The Voice Source in Forensic-Voice-Comparison: a Likelihood-Ratio based Investigation with the Challenging YAFM Database

David Vandyke[1], *Phil Rose, and Michael Wagner*[1, 2]
[1]*Human-Centered Computing Lab, University of Canberra, Australia*
[2]*College of Engineering and Computer Science, Australian National University*
{david.vandyke | michael.wagner}@canberra.edu.au
philjohn.rose@gmail.com

The glottal volume-velocity (voice-source) waveform, describing the airflow from the lungs shaped by the vibratory motion of the speakers' vocal folds and input to the vocal tract during the speech production process, has been shown to contain speaker discriminatory information (Vandyke, 2012 & references therewith).

Despite there having been various parameterisations proposed within automatic speaker recognition (AS_kR), the voice-source remains to be regularly employed as a information source. This is also the case in forensic voice comparison (FVC), where statistical measures have been less common still, although vocal fold vibration patterns have been used to inform subjective opinions of voice quality outside of the likelihood-ratio (LR) FVC paradigm. When employed, AS_kR has achieved better results with data-driven representations of the voice source (Vandyke, 2012 & references therewith). Fitting theoretic models (Rosenberg, Liljencrants-Fant) to observed voice-source data tends to only capture an average pattern and is better suited for use in speech synthesis, rather than for speaker differentiation tasks.

Investigations into the discriminative ability of the voice-source within the FVC-LR paradigm have to date found little to no benefit. The commercial GLOTTEX® package has been used by (Enzinger, 2012), with catastrophic fusion (decreases in both accuracy:C_{llr} and precision:95% CI) with a MFCC-GMM-UBM system observed with the features it provided. GLOTTEX® measures some descriptive statistics of the glottal waveform, and we suspect misses small idiosyncrasies that are key to informing identity. We introduce here a data-driven representation of the voice-source that is obtained via closed-phase inverse filtering of voiced pitch-periods. We also introduce the forensically challenging YAFM database.

Forensically challenging YAFM database & Experimental study

For our investigation into the use of a data-driven parameterization of the voice-source waveform as a useful feature for FVC we used the Young Australian Female Map-task (YAFM) database (Rose, 2011). YAFM consists of 26 female speakers with Australian English as L1, all within their twenties and all from the same university-attending social group. There exists 2 sessions for the majority of speakers, where each time the same guide (also a member of the social group) lends the participant through a verbal repetition of place names and locations, and a "map-task". The map used was synthetic and place names and the map-task route where designed to elicit several tokens for a range of phonemes. The guide and participant speech was manually diarised using Praat, leaving on average approximately 1.5 minutes of participant speech per session.

General convergence of speech signifying group membership is believed to be occurring, and several speakers possess significant vocal creak. Speakers are very similar to the authors' ears, and we believe YAFM is a forensically challenging database. Results of human listening experiments will be presented at the conference and our hypothesis is that chance recognition rates will be observed.

A mel-cepstral baseline experiment was performed using 12 MFCC + log energy with first order deltas, feature warped and then modeled with a 32 mixture GMM-UBM. The UBM was trained on features from all 54 female speakers from the ANDOSL[1] corpus, which contained microphone recordings of 200 phonetically varied sentences. Intersession variation existed via MAP adapting suspect models from the background using YAFM session 1 data. Offender speech was taken from the 2nd YAFM session.
The voice-source features, termed source-frames, are obtained by closed-phase inverse filtering and represent the derivative of the voice-source waveform (derivative due to modelling of radiation at the lips in the source-filter theory of speech production, but referred to as the voice-source for simplicity from here). Sections of the voice-source corresponding to voiced pitch-periods are then framed and prosody normalized; a scaling is performed in both length (pitch) and amplitude (energy in source wave). The feature extraction process is outlined in detail in (Vandyke, 2012).
Source-frames are modelled by 3 non-overlapping piece-wise polynomial regressions, and the polynomial coefficients concatenated and modelled via a GMM-UBM (with source-frames extracted from the ANDOSL background corpus for the UBM).
MFCC and source-frame scores were fused via logistic regression using the FOCAL toolkit[2]. Benefits are shown from incorporating the voice-source information in this parameterisation via a significant decrease in log likelihood-ratio cost C_{llr}. Further work remains to also include a confidence interval on LLR values.

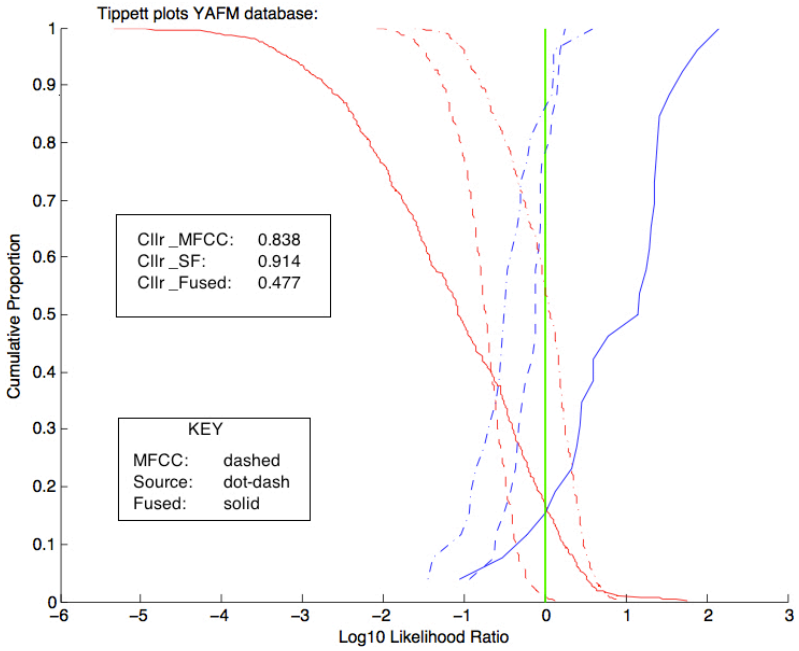

Figure 1 Tippet plots for the MFCC, voice-source and score fusion LLRs.

References

Enzinger, E., Zhang, C. & Morrison, G. S. (2012). Voice source features for forensic voice comparison – an evaluation of the GLOTTEX software package. *Proceedings of Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, pp. 78–85n.

Rose, P. (2011). Young Australian Female Map-Task (YAFM) database. *Available upon request.* Collected as part of a forensic voice comparison project at the ANU school of language studies & via the *Australian Research Council Discovery Grant* No. DP0774115.

Vandyke, D., Wagner, M. and Goecke, R. (2012). Speaker Identification Using Glottal-Source Waveforms and Support-Vector-Machine Modelling. *Proceedings of Speech Science and Technology*, Sydney, 3-6 December, pp 49-52.

[1] http://andosl.rsise.anu.edu.au/andosl [2] https://sites.google.com/site/nikobrummer/focal