Auditory detectability of vocal ageing and its effect on forensic automatic speaker recognition

Finnian Kelly [1], *Naomi Harte* [1]

[1] Department of Electronic & Electrical Engineering, Trinity College Dublin, Ireland
kellyfp@tcd.ie, nharte@tcd.ie

Abstract

The comparison of non-contemporary speech samples is common in forensic speaker recognition cases. It has yet to be established however, to what extent the time interval between non-contemporary samples can increase before a problem is created for forensic automatic speaker recognition. This paper presents results of a human listener test designed to evaluate the detectability of vocal ageing over increasing intervals of up to 30 years. Subsequently, a forensic automatic speaker recognition evaluation of 15 ageing males at increasing intervals of up to 60 years is presented. It is shown that at intervals of around 10 years, the average detectability of vocal ageing by humans is just above chance. As the interval rises to 30 years, vocal ageing is detected 90% of the time. In the automatic system, vocal ageing is manifested as a drop in intra-speaker likelihood ratios (LRs) as the time interval between non-contemporary samples increases. At an interval of 30 years, LRs for the vast majority of intra-speaker comparisons fall below a value of 100 - commonly interpreted as 'moderate support' on a verbal LR scale. Our findings indicate that at a time-lapse of 30 years, vocal ageing creates significant problems for forensic automatic speaker recognition.

Index Terms: vocal ageing, forensic automatic speaker recognition

1. Introduction

Vocal ageing, or the acoustic changes to the human voice as a result of ageing, has received significant research attention [1, 2, 3, 4, 5, 6, 7].

Vocal ageing is of direct relevance to forensic speaker recognition, where the goal is to determine whether an unknown voice in a questioned recording belongs to a suspected speaker [8, 9]. The nature of recovering voice evidence, which may be from telephone surveillance for example, and acquiring recordings of a suspect, means there is always some degree of time-lapse between recordings to be compared [10]. In some cases the time-lapse runs into years, with an extreme example being the Yorkshire Ripper Hoaxer trial [11], where the interval between the questioned and suspected speaker recordings was 26 years. There is a concern that vocal ageing may weaken the outcome of forensic speaker recognition in the presence of such long time-lapses.

Evaluating the effect of ageing is of particular interest in forensic automatic speaker recognition (FASR), where the task is to quantify the degree of similarity between the speaker-dependent features extracted from the questioned recording and the speaker-dependent features extracted from a suspected speaker recording [9]. There is some evidence that traditional auditory speaker recognition by a trained expert is unaffected by ageing - in the Yorkshire Ripper Hoaxer investigation [11], French et al. concluded that ageing did not undermine the comparison of the questioned and suspect recordings. It is not clear however, the extent to which ageing would have affected an automatic comparison in that case. In addition, the ageing effect is of relevance to speaker profiling and reference population composition [12], as well as FASR in an investigative scenario [13, 18].

There has been little previous research into the effect of ageing on forensic speaker recognition, automatic or otherwise. Künzel [10] completed a study for which 10 male speakers were recorded several times in one year and then again 11 years later. Over this interval, the effect of vocal ageing was not detectable by a group of listeners. In the same study, an FASR experiment with the 10 ageing males was presented. The performance of the automatic system did not degrade as a result of the 11 year interval. There were one or two exceptions in both the listener and automatic experiments, but it was concluded that ageing over 11 years is not a problem for forensic speaker recognition. In a study by Hollien and Schwartz [14], it was found that speaker recognition of 10-11 ageing males by a group of human listeners was not significantly affected by a 6 year interval between samples. At an interval of 20 years however, the correct recognition rate dropped sharply.

Our previous work has investigated the effect of vocal ageing on speaker verification [15, 16]. Over time intervals of up to 60 years, ageing was demonstrated to significantly degrade verification performance. Although speaker verification is not directly applicable to a forensic evaluative scenario [17, 18], the underlying signal processing and pattern recognition is shared with FASR. Thus it can be assumed that ageing will also present problems for FASR. It is not clear at what time interval this will become apparent however.

In this paper, the auditory detectability of ageing in 10 male and 10 female speakers by human listeners is evaluated, and the time interval at which ageing becomes strongly detectable is observed. An FASR evaluation of 15 male speakers at intervals of up to 60 years is then presented. The effect of ageing on likelihood ratio estimation is shown, and a time interval between non-contemporary samples at which ageing becomes a significant issue for most speakers is proposed.

2. Speaker Ageing Data

Ageing speaker recordings were taken from the Trinity College Dublin Speaker Ageing (TCDSA) Database [15, 16], which contains recordings spanning a 30-60 year range per speaker. To minimise variability due to recording quality, only samples meeting certain quality requirements were included [15]. The database has recently been expanded with 8 speakers sourced from 'The Up series' television documentary series [19], bringing the total number of speakers to 26 (15 males and 11 fe-

males). All 15 males were used as subjects in the FASR experiment, and a subset of 10 males and 10 females were used for the auditory detectability test.

In addition to the longitudinal TCDSA database, a new database of male speakers, the TCDSA-FD (Forensic Development) database, was compiled for development of the FASR system. It was designed to provide suitable background and population modelling for the TCDSA male subjects. It contains approximately 1 hour of speech from 25 speakers in each of 5 age ranges: 25-35, 36-45, 46-55, 56-65 and 66+. All are microphone recordings of interviews and speeches sourced from YouTube. All speakers are Irish-accented, with some variation in accent source and strength. As with the TCDSA database, most speakers are professional speakers (politicians, presenters or public figures). All data used in this paper is freely available for academic research.

3. Auditory Detectability of Ageing

The degree to which human listeners can detect vocal ageing is of interest to forensic speaker recognition from a number of perspectives. The ability of listeners to compare non-contemporary speech samples, and hence perform auditory speaker recognition, will be influenced by the detectability of ageing. If there is a relationship between a speaker's ageing detectability and the output of an automatic system then this may help identify ageing-robust features.

3.1. Ageing comparison experiment

An experiment was designed to evaluate the detectability of ageing at different time intervals. A group of 25 listeners (8 females, 17 males) were recruited for the experiment. The majority (20) were native English speakers. The 5 non-native speakers were fluent in English. The age range of the listeners was between 17 and 47, with a mean age of 28.8 years. All were University students or professionals (apart from one: a 17 year old male). None had a history of hearing or speech impairment.

The test was similar in its aim to that of Künzel [10], and thus followed a similar protocol. 10 males and 10 females from the TCDSA database were chosen as test subjects. The focus of this paper will be on male subjects (our automatic experiment is male-only), but female results will also be presented for comparison. For each test speaker, age ranges 1-4 were defined as: 28-39, 40-45, 46-54, 55-64. A sample was extracted for each speaker from each age range, resulting in a minimum age interval of 6 years, a maximum of 14, and an average of 9.7 years.

For each speaker, a set of 6 comparison tests were created by pairing age-range combinations in the following way: 1-2, 1-3, 1-4, 4-1, 4-2, 4-3, where 1-2 denotes age range 1 followed by age range 2. For each occurrence of an age range in this set of comparison tests, a different 5 second sample was used - thus there were 12 different 5 second samples used to compile the 6 comparison tests. Each pair of samples were separated by a 0.5 second beep, with 0.25 seconds of silence either side.

A simple web application was designed to administer the listener test (accessible at [20]). 24 listeners used headphones, 1 used loudspeakers. Of a total of 120 possible comparison tests, each listener was presented with 48 (male and female) tests in a random order. Each comparison test was comprised of an audio player object and a two-part question. The audio player object allowed multiple plays of each recording, and listeners were instructed that they could listen to a comparison more than once if necessary. The number of plays for each question was recorded. The question was in a two-alternative forced-choice format, whereby listeners were asked whether the speaker was older in part 1 or part 2 of the sample. They were also asked to give the certainty of their decision on a 4 point scale: "completely certain", "quite certain", "quite uncertain", "completely uncertain (had to guess)". The average test duration was just over 20 minutes.

3.2. Ageing comparison results

The responses of the listener test were analysed by calculating the percentage of correct decisions for different question subsets: per-speaker, per-gender and per-age interval. The percentage of correct decisions - i.e. those correctly identifying in which part of the comparison sample the speaker was older - was calculated by dividing the total number of correct decisions by all listeners by the total number of decisions.

In Table 1, the percentage of correct decisions for males and females at different mean ageing intervals in forwards (younger-older comparison) and backwards (older-younger comparison) directions are given. Ageing detectability for males at an interval of 10 years is not far above chance level (58.9%) for a combination of forwards and backwards directions. With increasing age interval, detectability rises to 66.8% at 20 years and 85% at 30 years. Interestingly, detectability in a forwards direction is noticeably greater than in a backwards direction, particularly in males. This agrees with the findings of Künzel's experiment [10]. Also noticeable is the higher detectability of female subjects than males, in both directions. The combined percentage correct for females is 8-13% greater than males at all age intervals. We will discuss these observations in Section 5.

		Correct Decisions (%)		
speaker	interval	forwards	backwards	combined
males	10 years	66.7	51.5	58.9
	20 years	76.1	58.9	66.8
	30 years	84.9	85.1	85.0
females	10 years	74.7	64.1	70.0
	20 years	78.1	82.1	80.2
	30 years	94.6	92.31	93.4
all	10 years	70.7	57.6	64.1
	20 years	77.2	70.7	73.7
	30 years	89.5	88.6	89.0

Table 1: Percentage of correct decisions for different speaker genders, mean age intervals and comparison directions.

In Figure 1 (a) and (b), the percentage correct values for each male speaker at each age interval in forwards and backwards directions is shown. The variability in detectability between speakers is evident. The detectability of the second speaker, 'bowman', for example, is low at all intervals relative to the speaker average. The general within-speaker consistency of increasing detectability with age interval confirms a speaker-dependent, rather than sample-dependent stimulus is informing the listener decisions. In Figure 1 (c) and (d), the percentage correct value is plotted against age interval for each male comparison test, in forwards and backwards directions. A general trend of increasing detectability with age interval can be seen.

4. Forensic Automatic Speaker Recognition of Ageing Speech

Using the 15 males from the TCDSA database (Section 2) as subjects, an experiment was designed to evaluate the effect of ageing on forensic automatic speaker recognition (FASR).

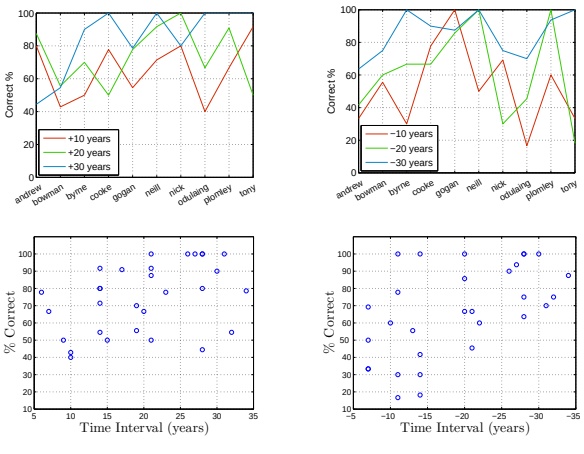

Figure 1: Percentage correct values for male speakers; Top Left, (a): Forwards comparison. Top right, (b): Backwards comparison. Bottom left, (c): vs forwards comparison time intervals. Bottom right, (d): vs backwards comparison time intervals.

4.1. Speaker Recognition System

The application of automatic speaker recognition techniques to forensics is well established [21, 9, 22, 23]. The accepted approach is to use a Bayesian framework for the interpretation of voice evidence, via the formulation of a likelihood ratio (LR) measure. The LR measure can be interpreted as the ratio of two hypotheses - the hypothesis that a suspected speaker is the source of the speech in the questioned recording, and the hypothesis that he/she is not. With the LR measure, speech evidence can be combined with prior probabilities (e.g. background information provided by police) to calculate posterior probabilities that can be used for judicial decisions [21].

In this work, a Gaussian Mixture Model (GMM) system, similar to [23], was used for the automatic experiments. Recordings of a suspected speaker are tested against his GMM. The resulting similarity-scores are used to estimate a (normally distributed) within-source (WS) similarity-score distribution. The questioned recording is tested against the GMMs of a set of speakers from the same population as the suspected speaker, and the resulting similarity-scores are used to estimate the between-source (BS) distribution. The questioned recording is then tested against the suspected speaker GMM, and the resulting similarity-score is evaluated on both the WS and BS distributions. The ratio of these values is the LR. Here, 'similarity-score' refers to the log-likelihood ratio between a GMM and a Universal Background Model (UBM) [24].

4.2. Forensic Automatic Speaker Recognition Experiment

A UBM was trained with 2.5 hours of speech from the TCDSA-FD database, distributed evenly across age ranges. For each speaker, two GMMs were trained, using 1 minute of speech from their first and last year of data respectively. All years of data in between were taken as questioned recordings.

The WS distribution was estimated by testing each GMM with data from the training year. As data availability was limited, a bootstrapping procedure [23] was applied. A set of 30 speakers, distributed evenly across age ranges, were taken as population speakers. Given GMMs trained for each with 1 minute of speech, the BS distribution was estimated from the log-likelihood scores of 30 second excerpts from each questioned recording. Within-source degradation prediction (WDP), within-source minimum variance limiting (WMVL) and outlier removal were applied [23]. WDP enables the WS distribution to be adequately estimated when there only one recording available (at a given age) for the suspected speaker.

Finally, a 30 second excerpt of each questioned recording (QR) was tested against suspected speaker GMMs, and evaluating on the BS and WS distributions, LRs in forwards and backwards directions were calculated. 'Forwards' indicates that the year of the QR is after the year of the suspected speaker recording (used to estimate the WS), and 'backwards' indicates that the year of the QR is before the year of the suspected speaker recording.

4.3. Front-end processing and GMM system parameters

All speech was pre-processed by downsampling to 8kHz, removing silences with an energy-based voice activity detector and applying pre-emphasis. MFCC features of length 12 were extracted over 20ms windows with 50% overlap and a Mel filterbank of 26 bands. Mean and variance normalization were applied to the features followed by RASTA filtering. Finally, delta coefficients were appended. See [24] for details.

The UBM contained 512 components, and was trained with 10 iterations of the expectation maximisation (EM) algorithm [25]. GMMs for each speaker in were trained by adapting the UBM means [26]. A relevance factor of 16 was used, and all UBM components were considered in adaptation and scoring. A Z-normalisation [24] was applied to all scores, using statistics calculated from a set of 25 speakers from the TCDSA-FD database, distributed evenly across age ranges. There was no overlap between TCDSA-FD speakers used for the UBM, population, or Z-normalisation datasets.

4.4. Speaker Recognition Results

An example of the ageing likelihood ratio (LR) calculation is shown in Figure 2. As the time interval between the suspected speaker recording (used to estimate WS) and the questioned recording (QR) increases, the log-likelihood ratio of the QR evaluated on the WS distribution decreases, while increasing on the BS distribution. Thus, the LR decreases. For comparison, a LR calculated for an imposter recording has been included. As the time interval increases, the imposter and suspected speaker LRs converge.

In Figures 3 and 4, the LRs for each speaker are plotted against the time intervals between the year of the questioned recording and the year of the suspected speaker recording, in forwards and backwards directions. A (base 10) log and a linear fit has been applied to the scores of each speaker. With only one exception, the LR scores of all speakers degrade at a relatively consistent rate. Künzel [10] reported no great LR degradation for the majority of speakers after 11 years. Our results show a more significant degradation after this interval. The results here also show more variability, which may be attributable to the conditions being less controlled than in Künzel's experiment.

5. Discussion

The listener test presented in this paper aligns with previous work by Künzel [10], in that detectability was just above chance level at a 10 year time lapse, and was higher in a forwards than a backwards direction. Our extension of the test to an interval of 30 years demonstrates that ageing effects are strongly detectable (90% correct) over this interval. Thus, auditory speaker recognition over this interval will be significantly affected.

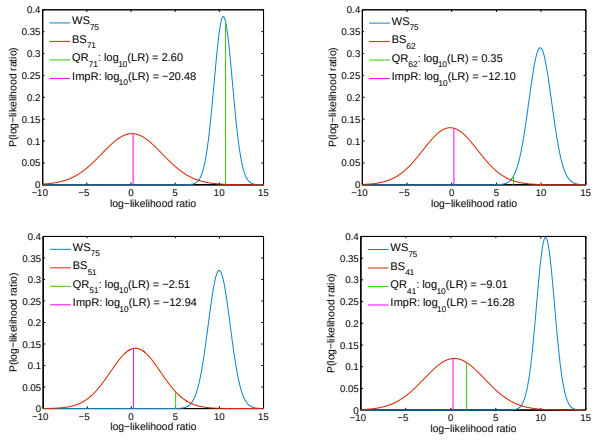

Figure 2: LR calculation for 'odulaing' in a backwards direction at increasing time intervals (see trend for this speaker in Figure 4). WS_{xx} denotes the within-source distribution estimated with a recording of the suspected speaker at age xx. BS_{yy} is the between-source distribution estimated with the age yy questioned recording. QR_{yy} is the log-likelihood ratio of the age yy questioned recording given the suspected-speaker GMM, and ImpR is the log-likelihood ratio of an imposter recording given the suspected-speaker GMM. The resulting LRs for the QR and ImpR are shown.

It was suggested by Künzel [10] that the difference in forwards and backwards ageing detectability may be as a result of humans being more adept at recognising chronological ageing (younger-older) than the reverse. Although not the focus of this paper, the fact that ageing in females was more detectable than in males may be attributable to the differences in male and female ageing. In females, for example, speaking fundamental frequency decreases with age in a more consistent manner than in males [6, 7]. It has been suggested that males show more marked effects than females with ageing [8]. This may bring with it a variability that makes it difficult to find ageing cues consistent across males in general. Aspects of the test, including listener age/gender effects, the certainty of their decisions, and the number of sample plays, are beyond the scope of this paper and will be addressed in future work.

The FASR experiment demonstrated the detrimental effect

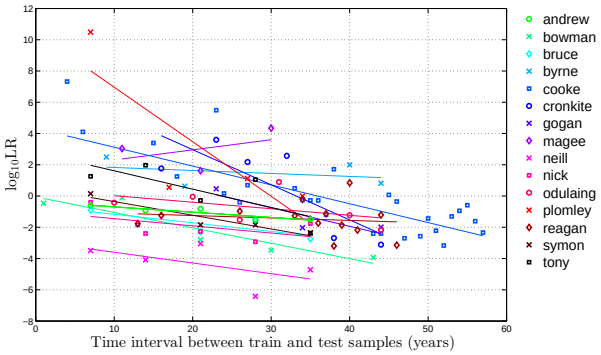

Figure 3: LR scores against the time interval between the year of the questioned recording and the year of the suspected speaker recording, in a forwards direction.

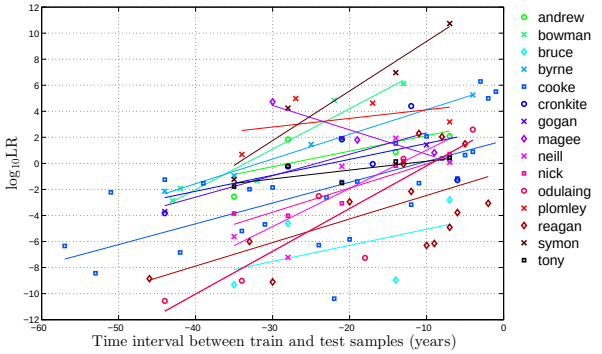

Figure 4: LR scores against the time interval between the year of the questioned recording and the year of the suspected speaker recording, in a backwards direction. Note that there is a different y-axis scale in Figure 3 for the purpose of visibility.

of ageing on LR estimation. All speaker LR trends in Figures 3 and 4 fall below 100 ($\log_{10} 100 = 2$) at a time interval of 30 years, with one or two exceptions. An LR value of 100 corresponds to the minimum value for 'moderate support' on a widely used verbal LR scale [8]. There is much inter-speaker variability however, and after 10 years, approximately half of the speakers have an LR value less than 100. In can be concluded from this, that after 30 years (but likely sooner), the LR of a typical male speaker will degrade to a point where its evidential value is significantly weakened.

It is unclear how to select speakers for the UBM, population and normalization sets in the case of long-term ageing, as there is a choice to be made of whether to select speakers of similar age to the suspected speaker or to the questioned recording. In this paper, a compromise was made by selecting speakers covering a range of ages. The use of speakers of a similar age to the suspected speaker was also investigated - there was no significant difference in the resulting LRs than those presented here.

If there were an overlap between the auditory features used by listeners to detect age and the features (MFCCs) used by the automatic system, some correlation between the output of the automatic system and the listener test results would be observed. It would be expected that those speakers with a high ageing detectability would be those with the most LR degradation. Some instances of this can be observed - 'andrew' for example, in a forwards direction, has poor ageing detectability and a slow rate of LR degradation. In general however, there appears to be little correlation between automatic and listener testing results.

6. Conclusions

Auditory detectability of ageing increases, on average, from just above chance level at 10 years, to 90% after 30 years. For a few speakers however, ageing is strongly detectable ($\geq$80%) after 10 years (Figure 1 (a) and (b)). The LR values of ageing speakers calculated with an automatic system degrade as the time interval between the questioned and suspected speaker recordings increases. After 30 years the LR value of almost all speakers drops below a 'moderate support' level. For many speakers the LR drops to this level after just 10 years. It is concluded that ageing will have a detrimental effect on forensic automatic and auditory recognition within 10 years for some speakers, and certainly within 30 years for most.

7. References

[1] W. Decoster and F. Debruyne, "Longitudinal voice changes: Facts and interpretation," *Journal of Voice*, vol. 14, no. 2, pp. 184–193, 2000.

[2] J. D. Harnsberger, W. S. Brown Jr, R. Shrivastav, and H. Rothman, "Noise and tremor in the perception of vocal aging in males," *Journal of Voice*, vol. 24, no. 5, pp. 523–530, 2010.

[3] S. E. Linville, "The Aging Voice," *The American Speech-Language-Hearing Association (ASHA) Leader*, pp. 12–21, 2004.

[4] P. B. Mueller, "The Aging Voice," *Seminars in Speech and Language*, vol. 18, no. 2, pp. 159–169, 1997.

[5] S. Schötz, "Perception, Analysis and Synthesis of Speaker Age," Ph.D. dissertation, 2006, Lund University, Sweden.

[6] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman, "Changes in Acoustic Characteristics of the Voice across the Life Span: Measures from Individuals 4-93 Years of Age," *Journal of Speech, Language, and Hearing Research*, vol. 54, pp. 1011–1021, 2011.

[7] P. Torre III and J. A. Barlow, "Age-related changes in acoustic characteristics of adult speech," *Journal of Communication Disorders*, vol. 42, no. 5, pp. 324–333, 2009.

[8] P. Rose, *Forensic Speaker Identification*. Taylor & Francis, 2002.

[9] A. Drygajlo, "Forensic Automatic Speaker Recognition," *Signal Processing Magazine, IEEE*, vol. 24, no. 2, pp. 132–135, 2007.

[10] H. J. Künzel, "Non-contemporary speech samples: Auditory detectability of an 11 year delay and its effect on automatic speaker identification," *The International Journal of Speech, Language and the Law*, vol. 14, no. 1, pp. 109–136, 2007.

[11] J. P. F. French, P. Harrison, and J. Windsor-Lewis, "R v John Samuel Humble: The Yorkshire Ripper Hoaxer trial," *The International Journal of Speech, Language and the Law*, vol. 13, no. 2, pp. 256–273, 2006.

[12] R. Rhodes, "Assessing the strength of non-contemporaneous forensic speech evidence," Ph.D. dissertation, 2012.

[13] D. Meuwly, "Forensic speaker recognition: An evidence odyssey, summary," in *Odyssey 2004*, 2004.

[14] H. Hollien and R. Schwartz, "Speaker identification utilizing non-contemporary speech," *Journal of Forensic Sciences*, vol. 46, no. 1, pp. 63–67, 2001.

[15] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score-ageing-quality classification space," *Computer Speech & Language*, vol. 27, no. 5, pp. 1068–1084, 2013.

[16] ——, "Compensating for ageing and quality variation in speaker verification," in *Interspeech 2012*.

[17] C. Champod and D. Meuwly, "The inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, no. 2-3, pp. 193–203, 2000.

[18] G. Jackson, S. Jones, G. Booth, C. Champod, and E. W. Evett, "The nature of forensic science opinion – a possible framework to guide thinking and practice in investigations and in court proceedings." *Science & Justice*, vol. 46, no. 1, pp. 33–44, 2006.

[19] "The Up series," 1977-2012, directed by Michael Apted, produced by Granada Television.

[20] "http://edu.surveygizmo.com/s3/1172145/Age-Comparison-Remote-Firefox," Age Comparison Experiment - accessible as of 29/05/2013.

[21] D. Meuwly and A. Drygajlo, "Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM)," in *Odyssey 2001*, 2001.

[22] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech & Language*, vol. 20, no. 2/3, pp. 159–191, 2006.

[23] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia, "Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 331–355, 2006.

[24] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[25] J. Bilmes, "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," International Computer Science Institute, Tech. Rep., 1988.

[26] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.