were determined auditorily through careful listening.

The second important decision for the measurement of AR relates to the linguistic unit (the syllable in this case) as being defined either phonologically or phonetically. A phonologically defined syllable is "defined in terms of the lexicon and grammatical rules of the language", where as a phonetically defined syllable is one that is "manifested in phonetic reality" [11]. Jessen gives an example using the phrase "did you eat yet?" Phonologically we would count this as having four syllables; however, in reality the phonetic number of syllables may be reduced or in some rare cases even increased. If the phrase was to be reduced it may be realized as perhaps two syllables as in "jeet yet" [11]. For this reason, it is important to note that phonological versus phonetic syllables can have a large impact on the number of syllables in a given interval. In a case where a phrase is phonetically only two syllables, AR will obviously be lower than if the same phrase was counted on four phonological syllables (see [11] for further discussion). Jessen [11] suggests that syllables are best defined phonologically, rather than phonetically, therefore the present study is based on phonological syllables.

The final methodological decision, and perhaps the most influential on the results, involves the kind of speech interval that is selected for measuring AR. The AR can be calculated for the entire duration of fluent portions in a recording, known as "global AR", or by taking multiple pieces of fluent speech segments in order to calculate "local ARs" [6]. Miller et al. [15] showed that speakers often change their speech tempo over the course of longer utterances. Therefore, in order to capture such changes in tempo that may occur within a single recording it is more beneficial to obtain local ARs. Previous AR research has used "interpause stretches" and "intonation phrases" to identify speech intervals over which to calculate local ARs [12]. However, Jessen [11] uses an experimental method (memory stretches) which is also the method chosen here for the current study. Jessen [11] suggests that with memory stretches, one avoids empirical or methodological problems associated with previously used methods. He also states that by selecting speech intervals using memory stretches, it allows for "a much simpler and more pragmatic approach."

Sound Forge Audio Studio 10.0 was used for analysis and speech segments were only selected at least two minutes into the recording, to allow the speaker to become comfortable speaking to their accomplice and in the presence of the recording equipment. Similar to [11], only speech segments with fluent speech were chosen and the region marked out. Following the memory stretch procedure, each fluent segment was listened to several times and the speech phrase was then typed out onto the region marker tag (in Sound Forge), along with the number of phonological syllables. After collecting a minimum of 26 local ARs, it was possible to view all recorded regions that listed the number of syllables and also included the length of the speech segment. Those figures were entered into

Microsoft Excel and the mean of the local ARs as well as standard deviations were computed for all speakers.

The maximum number of syllables in a memory stretch was in the very low 20s, but for the majority of speaker it was between 7 and 11 syllables included in a memory stretch (in order not to "push the limits", and avoid mistakes [11]). In keeping with the methodology of Jessen, no fewer than four syllables were used per memory stretch. A four syllable minimum threshold is in place in order to avoid the "inclusion of very short interpause stretches that could unduly increase the effect of phrase-final lengthening on the calculated articulation rate" [11]. It is important to emphasize here that each memory stretch consisted of only fluent speech, which excluded any kind of pauses, either filled or unfilled, repeated syllables, and any syllable lengthening that went beyond phonological requirements in English. A total of 2993 AR measurements were taken across the 100 speakers. The average number of memory stretches measured per speaker was approximately 30, with a standard deviation of 2.1 and a range of 26-32.

3.3. Calculating likelihood ratios

In order to examine the discriminant power of AR for forensic purposes, likelihood ratios (LR) were calculated. The LR calculations for AR were performed using a MatLab implementation of Aitken and Lucy's [16] Multivariate Kernel-Density (MVKD) formula [17]. The MVKD formula by Aitken and Lucy [16] assumes that within-speaker variability is normally distributed (numerator). The between-speaker variation, however, is not assumed to be distributed normally and is estimated using kernel-density, which accounts for skewed distributions. A MatLab script [18] was used to run multiple same speaker same speaker (SS) and different speaker (DS) LR calculations for AR. The script calls for the 100 speakers to be split in half, such that SS comparisons may be performed (50 SS comparisons), which in turn results in 2,450 DS comparisons (50*49). Speakers 001-050 acted as the speaker comparisons, while speakers 051-100 acted as the background population. The calculated raw LRs were transformed using natural and base$_{10}$ logarithms – log likelihood ratios (LLR). The transformation allows zero to act as the center point between the support for H_p and H_d.

Performance of the parameter is examined with respect to log-LR cost (Cllr) and equal error rate (EER), which are both metrics of system validity. The Cllr is a Bayesian error metric that quantifies the ability of the system to output LRs that align correctly with the prior knowledge of whether speech samples were produced by the same or different speakers. The Cllr acts as an error measure that captures the "gradient goodness of a set of likelihood ratios derived from test data" [19,20]. Cllr was calculated using Brümmer's FOCAL toolkit [21] function *cllr.m* with the log-LRs as input. Values of Cllr that are closer to zero indicate that error is low. For values nearing one the error is considered poor, while values above one indicate a very poor performance [22]. EER, unlike Cllr, provides a "hard" accept-reject measure of validity. This is based on the point at which the percentage of false hits (DS pairs that offer support for the H_p) and the percentage of false misses (SS pairs that offer support for the H_d) are equal [23].

4. Results

The distribution of mean ARs for all 100 speakers are presented in Figure 1. The mean articulation rate across the 100 speakers is 6.02 sylls/sec, with a range from 4.57 to 7.79 sylls/sec. The mean standard deviation across the 100 speakers is 1.2 sylls/sec overall, with a range from .68 to 9.17 sylls/sec (the three highest standard deviations, as seen in Figure 1, are outliers). Overall, these results indicate that there is a higher level of variation occurring within a speakers' AR than there is between different speakers' ARs.

4.1. Likelihood ratio results

The results examining the discriminant power AR are summarized in Table 1. The second row of Table 1 contains the results from SS comparisons and the third row contains DS comparison results. The percentage of correct SS and DS comparisons is found in the second column, followed by the Mean LLRs in the third column. A correct LRs is achieved if a LLR for SS comparisons is a positive value (providing support for the prosecution hypothesis), while an incorrect LR is the result of a negative value for a DS comparison (providing

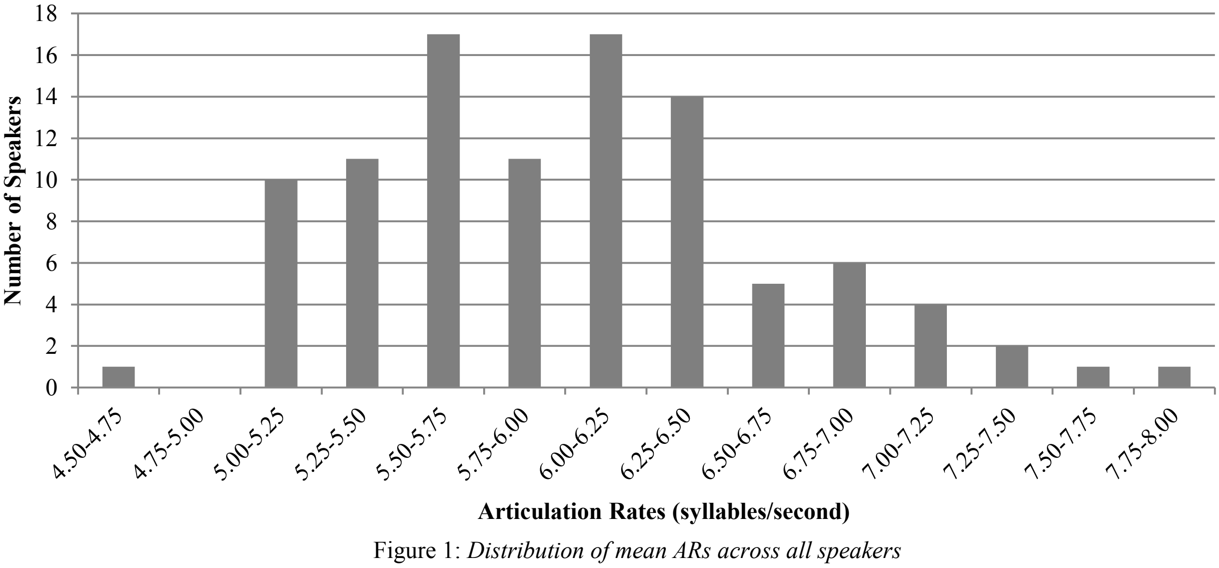

Figure 1: *Distribution of mean ARs across all speakers*

support for the defense hypothesis). Finally, EER and Cllr for AR as system are presented in the fourth and fifth columns.

Table 1 shows that AR is performing much better with SS comparisons than DS comparisons. The results may seem counterintuitive since there is higher within speaker variability than between speaker variability for AR, and it might be assumed that the high within speaker variation would cause DS pairs to perform better than SS pairs. However, it appears that because the degree of variation in AR is so high within speakers overall, the system tends to allocate higher degrees of similarity if two speakers have similar degrees of (high) within speaker variation.

Table 1: *Summary LR-based discrimination for mean articulation rate (100 speakers)*

Comparisons	% Correct	Mean LLR	EER	Cllr
AR SS	90.0	0.18	.3340	.8981
AR DS	46.2	-2.94		

This is evident in the fact that for DS comparisons, the system is performing slightly worse than chance (50%; since a LLR correct/incorrect response is categorized as either for or against the H_p) as the AR system tends to over-predict pairs being the same speaker than different speakers (note the high error rate in correct DS judgments). Following [22], the Cllr for the AR system would classify itself as having a 'poor' performance. The EER is also high at 33.4%, and the mean SS LR offers only limited evidence to support the prosecution hypothesis (H_p). The mean DS LR is slightly stronger, and offers moderate evidence to support the defense hypothesis (H_d).

The Tippett plot in Figure 1 provides a visual measure of the performance of AR as a discriminate feature. The *x-axis* displays $\log_{10}$ LRs where zero is the division between support for H_p (>0) and support for H_d (<0). The y-axis displays cumulative proportion. Contours that are more flat indicate a higher proportion of pairs that achieve a stronger strength-of-evidence, and contours that are steeper indicate a weaker strength-of-evidence. The results for SS and DS comparisons are assessed together. Figure 1 shows that error rates are higher for DS comparisons than they are for SS comparisons.

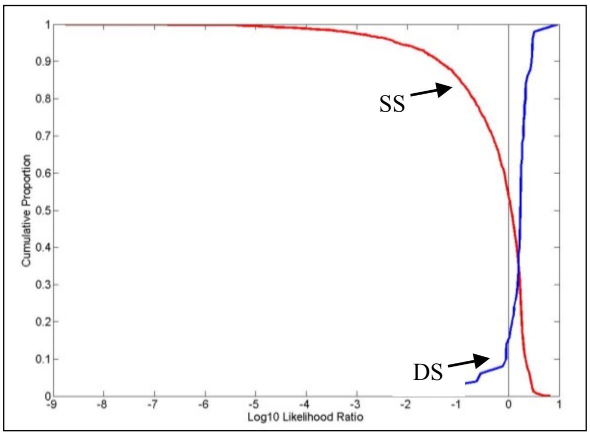

Figure 2: *Tippett plot of articulation rate*

The SS (red) line is steeper than that of the DS (blue) line and provides a relatively low strength of evidence. DS on the other hand can attain higher strength of evidence (a Log_{10} LR above -5), although these values are reserved for a very small percentage of DS comparisons. It is important to remember when analyzing SS and DS LR results that "two samples cannot get more similar for a feature than identical" [24], therefore, DS comparisons will carry the potential for achieving a higher strength of evidence than SS comparisons. The Tippett plot provides an overall picture that AR as an individual parameter is relatively weak at discriminating between individuals, and only produces higher strength of evidence for a very small proportion of DS comparisons.

5. Discussion

Overall, AR can be classified as a speech parameter that carries higher intra-speaker variation than it does inter-speaker variation. AR as a discriminant parameter has proved to be poor, and it is not close to being as good at discriminating between individuals as experts have reported [1]. Results have also shown that AR offers a very weak strength of evidence for SS comparisons, however, DS comparisons can potentially offer a higher level of strength of evidence. It is important to note that although the strength of evidence for DS comparisons is stronger than SS comparisons, there is a higher rate of incorrect DS judgments (~54%).

The results of the analysis of AR as a parameter under an LR framework in forensic speech science signals caution for casework, insofar as parameters previously thought to be good speaker discriminants might transpire to carry higher intra-speaker variation than inter-speaker variation (like AR), which will potentially result in a lower strength of evidence for a given parameter. Further research on speaker discriminants is still needed for other commonly used parameters in forensic casework, because it appears that some experts in the field are analyzing certain features that have not been previously tested empirically. As a result, forensic phoneticians may be giving undue weight to features which provide little in terms of discrimination. This is shown by the fact that 93% of experts analyze speech tempo, despite AR contributing little to discriminating between individuals with *average* ARs.

Although AR is not the discriminant shibboleth experts may have hoped for, it is important that AR is still considered in forensic speaker comparisons in conjunction with other speech parameters. There are instances when speakers may have a very low or high AR, and the parameter can be considered useful. As Rose [24] points out, "not all speakers differ from each other in the same way". Therefore, there will be those few individuals where AR is potentially a good discriminant parameter.

6. Acknowledgements

This project was funded by the European Commission's Seventh Framework Program (FP7-PEOPLE2007) under grant agreement number 238803 (Marie Curie Initial Training Network 'Bayesian Biometrics for Forensics (BBfor2). Thanks also to Professor Peter French for his guidance and support on this project.

7. References

[1] E. Gold and P. French, “International practices in phonetic speaker comparison,” *International Journal of Speech, Language and the Law*, vol. 18, no. 2, pp. 293-307, 2011.

[2] K. Earnshaw, (2014). *Assessing the Discriminatory Power of /t/ and /k/ for Forensic Speaker Comparison using a Likelihood Ratio Approach*. MSc thesis, University of York, 2014.

[3] V. Hughes, *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*. PhD thesis, University of York, 2014.

[4] V. Hughes, P. Foulkes, and S. Wood, “Filled pauses as variables in forensic voice comparison,” *International Journal of Speech, Language and the Law*, vol. 23, no. 1, pp. 99-132, 2016.

[5] C. Kavanagh, *New consonantal acoustic parameters for forensic speaker comparison*. Unpublished; University of York. PhD, 2013.

[6] K. McDougall, “The role of formant dynamics in determining speaker identity,” Ph.D. abstract in the *International Journal of Speech, Language and the Law*, vol. 13, no. 1, pp. 144-145, 2006.

[7] K. McDougall, “Speaker-specific formant dynamics: an experiment on Australian English /aɪ/,” *International Journal of Speech, Language and the Law,* vol. 11, no. 1, pp. 103-130, 2004.

[8] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sönmez, E. Shriberg, A. Stolcke, H. Bratt, and R. R. Gadde, “Speaker recognition using prosodic and lexical features,” in *Proc. IEEE ASRU*, Dec. 2003, pp. 19–24.

[9] J. Laver, *Principles of Phonetics*. Cambridge: Cambridge University Press, 1994.

[10] H. Künzel, “Some general phonetic and forensic aspects of speaking tempo,” *Forensic Linguistics,* vol. 4, pp. 48-83, 1997.

[11] M. Jessen, Forensic reference data on articulation rate in German. *Science and Justice,* vol. 47, pp. 50–67, 2007.

[12] J. Trouvain, Tempo variation in speech production. Implications for speech synthesis. Doctoral dissertation published as Report Nr. 8 of *Reports in Phonetics*, University of the Saarland, 2004.

[13] H. Cao and Y. Wang, “A forensic aspect of articulation rate variation in Chinese,” *Proc. of the 17th International Congress of Phonetic Sciences*, August 17-21, 2011, Hong Kong, China, pp. 396-399, 2011.

[14] F. Nolan, K. McDougall, G. de Jong & T. Hudson, “The DyViS database: style-controlled recordings of 100 homogenous speakers for forensic phonetic research,” *International Journal of Speech, Language and the Law*, vol. 16, no. 1, pp. 31-57, 2009.

[15] J.L. Miller, F. Grosjean, and C. Lomanti, “Articulation rate and its variability in spontaneous speech: a reanalysis and some implications,” *Phonetica,* vol. 41, pp. 215-225, 1984.

[16] C.G.G Aitken and D. Lucy, “Evaluation of trace evidence in the form of multivariate data,” *Applied Statistics,* vol. 54*,* pp. 109-122, 2004.

[17] G.S. Morrison, MatLab implementation of Aitken and Lucy’s (2004) forensic likelihood ratio software using multivariate-kernel-density estimation, 2007. Downloaded: December 2011.

[18] P. Harrison, MatLab script, ss_ds_lrs.m, Downloaded: May 2012.

[19] G.S. Morrison, “The place of forensic voice comparison in the ongoing paradigm shift”. Written version of an invited presentation given at the 2nd International Conference on Evidence Law and Forensic Science. 25-26 July 2009, Beijing, China, pp. 1-16.

[20] G.S. Morrison, “Likelihood-ratio voice comparison using parametric representations of the formant trajectories of diphthongs,” *Journal of the Acoustical Society of America,* 125, pp. 2387-2397, 2009.

[21] N. Brummer, “Focal toolkit,” http://sites.google.com/site/niko brummer/focal, Downloaded: 13 August 2012.

[22] D.A. van Leeuwen and N. Brümmer, “An introduction to application-independent evaluation of speaker recognition systems,” In Müller, C. (ed.) *Speaker Classification I*, LNAI 4343. Berlin: Spinger-Verlag, pp. 330-353, 2007.

[23] N. Brümmer, and J. du Preez, “Application-independent evaluation of speaker detection,” *Computer Speech and Language* vol. 20, no. 2-3, pp. 230-275, 2006.

[24] P. Rose, “The intrinsic forensic discriminatory power of diphthongs” in *Proc. of the 11th Australasian International Conference on Speech Science and Technology*. 6-8 December 2006, University of Auckland, New Zealand, pp. 64-69.