2013 8th International Conference on Information Technology in Asia (CITA)

Voice Source Waveforms for Utterance Level Speaker Identification using Support Vector Machines

David Vandyke[1], Michael Wagner[1,2], Roland Goecke[1,2]
[1]University of Canberra, Australia
[2]Australian National University, Australia
{david.vandyke, michael.wagner}@canberra.edu.au, roland.goecke@ieee.org

***Abstract*—The voice source waveform generated by the periodic motion of the vocal folds during voiced speech remains to be fully utilised in automatic speaker recognition systems. We perform closed-set speaker identification experiments on the YOHO speech corpus with the aim of continuing our investigation into the level of speaker discriminatory information present in a data driven parameterisation of the voice-source waveform obtained by closed-phase inverse filtering. Discriminatory modelling using support-vector-machines resulted in utterance level correct identification rates of 85.3% when using a multi-class model, and 72.5% when using a binary, one-against-all regression model, each on cohorts of 20 speakers respectively. These results compare well with other speaker identification experiments in the literature employing features derived from the voice source waveform, and are positive when observed under the hypothesis that they should be complementary to the common magnitude spectral parameters (mel-cepstra).**

Keywords*—Glottal Waveform, Voice Source, Speaker Identification

I. Introduction

Advances in the performance of automatic speaker recognition systems over the last five years have come from the improvements in adjusting to nuisance variations such as channel and/or microphone distortions and background variability. In particular joint factor analysis (JFA) [1] has provided a theoretic framework for attempting to separate out these problem variations from the intrinsic variation of a speakers voice. As demonstrated in systems submitted in recent years NIST Speaker Recognition Evaluations [2], these baseline systems continue to model predominately only magnitude spectral information (mel-cepstra) . These features relate most strongly to the speakers vocal tract physiology and articulator use.

The linear source-filter theory of speech production [3] says that an air pressure wave, produced by the diaphragm, is shaped by the vibratory motion of the vocal folds (creating the voice-source waveform), and modulated by the vocal tract, which is modelled as a time-invariant linear filter. Whilst the largest source of variation between speakers voices originates from their different vocal tract and articulator physiologies, speaker identifying characteristics have been found in several parameterisations of the voice-source waveform [4], [5], [6], [7]. Further, this voice-source information should be complementary (based on the physiology and speech production theory) to standard spectral features, and when combined with mel-frequency cepstral coefficients (MFCC) it has indeed been shown to raise recognition accuracy above the baseline system performance [6], [8].

Despite these results the source-waveform remains to be regularly and efficiently utilised in automatic speaker recognition systems. We continue here our investigation of a data driven parameterisation of the voice-source waveform first proposed for speech synthesis [9]. Our parameterisation of the voice-source waveform is derived from closed-phase inverse filtering and is based on the deterministic component of the Deterministic plus Stochastic model (DSM) [9], as we describe in Section III.

II. Relation to Prior Work & Limitations

This work continues the examination of discriminatory separation by support vector machines of the source-frame features (describe below in Section III) which was begun in the proceedings of Speech Science & Technology 2012 [7]. In this earlier exploration, identification results where reported on a per source-frame level, obtaining identification rates replicated for convenience below in Table I.

TABLE I. *Best source-frame ID rates for each cohort size (over different Principal Component dimensions). The number of source-frames refers to the amount used per speaker for training.*

Cohort Size	No. Source-Frames	ID Rate
5	300	70.8
10	500	64.3
15	800	65.0
20	1000	64.7

Using multi-class Support Vector Machines (SVM), cross-validation experiments were performed here [7], with training and testing partitions selected randomly and in the process removing the inter-session variation of the YOHO database [10]. It was left as a hypothesis that these source-frame level correct identification rates would translate into strong utterance level systems. Indeed this is the behaviour observed in moving from the micro level (speech frame, visual frame) to the macro level (utterance recording, visual sequence) in the majority of automatic recognition systems, and logically so for any non-trivial distributions of micro level recognition rates.

We extend this initial work by introducing a temporal divide between training and testing speech, and reporting correct identification rates at the utterance (.wav file) level. These

978-1-4799-1092-2/13/$31.00 ©2013 IEEE

experiments employ discriminative modelling to conclude our investigation into the speaker-identity-related information content of the source-frame features.

There are benefits to be had by developing a generative model of these features, including reducing the amount of speech data required for enrolment and testing, as well as providing a stronger theoretical framework which may be better able to relate back to the physical motions of the vocal folds shaping the source waveform. This work remains ongoing.

Limitations of this study to be build upon in future work include (a) testing our hypothesis that these features do provide complementary information to MFCC based systems, (b) increasing the number of speakers in the closed cohort, and (c) moving from closed-set identification experiments to open-set identification and then speaker verification. The verification paradigm would require measuring not just model and probe similarity but also typicality, that is how the test probe fits within the speaker population variability. This requires developing and incorporating background or population models of the voice-source features, fitting into the Bayesian likelihood ratio theory [11].

III. Glottal Source-Frames

We begin by describing how the pitch-synchronous glottal waveforms are extracted from the speech signal. These glottal, or voice-source, waveforms are obtained by inverse filtering and these signals are normalised in both pitch and amplitude. We refer to these features as source-frames. They are described in detail now in Section III.

A. Feature extraction

The speech signal (utterance) is segmented into 25ms frames with a 5ms shift. We aim to perform closed-phase linear predictive analysis, where the assumptions of the source-filter model of speech production [3], [12] are most valid due to the maximal separation of interaction between the vocal tract and vocal folds.

To then determine the instants of glottal closure over each voiced pitch period we determine the autocorrelation linear predictors over all frames, before employing the overlap-add method to construct the linear predication residual over the entire utterance. Extrema of this residual signal are then located within containers demarcated by an averaged version of the speech signal, which give the locations of both glottal closure and glottal opening instants. This glottal instant detection algorithm is described in detail in [13], and has been shown to be superior to alternative glottal instant detection methods such as DYPSA [14]. It also has the advantage of estimating the point of glottal opening, when commonly it is simply assumed that a fixed portion of each pitch period is closed.

Closed-phase linear prediction (autoregession solution) is then performed in order to determine as accurately as possible the linear filter representing the vocal tract at each moment of voicing of the speech signal.

These all-pole filters are then used to inverse filter the speech signal, determining the pitch-synchronous error signal representing the voice-source waveform. All consecutive

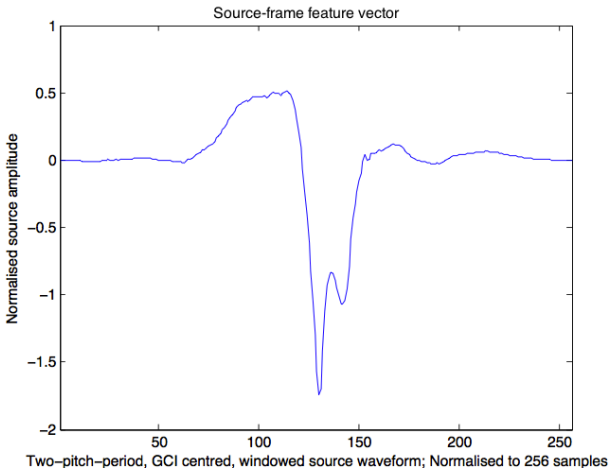

Fig. 1. *Single source-frame vector extracted by inverse filtering. A single period of the source waveform is evident centrally.*

two pitch-periods, where each pitch-period begins and ends on a glottal closure instant (GCI), are then gathered. This creates a set of vectors of different lengths (the length depends upon the fundamental frequency of the pitch-period region they are extracted from). These are then normalised in both x (length/pitch) and y (amplitude/voice-waveform-energy) dimensions. This normalisation enables the signals to be analysed statistically with discriminative (and generative) models. It retains information pertaining to the overall periodic motion of the vocal folds at the cost of shimmer and jitter information.

Due to the modelling of the radiation of the speech waveform at the lips in the source-filter model of speech production as a single pole filter, the obtained voice-source waveform represents the first derivative of the volume-velocity waveform of air produced by the diaphragm/lungs and modulated by the vocal tract filter [12]. This is the driving function of human speech.

We refer to these normalised, double pitch-period, GCI centred voice-source waveforms as source-frames. The amplitude scaling is done by normalising by the standard deviation of the source-frame data, and the frame length is mapped to a constant number of samples by interpolation or decimation as necessary, along with the required anti-aliasing, low pass filtering. Finally the source-frames are Hamming windowed to emphasise the shape of the signal around the central glottal-closure instant. As such each source-frame vector contains information about a single pitch-period. These features are based upon the deterministic component of the Deterministic plus Stochastic model (DSM) proposed by Drugman et al. [4], [5], [9]. One such source-frame feature vector is shown in Figure 1.

IV. Support-Vector-Machine Modelling Experiments

We investigate the ability of Support Vector Machines (SVM) to discriminate between speakers based on these source-frame features. We examine the ability of both multi-class SVM and single class SVM regression to separate speakers in closed-set speaker identification experiments.

We use male speakers from the YOHO American speech corpus [10], containing microphone speech sampled at 8000 samples/second and stored by single channel PCM compression. YOHO contains multisession recordings and in all experiments training and testing speech is taken from different recording sessions. YOHO, whilst non-challenging for current baseline automatic speaker recognition systems (such as factor analysis [1] and even GMM-UBM [11]), permits the voice-source waveform to be initially examined in the absence of channel and noise variations that can impact negatively on the linear prediction and inverse filtering processes. This is the approach of several significant papers in their preliminary investigations of the voice-source [6], [8], [15], [16].

Source-frames are normalised to $N = 256$ samples. The dimensionality is an issue for computation considerations, and Principal Component Analysis (PCA) is used purely for dimension reduction. A disjoint set of 10 male speakers from the YOHO dataset, not used in any identification experiments as clients or impostors, were selected and source-frames extracted from all of their enrol data. We shall refer to this set as the background set. Using this background set a basis of principal components was determined against which experimental features could be projected into for dimension reduction.

The percentage of variation retained from the background sets' data by increasing the number of principal components is shown in Figure 2. We see that more than 90% of the variation within the data is covered by retaining the first 50 principal components. This is expected as the windowing of the source-frame produces many near zero samples shared at each end of all source-frame vectors.

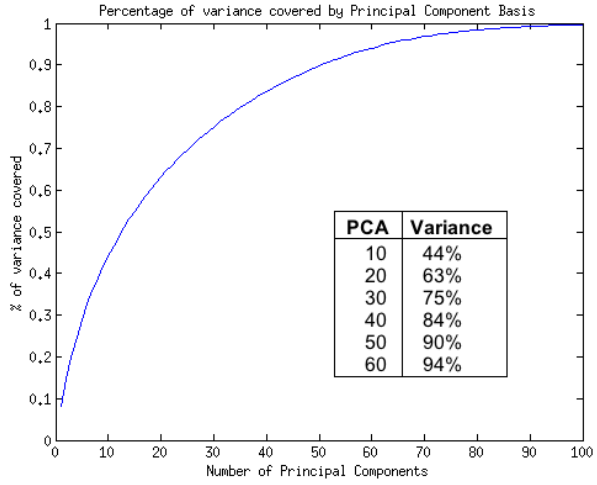

Fig. 2. *Variance of Source-Frame data covered by principal component basis.*

The number of principal components used was varied from 30 to 60 in the multi-class SVM experiments detailed below (Sub-Section IV-A). It can be seen from the statistically non-significant increase in correct identification rates on the multi-class SVM training as the PCA dimension was increased (see Table II) that there is also a large amount of noise variation in the source frame data not related to speaker identity, and that retaining larger numbers of principal components is not beneficial to recognition accuracy.

The LIBSVM package [17] was used to implement all SVM experiments. Cross validation on a further disjoint set of male YOHO speakers was used to determine the optimal kernel function (radial basis function) and kernel parameters (gamma = 0.0325, cost = 32). Data was not scaled further than the prosody normalisation step during feature extraction. Identification rates on a per utterance (and per source-frame) level are given below for investigations of multi-class SVM (Section IV-A) and regression SVM (Section IV-B).

A. Multi-Class SVM Modelling

Multi-class support vector modelling is explored with closed cohorts of $5, 10, 15, 20$ and 30 speakers. For each experiment hyperplanes (SVM models) are trained on speech coming only from the 'Enrol' partition of YOHO, and test probes are taken only from the 'Verify' partition. Identification rates are given at the source-frame level and the utterance level. For all source-frames, from all probe utterances, probabilities measuring class membership likelihood are output. Frame level identification rates are based upon assignment of source-frames to a speaker/class whose model generates the maximal probability. Utterance level scores are determined by calculating the mean probability value over all source-frames from the utterance. Utterance level identification decisions are then done by assigning the utterance to the model/speaker with the maximum score.

Training and testing source-frame data is projected against the background data PCA basis for dimension reduction. Experiments are performed when retaining $30, 40, 50$ and 60 principal component dimensions. Table II reports average (across the four PCA dimension results) utterance and frame level correct identification rates.

TABLE II. *Summary results: Multi-class SVM.*

Cohort Size	# Source-Frames	Frame %	Utterance %
5	1100	38.3%	71.74%
10	2200	30.9%	86.7%
15	6000	39.4%	89.5%
20	6000	25.6%	85.3%
30	6000	20.5%	80.8%

Identification rates do not evolve in relation to the chance identification rate $(1/\text{CohortSize})$. We believe this is due to the amount of data available for training the SVM model, where over training and under training is likely occurring on either side of 15 speakers. Limitations in available computational power necessitated training on 6000 source-frames for the 20 and 30 speaker cohorts, as done for the 15 speaker group.

Frame level identification rates for each cohort and each PCA dimension size are shown in Figure 3. The influence of the PCA dimension is shown to be minimal.

Utterance level identification rates, again for each cohort and PCA dimension, are shown in Figure 4. The identification rates are promising, especially under the working assumption that the voice-source information is orthogonal to common spectral magnitude features (mel-cepstra).

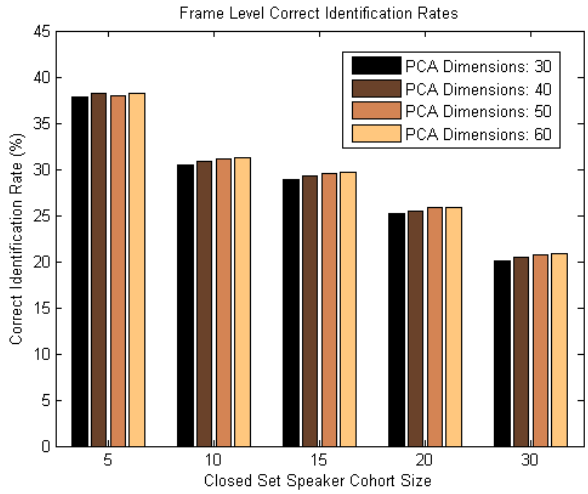

Fig. 3. *Frame level correct identification rates for each closed set speaker size and PCA dimension.*

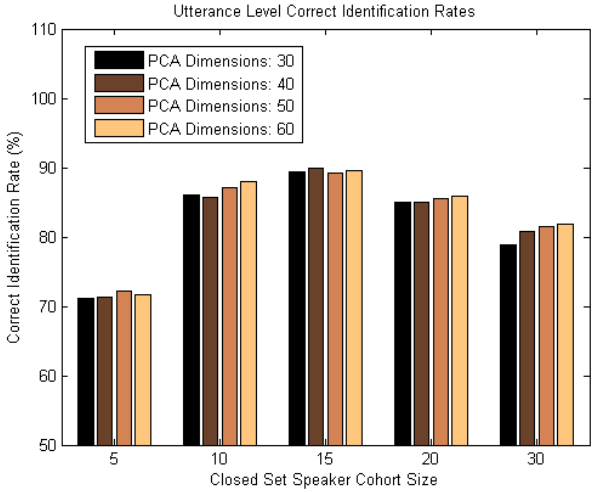

Fig. 4. *Utterance level (.wav file) correct identification rates for each closed set speaker size and PCA dimension.*

Misidentifications are found to be approximately uniform across speakers in all cohorts. Figure 5 demonstrates how the frames from speaker 2's test utterances are assigned when scored against the multi-class SVM model for the size 15 cohort. While the majority are correctly assigned to speaker 2, the misclassifications are approximately uniform (speakers 5 and 6 take a slightly larger amount). This behaviour, shown in Figure 5, is typical of the distribution of misclassifications in all multi-class experiments.

B. Regression SVM Modelling

We also examine the ability of binary SVM regression models in closed set speaker identification experiments on the YOHO corpus. For each speaker within the cohort, a regression SVM model is trained on the pooled training data of all speakers. Training data was assigned the class $+1$ for frames belonging to the speaker whose SVM model is being trained, and -1 for all other speakers present within the training

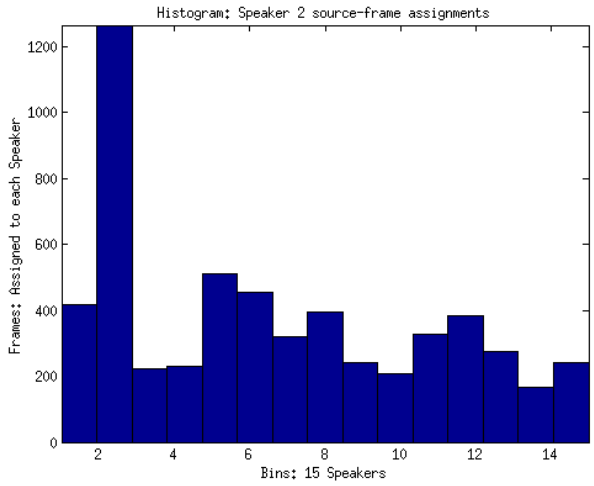

Fig. 5. *Histogram of frame level assignments to each speaker in cohort of size 15. Using multi-class C-SVM model. Probe source-frames all belong to speaker 2. Misclassifications are approximately uniform.*

set. Source-frames presented at testing time are assigned a predicted labeled by the regression model with a value on the continuum between these training class labels $[-1, +1]$.

Closed set speaker identification experiments are performed as follows. In reporting frame level identification rates, source-frames are assigned to the speaker whose regression model output the largest score. The more important utterance level identification rates were determined as follows. The mean of all the regression outputs of all the frames of the test utterance is taken to create an utterance score against the speaker SVM model. The speaker whose model outputs that largest utterance level score is identified as the speaker of the test utterance.

Taking the mean of the frame scores was empirically determined to achieve higher identification rates than other statistics such as maximums or taking the product.

This experimental process was performed for cohorts of size $5, 10, 15$ and 20 speakers. In all regression SVM experiments we retain only the first 30 principal component dimensions, informed by the results of the multi-class SVM experiments which provided strong evidence for the proposal that there is little benefit in retaining more principal component coefficients. The speakers and their training and testing data remained the same for each cohort size, as done for the multi-class SVM experiments.

Figures 6 and 7 show utterance scores for the cohort group of 5 speakers, where there were 184 test utterances (roughly split between speakers). Figure 6 shows the test probes from the 5 speakers scored against the regression model for speaker 1 whilst Figure 7 shows the same utterances scored against the regression model of speaker 2.

A verification style threshold is drawn on each figure along with points demarcating the continuous section of utterances coming from the speaker whose model is being tested against. This threshold line is drawn only to indicate the typical distribution of scores observed in all experiments; we perform speaker identification experiments which make no reference to any such thresholds.

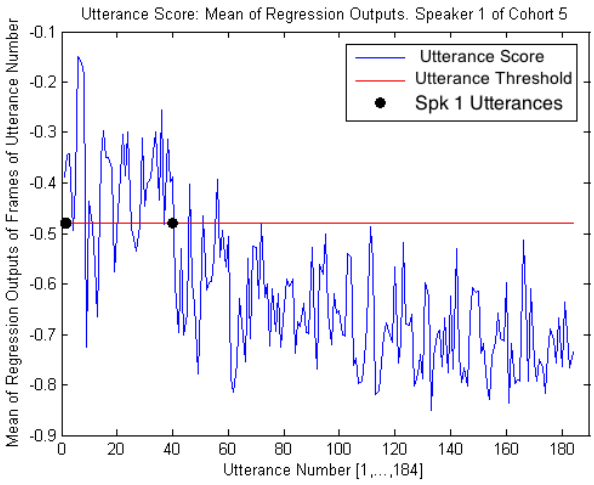

Fig. 6. *Utterance scores; taken as mean of regression output over all frames from each utterance. Test performed against model for speaker 1 of cohort size 5. The first 40 utterances belong to speaker 1; marked by black dots. The horizontal line displays a verification type threshold against which utterance scores could be compared for accept or reject decisions. We perform identification experiments and this threshold is shown only to indicate the typical difference in scores for utterances coming from the same speaker whose model is being tested against ('client'), compared to non 'client' probes.*

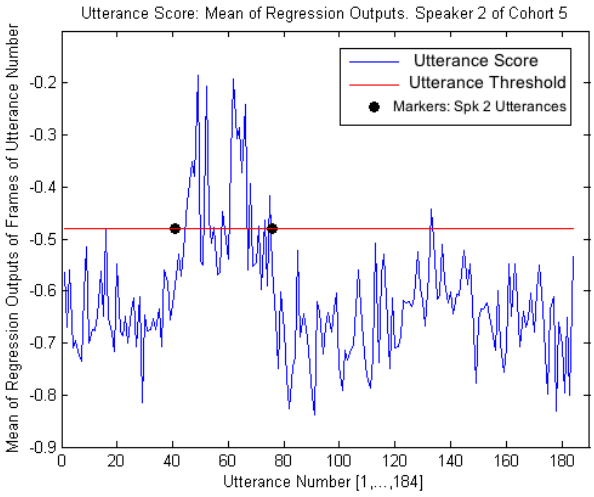

Fig. 7. *Utterance scores; for probes against model for speaker 2 of cohort size 5. Speaker 2's utterances begin at utterance number 41 and run through to utterance number 76.*

Regression SVM correct identification rates are reported in Table III for all cohort sizes.

For the cohort of 20 speakers (the largest used in both experiments), correct identification rates of 85.3% for the multi-class system and 72.5% for the regression system compare well to previous investigations of voice-source features using the YOHO [10] corpus, although it must be noted we use smaller cohort sizes here. Plumpe et al. [6], using a analytic model of the glottal wave based on parameterising its' opening, closing and return phases, obtained a misclassification rate across all of YOHO (averaged over male and females) of 28.6%. Gudnason et al. [8] achieved a misclassification rate on all of YOHO of 36% using a cepstral parameterisation of the spectrum of the voice-source.

TABLE III. *Summary results: mean identification rates using SVM regression. Reported are the mean correct identification rates on a per frame level (column3), and on a per utterance level (column4). The average number of source-frames used per speaker for SVM regression training for each cohort size are given in column2.*

Cohort Size	# Source-Frames per Speaker	Frame %	Utterance %
5	1100	37.4%	90.0%
10	2000	30.4%	89.3%
15	1000	21.7%	80.2%
20	1000	17.8%	72.5%

Note that the number of source-frames available per speaker (shown in Column 2 of Table III) ideally should increase with the cohort size for reliable training of the SVM regressor. We were limited to using the quantities given in Table III due to constraints on computational resources. The maximum cohort size was limited to 20 speakers for similar reasons.

Correct identification rates were reasonably consistent for each individual speaker in all experiments. Figure 8 shows the breakdown of correct identification rates for each speaker of the cohort of size 20.

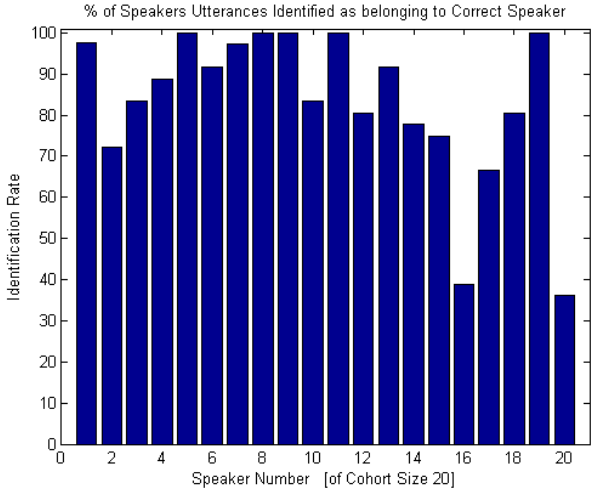

Fig. 8. *For each speaker of the cohort of size 20, the percentage of that speakers utterances correctly identified as belonging to that speaker (that is scoring highest against that speakers regression model) is shown.*

For the majority this rate is above 70%, however for speakers 16 and 20 the performance is significantly below the average for the cohort. Upon inspection these identification rates are strongly correlated with the total number of source-frames extracted from each speakers training utterances. Whilst the training utterances where the same in number for each speaker, the number of voiced pitch-periods in total over these utterances differed for each speaker. This affected the SVM regression model accuracy for certain speakers.

V. Discussion and Future Work

Identification results in both experiments have shown further evidence that the voice-source waveform obtained by inverse filtering of the speech signal contains significant information pertaining to speaker identity. Using a multi-class SVM model 85.3% of test utterances, coming from a different session to those used for training, were correctly identified for the cohort of 20 speakers. Using a binary SVM regression model with the same closed set of 20 speakers, 72.5% of test utterances were correctly identified. These results are similar to previous investigations of the voice-source waveform for speaker identification using the YOHO corpus [6], [8].

The identification rates of the regression model are inferior to the multi-class support-vector model, and there are logical reasons for this. Each single regression model attempts only to differentiate between binary classes, and doesn't model the variations of non-client speakers when training a client model. Such a model structure is better suited to a speaker verification paradigm where accept & reject decisions are required, and not selection from a group. Such a paradigm would require the introduction of some measure of typicality, that is a measure of how the speakers source-features are distributed over the speaker population of interest for the system [11].

To implement such identification systems as explored here, especially on large or open ended cohorts, would required significant computational power. Clients would also be required to give more enrolment speech than usability requirements on time would deem acceptable. These points are acknowledged, however the focus of this voice-source investigation using discriminatory models has been on exploring the identity information content of the source-frame features. To this end our aims have been achieved.

The hypothesis that this identity discriminating information complements common magnitude spectral features shall be explored in future work where the source-frames are modelled generatively. There are several significant reasons for developing a generative model of these features. This would alleviate the requirement for excessive amounts of enrolment speech, also allowing adaptation of distribution models using common Maximum A Posteriori (MAP) methods [18]. Further advantages include employing scoring based on probability measures which are logically more rigorous than distances.

This point particularly holds for the use of such features in a forensic context, which is of interest to our research group, where reporting methodology should be consistent across practitioners and cases at the base level. In particular this means applying a Bayesian framework to update beliefs based upon the presented evidence [19], and this is better adhered to by generative, probabilistic models.

Finally we believe these results further support the hypothesis that data driven models of the voice-source [15], [20] are more useful for speaker recognition than analytic models parameterising the sections of a pitch-period of the voice-source waveform (opening, closing, returning), such as those proposed earlier originally for speech synthesis such as Lijencrants-Fant [21] and Rosenberg [22]. These analytic models of the glottal waveform are suitable for speech synthesis but we believe they do not capture the nuance variations that differentiate speakers.

VI. Acknowledgment

The first author gratefully acknowledges the financial support provided by the Australian government's Australian Postgraduate Award provided to assist his doctoral studies.

References

[1] Patrick Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," in *Technical Report: CRIM*, 2006.

[2] National Institue of Standards and Technology, "Speaker recognition evaluations," [online] http://www.itl.nist.gov/iad/mig//tests/spk/, 2013.

[3] Gunnar Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, 1960.

[4] Thomas Drugman and Thierry Dutoit, "On the potential of glottal signatures for speaker recognition," in *INTERSPEECH*, 2010, pp. 2106–2109.

[5] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech & Language*, vol. 26, no. 1, pp. 20–34, 2012.

[6] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, Sept. 1999.

[7] D. Vandyke, M. Wagner, R. Goecke, and G. Chetty, "Speaker identification using glottal-source waveforms and support-vector-machine modelling," in *Proceedings of Speech Science and Technology*, 2012, pp. 49–52.

[8] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2008, p. 4821.

[9] Thomas Drugman, Geoffrey Wilfart, and Thierry Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *INTERSPEECH*, 2009, pp. 1779–1782.

[10] J.P Campbell Jr, "Testing with the yoho cd-rom voice verification corpus," in *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, may 1995, vol. 1, pp. 341 –344.

[11] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, pp. 19–41.

[12] J. Markel, "Digital inverse filtering-a new tool for formant trajectory estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 2, pp. 129–137, jun 1972.

[13] Thomas Drugman and Thierry Dutoit, "Glottal closure and opening instant detection from speech signals," in *INTERSPEECH*, 2009, pp. 2891–2894.

[14] Patrick A. Naylor, Anastasis Kounoudes, Jon Gudnason, and Mike Brookes, "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, jan 2007.

[15] Jon Gudnason, Mark R. P. Thomas, Daniel P. W. Ellis, and Patrick A. Naylor, "Data-driven voice source waveform analysis and synthesis," *Speech Communication*, vol. 54, no. 2, pp. 199–211, 2012.

[16] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 968–981, march 2012.

[17] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27, 2011.

[18] J.-L. Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291 –298, apr 1994.

[19] Geoffrey Stewart Morrison, "Forensic voice comparison and the paradigm shift," *Science and Justice*, vol. 49, no. 4, pp. 298–308, 2009.

[20] M.R.P. Thomas, J. Gudnason, and P.A. Naylor, "Data-driven voice soruce waveform modelling," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, april 2009, pp. 3965 –3968.

[21] G Fant, J Liljencrants, and Q Lin, "A four-parameter model of glottal flow," *Speech Transmission Laboratory QPSR*, vol. 4, no. 4, pp. 1–13, 1985.

[22] S Rosenberg, "Glottal pulse shape and vowel quality," in *Journal of the Acoustic Society of America, 49, 2*, 1970, pp. 583–590.