Please cite this study as: A. Tsanas, E. San Segundo, P. Gomez-Vilda: "Exploring pause fillers in conversational speech for forensice phonetics: findings in a Spanish cohort including twins", 8th International Conference on Pattern Recognition Systems, 11-13 July 2017, Madrid, Spain

Exploring Pause Fillers in Conversational Speech for Forensic Phonetics: Findings in a Spanish Cohort Including Twins

A. Tsanas[1,2]*, E. San Segundo[3], P. Gómez-Vilda[4]

[1]Usher Institute of Population Health Sciences and Informatics, Medical School, University of Edinburgh, UK, [2]Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, UK, [3]Department of Language and Linguistic Science, University of York, UK, [4]NeuVox Lab, Center for Biomedical Technology, Universidad Politécnica de Madrid, Madrid, Spain

*Asterisk indicates corresponding author: Nine Edinburgh Bioquarter, 9 Little France road, Edinburgh, UK, EH16 4UX (Athanasios.Tsanas@ed.ac.uk, tsanas@maths.ox.ac.uk)

Keywords: Forensic phonetics, fundamental frequency contour, pause fillers, speech signal processing.

Abstract

Pause fillers occur naturally during conversational speech, and have recently generated interest in their use for forensic applications. We extracted pause fillers from conversational speech from 54 speakers, including twins, whose voices are often perceptually similar. Overall 872 tokens of the sound [e:] were extracted (7-33 tokens per speaker), and objectively characterised using 315 acoustic measures. We used a Random Forest (RF) classifier and tested its performance using a leave-one-sample-out scheme to obtain probabilistic estimates of binary class membership denoting whether a query token belongs to a speaker. We report results using the Receiver Operating Characteristic (ROC) curve, and computing the Area Under the Curve (AUC). When the RF was presented with at least 20 tokens in the training phase for each of the two classes, we observed AUC in the range 0.71-0.98. These findings have important implications in the potential of pause fillers as an additional objective tool in forensic speaker verification.

1 Introduction

Speaker identification, speaker recognition, and speaker verification form part of an established area which has attracted considerable research interest over the years [1,2]. Similar to other biometrics, such as fingerprints, the aim is to determine or verify the identity of a person on the basis of some unique properties. More recently, researchers have worked on the considerably more challenging scenario of forensic speaker comparison, which focuses on human speech or acoustic clues of some kind in a forensic setting [3,4]. One of the aims may be to identify whether a suspect's voice matches some preceding voice recording that may be available such as ransom demand or bomb threat.

In many applications, speech is recorded under highly controlled acoustic conditions, with the option to repeat recordings if they are not of sufficiently good quality. For example, clinical speech assessment often takes place in an anechoic chamber to prevent external noise from distorting the signal, and uses high quality microphones [5]. Moreover, there is control over the speech activity, such as requiring speakers to read pre-specified sentences aloud, repeat phrases with linguistically meaningful units, or sustain vowels [5,6]. In a forensic context, recordings will almost certainly be recorded over the phone, which distorts speech signals due to bandwidth limitations; therefore, there are many more pragmatic considerations compared to controlled laboratory conditions.

Of particular interest to forensics is the study of twins, because they tend to exhibit very similar biometrics [7]. Monozygotic (MZ) twins present very similar anatomical characteristics, which are also marked on their voices. Similarly, same-gender DiZygotic (DZ) twins may have similar voices, although they share on average half their genetic information. Distinguishing their voices is difficult; research on twins is therefore worth exploring for forensic purposes [8].

Processing casual conversational speech even when high quality data is available may be challenging due to the lack of any standardization, which is the reason why in clinical settings sustained vowels are typically used [5]. However, natural conversational speech comprises a number of disfluencies and hesitations such as *uh* and *um*, which are often observed when speakers attempt to remember something or to form coherent sentences, for example. These sounds are known as *pause fillers*, and research has demonstrated there are personal variants people favor, suggesting they could be used to distinguish speakers [9]. There are different pause fillers depending on the language, region, or even generation of speakers; the most commonly used pause filler amongst speakers in Spain is the sustained "ehh", phonetically denoted as [e:] [10]. Recent research work has made an exciting link: pause fillers exhibit similar characteristics to (short) sustained vowels (relatively constant amplitude and frequency) [10], suggesting that research findings on sustained vowels might be applicable in the setting of processing pause fillers.

The fundamental frequency (F0) is considered to be the single most important characteristic of speech [5], and it is critical to estimate it accurately in diverse applications [5,11]. Although the concept of F0 is intuitively simple to understand as the dominating frequency in the signal, Roark has made a compelling case that there is no simple definition of F0 beyond

a single period and demonstrated that small changes in the hyper-parameters of F0 estimation algorithms often have large effects on the results [12]. Talkin provided an excellent summary of the objective difficulties in accurately estimating F0 in practice, emphasizing it is time-varying potentially widely over short lengths of time and the presence of harmonics [13]. It is difficult to conclusively decide on the best F0 estimation algorithm, and probably there is not a single best approach for *all* applications [13]. Commonly used algorithms for F0 estimation such as Praat [14], for example, rely on simple-to-use mathematical tools (auto-correlation), implicitly making strong assumptions regarding the stationarity of the speech signal which are violated in practice [5]. A recent study has thoroughly investigated 10 widely used F0 estimation algorithms in the sustained vowel /a/ setting using two speech databases and including analysis on degraded, telephone-quality signals [15]. It was reported that the Nearly-Defect-Free (NDF) [16] and the Sawtooth Waveform Inspired Pitch Estimator (SWIPE) [17] performed considerably better compared to some widely used competing F0 estimation algorithms; moreover, an adaptive framework combining the outputs of multiple F0 estimation algorithms could improve further the accuracy of estimating F0 [15]. Here, we use these findings to inform the processing of the pause fillers, focusing on understanding the time course of the F0 values for the duration of the signal (typically referred to as F0 *contour*).

The aims of this study were to: (a) verify the potential of using pause fillers in speaker verification using a more rigorous supervised learning setup compared to our recent previous study [10], and (b) investigate the required conditions which would enable accurate results to be achieved in this setting.

2 Data

We used the data previously described in San Segundo [10,18]. The dataset comprises 54 male native speakers of standard peninsular Spanish, aged 18-52, 24 of whom were MZ twins, 10 were DZ twins, 8 brothers and 12 unrelated speakers. The participants' health status was screened at the time of the recordings by standard health questionnaires.

All recording sessions were completed in the Phonetics Laboratory of the Consejo Superior de Investigaciones Científicas (CSIC) in Madrid. We used similar recording settings to those used in relevant studies in forensic phonetics by other research groups [19,20]. Each speaker was recorded on two timely distinct periods (2-4 weeks apart) to account for and assess intra-speaker variability. Participants were required to come in pairs for the recording sessions because several speaking tasks required collaborative exercises: twins came together, and other speakers joined with a friend/colleague, or their brother. Although several different speaking tasks were used, here we focus only on one of the speaking tasks: informal interview between a speaker and an experienced interviewer (E.S.S.) who stimulated a speaking style similar to those observed in forensic recordings. The interview lasted approximately 10 minutes and was carried over telephone to simulate realistic forensic conditions. The recordings were obtained using high-quality omnidirectional microphones, sampled at 44.1 kHz with 16-bit resolution. To better approximate a more realistic setting in a forensic setting, the voice signal was low-pass filtered at 3.4 kHz, high-pass filtered at 300 Hz, and down-sampled to 8 kHz. The interviewer elicited responses which required remembering past events, thus probing for hesitations which in practice lead to pause fillers during casual discussion. We confirmed that pause fillers typically took the phonetic form of a long [e].

3 Methods

The methodology used in this study comprises the extraction of pause fillers from conversational speech, the F0 estimation as an integral speech signal processing component, and finally the characterization of each signal by applying signal processing algorithms to extract acoustic measures.

3.1 Extracting pause fillers from conversational speech

We manually located and extracted tokens in each of the two recording sessions for each study participant. For each recorded session for each participant, 7 – 33 *tokens* of [e:] were extracted, with a mean duration of about 200 milliseconds. Ultimately, this led to the extraction of 880 tokens for all 54 participants. We excluded eights tokens which were too short; therefore, we used 872 tokens.

3.2 F0 estimation

We used the NDF algorithm to extract F0 [16], based on our previous findings in the analysis of sustained vowels [15]. We have found that particularly for short signals, NDF performed extremely well. NDF uses a fusion approach from both time-domain and frequency-domain-based instantaneous F0 candidates to determine the final F0 estimates. We refer to Kawahara *et al*. [16] for further algorithmic details. We computed F0 estimates every 1 millisecond for each signal, so for the average pause filler of 200 milliseconds we obtained a vector comprised of 200 successive F0 estimates, the F0 *contour*. This is a critical pre-processing step, and the F0 contour is used by many the acoustic measured described next.

3.3 Extracting additional acoustic measures

We have used the Voice Analysis Toolbox [21-23] to compute the acoustic measures (henceforth *features*) in this study. This toolbox was originally developed for processing sustained vowel /a/ phonations, but the similarity of the pause fillers with sustained vowels [10] suggests that these acoustic measures may be well adapted in this setting.

As indicated previously, many of the computed features rely on the accurate estimation of the F0 contour. These include the *jitter variants*, which quantify instabilities in F0. Similarly, we have computed F0 differences compared to normative data, as well as general F0-based statistics as part of the F0-related measures. The F0 contour was also used as the input signal in a wavelet decomposition scheme, which is a generic time-series approach that was previously shown to work very well in a related application [22]. Other features included the *shimmer variants*, i.e. amplitude perturbations to quantify whether the speaker retains a relatively stable volume in the

pause filler. Many of the other families of acoustic measures fall under the umbrella concept of signal to noise ratio. Finally, we have computed the 42 Mel Frequency Cepstral Coefficients (MFCCs) using the VoiceBox by Brookes [24]. MFCCs are widely used as the standard benchmark in speaker recognition applications [2].

Overall, we have characterised each token using 315 features, thus giving rise to a design matrix of 872×315. The design matrix contained no missing entries. Table 1 summarizes all the features used in the study.

Table 1: Summary of computed features

Feature	Description	Number
Jitter variants	Fundamental frequency perturbations	30
F0-related measures	F0-based statistics and comparisons against normative data	9
Wavelet-based measures	Wavelet decomposition methods of F0	182
Glottal quotient	Vocal fold cycle variability	3
Recurrence Period Density Entropy (RPDE)	Uncertainty in F0 estimation compared to normative data	1
Pitch Period Entropy (PPE)	Quantifying variability in F0 compared to normative data	1
Detrended Fluctuation Analysis (DFA)	Stochastic self-similarity of turbulent noise	1
Shimmer variants	Amplitude perturbations	21
Harmonics to Noise Ratio	Signal to noise ratio using autocorrelation	4
Glottal to noise excitation (GNE)	Noise synchronization in different frequency bands	6
Vocal Fold Excitation Ratio (VFER)	Noise synchronization in different frequency bands	9
Empirical Mode Decomposition Excitation Ratio (EMD-ER)	Decomposing the signal in multiple time series using EMD and quantifying energy and entropy	6
Mel Frequency Cepstral Coefficients (MFCC)	Amplitude and spectral fluctuations	42

3.4 Statistical analysis, mapping, and model validation

We computed the correlation coefficients between each feature and the binary outcome (samples from the same subject versus samples from the remaining 53 subjects); the process was repeated iteratively for each of the 54 subjects in the study. We used the standard rule of thumb that statistical correlations with a magnitude greater than 0.3 are *statistically strong* [25,26].

Subsequently, we used a standard supervised learning setup, employing a Random Forest (RF) classifier [27]. RF is a powerful tree-based ensemble approach, which has been described as one of the best off-the-shelf classification schemes [28]. In short, RF is comprised of multiple weak learners, the decision trees, each of which casts a vote on the class of a query sample. Aggregating the voting from all trees gives rise to the probabilistic estimate of class membership for each query sample. As an integral part of the tree-growing process, RF also provides an estimate of the importance of the features, henceforth referred to as *RF importance scores*, thus providing tentative insight into the characteristics which contribute most towards correctly determining class membership. In all cases, we focused on binary classification: in the practical forensic phonetics perspective, this would be the equivalent of testing whether the query sample belongs to a certain speaker, versus the alternative that it belongs to some other speaker.

Due to the limited number of samples, we used the leave-one-sample-out validation scheme: we trained the model using the $N-1$ samples (871 tokens), and computed its probabilistic performance denoting class membership on the left-out token. The process was repeated for all tokens in the dataset. We report results using the Receiver Operating Characteristic (ROC) curve, and computing the Area Under the Curve (AUC) approach: essentially this is a convenient metric to report in binary classification settings. ROC curves provide a compact visual impression of the trade-off between *sensitivity* (true positive rate, i.e. correct verification that the token belongs to the speaker investigated) and *specificity* (true negative rate, i.e. correctly identifying that the token belongs to a different speaker). The cut-off threshold can be adaptively set by the user depending on the trade-off they want to achieve. ROC curves and AUC are commonly used as performance metrics in binary classification settings reporting model accuracy [28,29]. In all cases, we only report the out-of-sample results, i.e. the results on the tokens not used in the training phase of the classification.

4 Results

Figure 1 provides a visual impression of the absolute values of the correlation coefficients. We observe that some correlation coefficients indicate statistically strong associations ($|R| > 0.3$), but there is no unique feature that is strongly correlated with the outcome for all subjects.

The computation of correlation coefficients provides an indication regarding how difficult the classification task is. The RF classifier fuses the information content from the features to form rules so that the tokens can be probabilistically accurately assigned to the correct class membership. Following standard supervised learning rules of thumb, a minimum number of 20 samples should be ideally available for each class [28]. Here, we have noted that for some participants there are fewer tokens to work on; therefore, the classification task is objectively more difficult to provide good results. Figure 2 presents the ROC curve for the speaker with the greatest number of extracted pause fillers [e:] (33 tokens). We observe that the AUC for this speaker is 0.93, indicating that we obtain an excellent trade-off between sensitivity and specificity.

Next, we repeated the statistical mapping process with the RF for all 54 speakers for completeness. It is well established that a minimum number of training samples (empirically 20-

25) is required to maximize the probability that a classifier would correctly differentiate classes [28]. Unfortunately, for many speakers there are relatively few tokens available, which hinders the statistical power of the supervised learning setup.

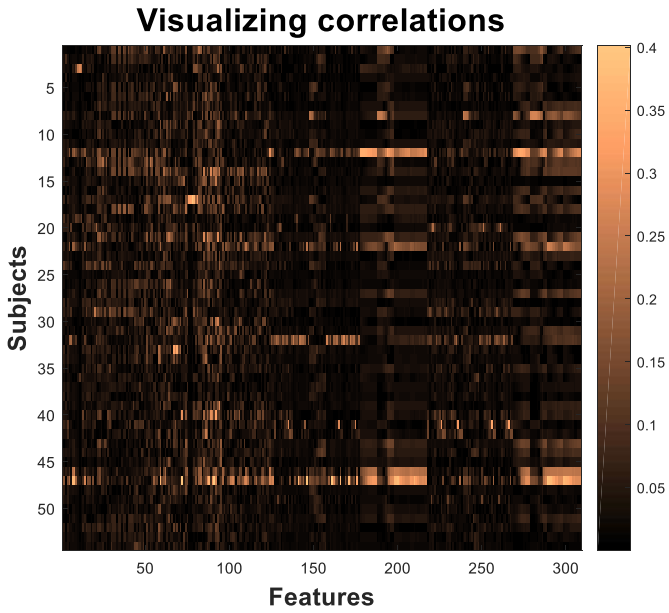

Fig. 1: *Visual presentation of the absolute correlation coefficients for the features used in the study, across all 54 subjects. The binary response investigated was the feature values to identify the jth subject versus the remaining 53 subjects. We remark that some univariate correlations are statistically strong ($|R| > 0.3$), but there is no unique feature that stands out consistently to identify a subject.*

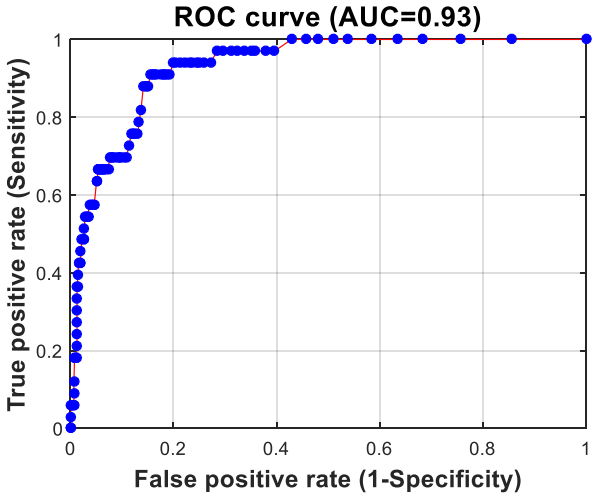

Fig. 2: *Indicative Receiver Operating Characteristic (ROC) curve for the speaker with the greatest number of extracted pause fillers [e:] (33 tokens) from the two interview sessions with an experienced interviewer.*

Figure 3 provides a summary of the AUC computed for each speaker as a function of the number of extracted tokens; we have also noted whether the speaker was a MZ or DZ twin, or if the tokens belonged to brothers or unrelated speakers. As expected, there is a clear trend for the classification setting to be more successful when having a minimum number of samples. Using the rule of thumb that at least 20 samples should be available for a supervised learning classification setup, we note that the AUC varies between 0.71-0.98. Unfortunately, the low number of speakers for whom more than 20 tokens is available does not allow further investigation on establishing whether twins are considerably more difficult to be probabilistically correctly detected.

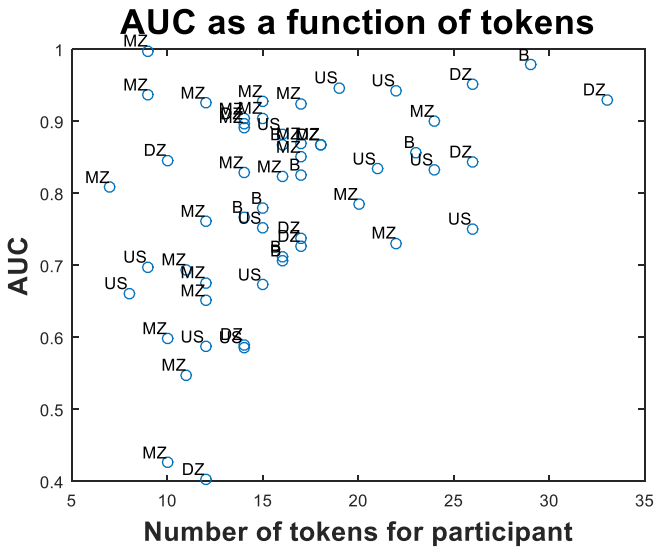

Fig. 3: *Presenting the AUC as a function of the number of tokens used for each participant. MZ: MonoZygotic twins, DZ: DiZygotic twins, B: Brother, US: Unrelated Speaker.*

Overall, the AUC findings suggest that important information is contained in pause fillers, which can lead to reasonably accurate results. As part of the RF computation process, we have also inspected the RF importance scores for the computations on the 54 speakers. There was no discernible pattern on some features having consistently high importance scores across all developed RF models.

5 Discussion

We investigated the potential of using pause fillers extracted from conversational speech for forensic speaker verification. We built upon our recent previous study's findings [10] by employing a supervised learning setup and applying a more rigorous validation process. We found that when provided with a sufficiently large number of tokens (20) for the training of a robust RF classifier, AUC results between 0.71-0.98 could be reached. These results further support previous explorations and improve upon results reported on this dataset [10,30,31], by using a principled supervised learning framework.

Frequency-related patterns are in general more robust than amplitude patterns which are sensitive to microphone placement, and hence more difficult to control in practical settings [5,6]. F0 estimation is often used as a precursor to the computation of more advanced acoustic measures [5,6,21,22,32]. We had previously shown that SWIPE and NDF are probably the best known F0 estimation algorithms for the sustained vowel /a/ [15]; since pause fillers are phenomenologically like sustained vowels [10], our conjecture was that these two F0 estimation algorithms should be very accurate in the present study's setting. We used NDF because in our experience it works better on short signals [15]. Nevertheless, future studies will need to verify this hypothesis by simultaneously obtaining an external objective measure of F0 using electroglottography during conversational speech, and focusing on the pause filler segments. Moreover, future work on pause fillers should validate previous findings where an information-fusion based F0 ensemble in sustained vowels performed even better than NDF [15].

We have relied on the inherent strength of the RF classifier to mitigate the well-known problem of the *curse of dimensionality* [33]. This refers to the insufficient population of the feature space with limited data, and hence calls for the use of techniques to reduce the size of the design matrix by selecting a subset of features, or using manifold embedding techniques to transform the feature space [33]. RF are generally very robust against the inclusion of redundant and noisy features, although there are key insights to be obtained by determining a compact feature subset, in terms of its interpretability, as can be seen in related applications [34,35]. There were no features which were consistently ranked with the highest importance scores for the different RF models developed; this could indicate the presence of multiple Markov blankets, or more likely reflect the limited number of tokens comprising one of the classes in this highly unbalanced problem. This topic is an active area of research in machine learning, and it needs to be further explored in this particular application to gain further insight.

The supervised learning scheme used in this study falls under the remit of speaker *verification*: given a token, we probabilistically assessed whether this token would belong to a certain speaker or not. We remark this is a considerably simpler problem compared to speaker *recognition*, where potentially a token could be used to identify a person out of a large pool of candidates. For the forensic phonetic application investigated here this is a promising first step which is meaningful in practice. For example, pause fillers could be extracted from a query (unknown) voice sample which may be available from ransom demand from the perpetrator; also an experienced interviewer may be able to elicit pause fillers during conversational speech with a suspect, which would be used to compare against the query sample and probabilistically determine whether the suspect and the perpetrator are the same person. In future work we aim to investigate how the present study's findings could be generalized across wider cohorts and larger sample sizes, and perhaps also tackle this problem from a more general, speaker recognition perspective.

The number of samples used in this study was a major limiting factor in doing full-scale comprehensive supervised learning explorations: there were few speakers with more than 20 extracted tokens. Nevertheless, we demonstrated that promising findings could be delivered even under these objective constraints. We envisage that in a practical setting longer-term interviews would be needed to elicit a greater number of pause fillers. Although there were four different groups represented in the studied cohort (MZ twins, DZ twins, brothers, unrelated speakers), the number of speakers in each group does not provide sufficient statistical power to enable direct group comparisons in a principled supervised learning setup. Moreover, this study only focused on a group of male Spanish native speakers; these findings would need to be verified in other cohorts including females and speakers with different native languages.

The findings reported in this study support the argument that pause fillers contain useful information from a forensic perspective. It is likely that combining information extracted from pause fillers, continuous speech, and additional clues which may be available in a forensic setting, would offer a more comprehensive framework upon which better-informed, accurate decision could be made.

Acknowledgements

This work was supported by the Arts and Humanities Research Council [grant number AH/M003396/1], and the Plan Nacional de I + D + i, Ministry of Economic Affairs and Competitiveness of Spain [grant numbers TEC2012-38630-C04-01, TEC2012-38630-C04-04, and TEC2016-77791-C4-4-R].

Conflicts of interest

We have no conflict of interest.

References

[1] J.P. Campbell: "Speaker recognition: A tutorial", *Proceedings of the IEEE*, Vol. 85(9), pp. 1437–62, 1997

[2] J.H.L. Hansen, T. Hasan: "Speaker recognition by machines and humans: a tutorial review", *IEEE Signal Processing Magazine*, Vol. 32 (6), pp. 74-99, 2015

[3] P. Rose, *Forensic speaker identification*, CRC Press, London, 2002

[4] A. Neustein, H.A. Patil (eds). *Forensic speaker recognition: law enforecement and counter-terrorism*, Springer, 2012

[5] Titze, I.R. (2000). *Principles of Voice Production*, National Center for Voice and Speech, Iowa City, US, 2nd ed.

[6] A. Tsanas, *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning*, D.Phil. (Ph.D) thesis, University of Oxford, UK, 2012

[7] A. Jain., S. Phrbhakar, S. Pankanti, "On the similarity of identical twin fingerprints", *Pattern Recognition*, Vol. 35 (11), pp. 2653-2663, 2002

[8] E. San Segundo: "Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic

analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics" (Thesis abstract). *International Journal of Speech Language and the Law*, Vol. 22 (2), pp. 249-253, 2015

[9] H.J. Künzel, Some general phonetic and forensic aspects of speaking tempo, *International Journal of Speech, Language and the Law*, Vol. 4(1), pp. 48-83, 1997

[10] E. San Segundo, A. Tsanas, P. Gómez-Vilda: "Euclidean distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics", *Forensic Science International*, Vol. 270, pp. 25-38, 2017

[11] M. Christensen, A. Jakobsson, *Multi-pitch estimation*, Morgan and Claypool, San Rafael, CA 94903 USA, 1-6, 2009

[12] R.M. Roark, "Frequency and Voice: perspectives in the time domain," *Journal of Voice*, Vol. 20, pp. 325-354 2006

[13] D. Talkin, "A robust algorithm for pitch tracking," chapter 14 in *Speech coding and synthesis* (Eds. W.B. Kleijn and K.K. Paliwal), Elsevier Science B.V., Philadelphia PA 19103-2879 USA, 495-518, 1995

[14] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of sampled signal", *IFA Proceedings* 17, 97-110, 1993

[15] A. Tsanas, M. Zañartu, M.A. Little, C. Fox, L.O. Ramig, G.D. Clifford: "Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive Kalman filtering", *Journal of the Acoustical Society of America*, Vol. 135, pp. 2885-2901, 2014

[16] H. Kawahara, A. de Cheveigne, H. Banno, T. Takahashi, T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," *Interspeech*, 537-540, Lisbon, Portugal (2005)

[17] A. Camacho, J.G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *Journal of the Acoustical Society of America*, Vol. 124, pp. 1638-1652, 2008

[18] E. San Segundo, "A phonetic corpus of Spanish male twins and siblings: Corpus design and forensic application", *Procedia - Social and Behavioral Sciences* Vol. 95, pp. 59-67, 2013

[19] F. Nolan, K. McDougall, G. de Jong, T. Hudson, The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research, International *Journal of Speech Language and the Law*, Vol. 16(1), pp. 31-57, 2009

[20] G.S. Morrison, P. Rose, C. Zhang, Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice, *Australian Journal of Forensic Sciences*, Vol. 44(2), pp. 155-167, 2012

[21] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig: "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity", *Journal of the Royal Society Interface*, Vol. 8, pp. 842-855, 2011

[22] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig: "New nonlinear markers and insights into speech signal degradation for effective tracking of Parkinson's disease symptom severity", *International Symposium on Nonlinear Theory and its Applications (NOLTA)*, pp. 457-460, Krakow, Poland, 5-8 September 2010

[23] A. Tsanas: "Acoustic analysis toolkit for biomedical speech signal processing: concepts and algorithms", *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pp. 37-40, Florence, Italy, 16-18 December 2013

[24] VOICEBOX, Speech Processing Toolbox for Matlab, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 2006

[25] A. Tsanas, M.A. Little, P.E. McSharry: "A methodology for the analysis of medical data", in *Handbook of Systems and Complexity in Health*, Eds. J.P. Sturmberg, and C.M. Martin, Springer, pp. 113-125, 2013

[26] G.J. Meyer, S.E. Finn, L.D. Eyde, G.G. Kay, K.L. Moreland, R.R. Dies, E.J. Eisman, T.W. Kubiszyn, G.M. Reed, "Psychological testing and psychological assessment: a review of evidence and issues," *American Psychologist*, Vol. 56, pp. 128-165, 2001

[27] L. Breiman, "Random forests," *Machine learning*, Vol. 45, pp. 5-32, 2001

[28] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer, 2nd ed., 2009

[29] A. Webb, Statistical Pattern Recognition, John Wiley and Sons Ltd, 2002

[30] E. San Segundo, P. Gómez-Vilda, "Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings", *Language and Law / Linguagem e Direito*, 1(2), 22-41.

[31] E. San Segundo, H.J. Künzel: Automatic speaker recognition of Spanish siblings: (monozygotic and dizygotic) twins and non-twin brothers, *Loquens*, 2(2), e021, 2015

[32] P. Gómez-Vilda, R. Fernández-Baillo, A. Nieto, F. Díaz, F.J. Fernández-Camacho, V. Rodellar, A. Alvarez, R. Martínez, "Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters", *Journal of Voice*, Vol. 21(4), pp. 450-476, 2007

[33] I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh (Eds.), *Feature Extraction: Foundations and Applications*, Springer, 2006

[34] A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig: "Robust parsimonious selection of dysphonia measures for telemonitoring of Parkinson's disease symptom severity", *7th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pp. 169-172, Florence, Italy, 25-27 August 2011

[35] A. Tsanas, M.A. Little, C. Fox, L.O. Ramig: "Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 22, pp. 181-190, 2014