The reassigned spectrogram as a tool for voice identification

Sean A. Fulop and Sandra Ferrari Disner

Dept. of Linguistics, California State University Fresno
sfulop@csufresno.edu

ABSTRACT

A precise imaging scheme, the reassigned (or time-corrected instantaneous frequency) spectrogram, holds out considerable promise for identifying the speaker of an utterance. Unlike conventional spectrograms, reassigned spectrograms can display a few tens of milliseconds of phonation in great detail, without blurring in the time or frequency domains, and they are also impervious to many forms of noise or channel contamination. They are thus able to reveal some unique time-frequency features of an individual's phonatory process. While further testing is needed to establish evaluation criteria and confidence estimates, it is encouraging to see how readily pairs of reassigned spectrograms can be matched in the set illustrated herein. At the very least, such images can augment the techniques that are currently in use for speaker identification and verification.

Keywords: Forensic phonetics, Speech processing, Speaker identification

1. INTRODUCTION

While conventional spectrograms have, for decades, been employed in the forensic context as a means of inferring speaker identity, aspects of the aural-spectrographic method have been challenged by a number of speech scientists [18, 7]. Early reports of 99% accuracy in identifying speakers from just four words [8] have not been replicated, and an early report commissioned by the F.B.I. [3] warned that the assumption of interspeaker variability exceeding intraspeaker variability was not adequately supported by scientific theory and data. A compromise suggested by Ladefoged [10] is to look for acoustic features in the spectrogram that appear to reflect the speaker's physical characteristics, and to express the results in terms of the likelihood ratio [18].

The *reassigned spectrogram,* a lesser-known method of imaging the time-frequency spectral information contained in a signal [12, 5], offers some distinct advantages over the conventional spectrogram. Reassigned spectrograms are able to show the instantaneous frequencies of signal components as well as the occurrence of impulses with increased precision compared to conventional spectrograms (i.e. the magnitude of the short-time Fourier transform). Computed from the partial phase derivatives (with respect to time and frequency) of the short-time Fourier transform, such spectrograms can reveal unique features of an individual's phonatory process by "zooming in" on a few glottal pulsations during a vowel. These images can thus highlight the individuating information in the signal and exclude most linguistic information. To date, no attempt has been made to apply this newer technology to the problems of speaker identification or verification.

As far as automatic speaker verification is concerned, state of the art techniques [16] use a number of parameters drawn from acoustic analysis of the voice such as mel-frequency cepstral coefficients, which must then be employed in a statistical model of voices that is trained from numerous utterances before any verification can be reliably performed. Such methods behave as a "black box," providing an output confidence metric, but no single representation that can be analyzed. A recent paper [19] reports an equal error rate (an equal rate of false matches and missed matches) of 14% using the commonplace mel-frequency cepstral coefficients, and 10.5% using a new combination of methods proposed therein, on a database of 149 male speakers. Clearly there is still room for improvement.

In a paper that presaged their 1976 report to the F.B.I., Bolt et al. [2] envisioned "a device with a display emphasizing those sound features that are most dependent on the speaker. The patterns could then be judged with greater confidence by human experts." A simple idea for achieving such a device has been suggested by Plumpe et al. [14], inspired by movies of vocal fold vibration which "show large variations in the movement of the vocal folds from one individual to another." A more complete, though for the moment qualitative, method of observing individuating features of the glottal pulsation is presented in this paper. The precise degree of confidence with which persons can be distinguished by this means remains the subject of current research, but it is by now clear that the reassigned spectrogram can provide a wealth of new information that can augment, or perhaps in the future even replace, existing voice identification techniques.

2. THE REASSIGNED SPECTROGRAM

The *time-frequency analysis* of a signal refers generally to a three-dimensional representation showing the passage of time on one axis, the range of frequencies on a second axis, and the amplitude found in each time-frequency intersection (or *cell* in the digital domain) on a third axis, which is most frequently shown by linking the values to a grayscale. The archetype of this kind of representation is the *spectrogram* which was originally developed using analog filters [15], but which was eventually represented mathematically as the squared magnitude of the short-time Fourier transform [11].

Over the past few decades, various efforts have been undertaken within the signal processing community to develop a new kind of spectrogram that would show the time course of the *instantaneous frequencies* [4] of the components in a complex signal, instead of an overall energy distribution. To make sense out of this, one must consider the signal as a superposition, not of sine waves as in Fourier theory, but rather of amplitude and/or frequency modulated sine waves (often called *line components*). Based in large part on signal processing theory published by Rihaczek [17], a new spectrogram that showed the instantaneous frequencies of line components instead of the short-time Fourier transform was first described by Kodera et al. [9]. More recently, however, the idea has been the subject of numerous publications (e.g. [12, 1]), and this has led to the adoption of the *reassigned spectrogram* by a handful of applied researchers. An historical and technical review of the reassigned (also called *time-corrected instantaneous frequency*) spectrogram, complete with a variety of algorithms for its computation, has recently appeared [5]. Figure 1 shows the increased precision of the technique over the conventional spectrogram, and exemplifies the possibility of close-up images of the resonances emitted during the glottal pulsations.

Recent developments to further this technique [13, 6] have shown that by applying numerical thresholds derived from higher-order partial derivatives of the short-time Fourier transform phase (complex argument), it is possible to eliminate most noise and computational artifacts and thereby present only the excited line components and/or impulses in the plot. It is this kind of post-processed reassigned spectrogram (v. Fig. 1), which may be described succinctly as "pruned" to show either components or impulses (or both), that has been found most useful as a biometric image.

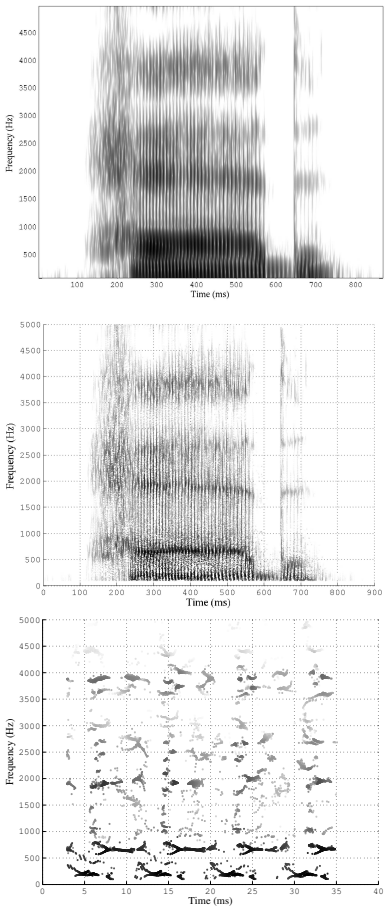

Figure 1: A conventional spectrogram of the English word *had* (top panel, 5.8 ms frames and 0.39 ms step), a reassigned spectrogram of the same (middle, 0.78 ms step), and a pruned reassigned spectrogram (see text) showing just the line components in a few glottal pulsations during the vowel (bottom, 78 μs step).

3. OBTAINING AN INDIVIDUATING IMAGE

In order to accurately image the speaker-specific details of the glottal excitation of vocal tract resonance frequencies, it is best to focus on a very short span, just a few pulsations of the vocal cords, for then the detailed excitation of head resonances by the glottal pulsations can be most clearly observed. After pruning the resulting analyses using the mentioned partial derivative threshold techniques, the specific pattern of impulsive resonance excitation present in the image turns out to be highly distinctive across even similar-sounding speakers uttering the same vowel in the same word.

When the examination of repetitions of the same utterance by the same speaker is undertaken, however, the resulting reassigned spectrograms are

found to be not nearly so variable. There is a certain amount of variation apparent for a typical speaker, depending on the degree to which the voice qualities of the repetitions match, but in every case so far examined it has been found that the vowel that was repeated by the same speaker using a similar voice quality could be visually recognized from among others spoken by different speakers—and all on the basis of just a single image showing a few glottal impulses. Figure 2 gives a sampling of four matched pairs of prints, one from each of four speakers (chosen at random from a database) repeating the same utterance. It can be seen that here we are much closer to the desideratum of Bolt et al., [2] viz. an image that is easily compared to others of its sort, and which highlights the peculiar acoustic attributes of each person's phonation process (although the filtering of this signal by the head is still evident, and may indeed add to the distinguishability owing to interspeaker differences in formant frequencies). What is more, the resulting images are impervious to many common forms of noise or channel contamination thanks to the pruning [6].

The nature of this report is preliminary, and the intent is simply to demonstrate the apparent potential of the reassigned spectrogram as a new speaker identification tool. It will be a much larger undertaking to formally evaluate the efficacy of this approach for forensic or security identification applications, as well as to pursue the automation of the comparison of the images, thereby providing confidence estimates in the manner of computer image matching. The relatively sparse plots shown in the figures give hope that automatic comparison by image matching techniques (such as support vector machines) can be achieved. In advance of these evaluations and developments, however, it is apparent at present that the reassigned spectrogram with higher-order phase derivative thresholding (pruning) could prove to be a very effective new tool for speaker identification. One can conclude that at the very least these kinds of images can complement existing voice identification techniques.

Acknowledgements

Thanks to Susan Guion and Doug O'Shaughnessy for encouragement, and Kelly Fitz for his stellar engineering work enabling this project. This paper was supported by a Fresno State award for Research, Scholarship, and Creative Activity.

4. REFERENCES

[1] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–89, 1995.

[2] R. H. Bolt, F. S. Cooper, E. E. David, P. B. Denes, J. M. Pickett, and K. N. Stevens. Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes. *J. Acoust. Soc. Am.*, 47(2 part 2):597–612, 1970.

[3] R. H. Bolt, F. S. Cooper, D. M. Green, S. L. Hamlet, D. L. Hogan, J. G. McKnight, J. M. Pickett, O. Tosi, and B. D. Underwood. *On the Theory and Practice of Voice Identification*. National Academy of Sciences, Washington, DC, 1979.

[4] J. R. Carson. Notes on the theory of modulation. *Proceedings of the Institute of Radio Engineers*, 10(1):57–64, 1922.

[5] S. A. Fulop and K. Fitz. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *J. Acoust. Soc. Am.*, 119(1):360–371, 2006.

[6] S. A. Fulop and K. Fitz. Separation of components from impulses in reassigned spectrograms. *J. Acoust. Soc. Am.*, 121(3):1510–1518, 2007.

[7] H. Hollien. *Forensic voice identification*. Academic Press, San Diego, 2002.

[8] L. G. Kersta. Voiceprint identification. *Nature*, 196:1253–1257, 1962.

[9] K. Kodera, C. de Villedary, and R. Gendrin. A new method for the numerical analysis of non-stationary signals. *Physics of the Earth and Planetary Interiors*, 12:142–150, 1976.

[10] P. Ladefoged. *A Course in Phonetics*. Thomson Wadsworth, Boston, 5th edition, 2006.

[11] L. K. Montgomery and I. S. Reed. A generalization of the Gabor-Helstrom transform. *IEEE Trans. Information Theory*, IT-13:344–345, 1967.

[12] D. J. Nelson. Cross-spectral methods for processing speech. *J. Acoust. Soc. Am.*, 110(5):2575–92, 2001.

[13] D. J. Nelson. Instantaneous higher order phase derivatives. *Digital Signal Processing*, 12:416–28, 2002.

[14] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Speech and Audio Processing*, 7(5):569–586, 1999.

[15] R. K. Potter. Visible patterns of sound. *Science*, 102(2654):463–470, Nov. 1945.

[16] T. F. Quatieri. *Discrete-Time Speech Signal Processing*. Prentice Hall, Upper Saddle River, NJ, 2002.

[17] A. W. Rihaczek. Signal energy distribution in time and frequency. *IEEE Trans. Information Theory*, IT-14(3):369–374, 1968.

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

[19] K. Sri Rama Murty and B. Yegnanarayana. Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Processing Letters*, 13(1):52–55, 2006.

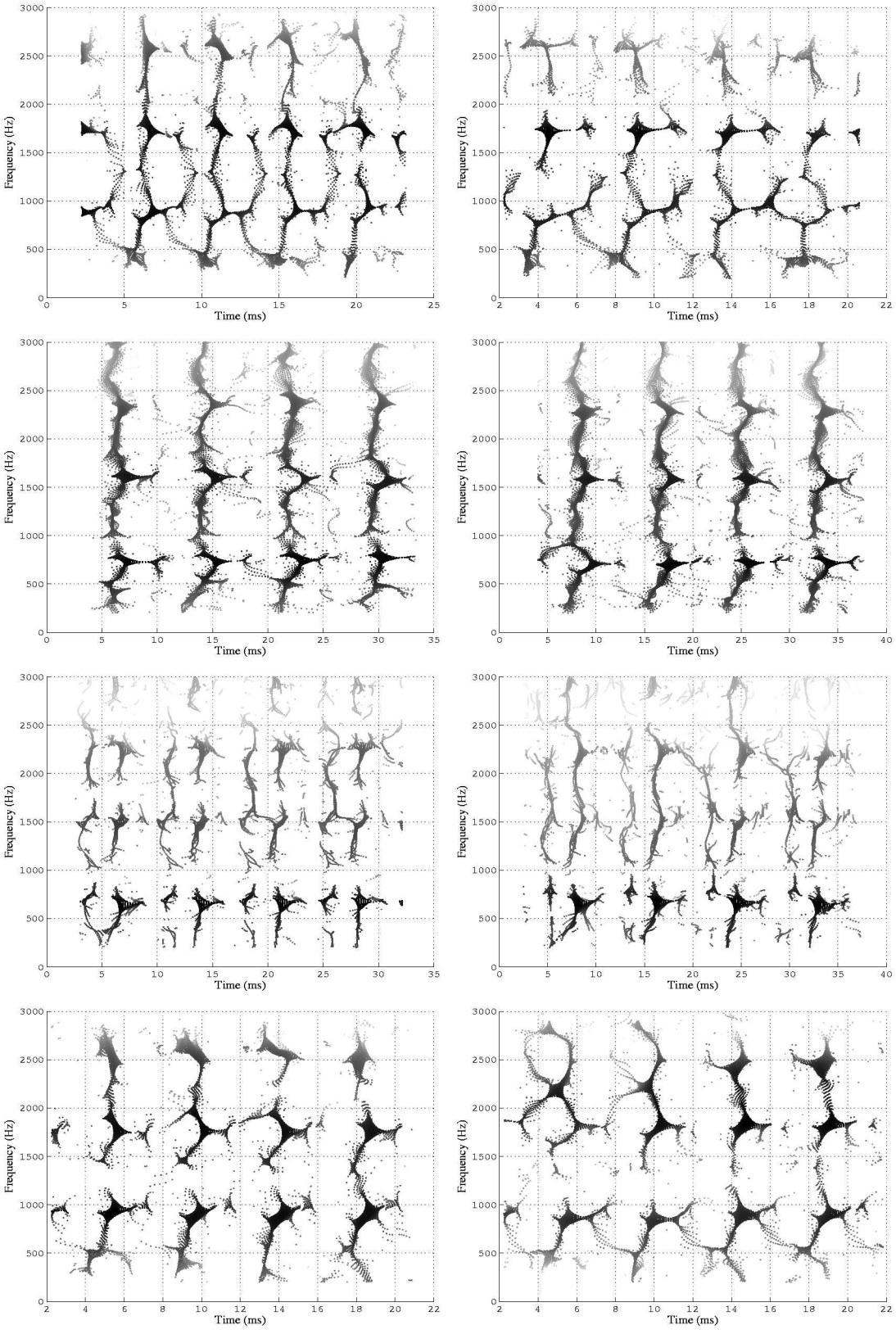

Figure 2: One pair of matching biometric images is shown in each row, taken from four different (randomly selected) speakers saying the phrase "password access" two times. Each print is a reassigned spectrogram, pruned to show both components and impulses, from the first vowel of "access." The frame length for each speaker is optimized for the fundamental frequency of the voice, and is around 75% of the period. The degree to which different speakers *do not* appear to match is perhaps more striking than the evident similarities in the matching prints.