Voice Source Waveform Analysis and Synthesis using Principal Component Analysis and Gaussian Mixture Modelling

Jon Gudnason[1]*, Mark R. P. Thomas*[1]*, Patrick A. Naylor*[1]*, Dan P. W. Ellis*[2]

[1]Comm. and Signal Proc. Group, Imperial College London, Exhibition Road, SW7 1BT
[2]LabROSA, Columbia University, New York, NY 10027

{jg, mrt102, p.naylor}@imperial.ac.uk, dpwe@ee.columbia.edu

Abstract

The paper presents a voice source waveform modeling techniques based on principal component analysis (PCA) and Gaussian mixture modeling (GMM). The voice source is obtained by inverse-filteirng speech with the estimated vocal tract filter. This decomposition is useful in speech analysis, synthesis, recognition and coding. Existing models of the voice source signal are based on function-fitting or physically motivated assumptions and although they are well defined, estimation of their parameters is not well understood and few are capable of reproducing the large variety of voice source waveforms. Here, a data-driven approach is presented for signal decomposition and classification based on the principal components of the voice source. The principal components are analyzed and the 'prototype' voice source signals corresponding to the Gaussian mixture means are examined. We show how an unknown signal can be decomposed into its components and/or prototypes and resynthesized. We show how the techniques are suited for both low bitrate or high quality analysis/synthesis schemes.

Index Terms: Voice source, inverse-filtering, closed-phase analysis, PCA, GMM

1. Introduction

This paper proposes a method for modeling the voice source waveform using Gaussian mixture modeling. The voice source waveform is used here to denote the glottal volume flow derivative [1, 2] and is considered to be the input signal in the source-filter representation of speech. Many existing models involve a piecewise fit to the voice source using standard mathematical functions. These include the Rosenberg model [3], the Liljencrants-Fant model [4], and the Klatt and Klatt model [5]. An extension to this method was provided where the coarse structure is modeled by function fitting and the fine structure modeled separately [2]. Other approaches to modeling the voice source include those motivated by physical modeling and include models such as Ishizaka and Flanagan [6] and Story and Titze [7]. The importance of accurately reproducing the voice source signal in speech synthesis is described in [8], where experimentation has shown that a parallel formant synthesizer can generate short speech segments indistinguishable from real speech provided it is driven by an inverse-filtered typical natural vowel from the same talker. A related approach is described in [9] where cepstrum coefficients are used to generate a single average voice source waveform from which any speech signal can be synthesized. The concept of voice source codebooks, derived from synthetic waveforms, has also been proposed for synthesis [10] and coding [11] with notable benefits over single-waveform models.

The motivation for modeling the voice source waveform, $u_{\mathrm{d}}(n)$, comes from the source-filter representation of speech production where an all-pole model of the vocal tract is excited by a source waveform [1],

$$s(n) = u_{\mathrm{d}}(n) + \sum_{k=0}^{p} a_k s(n-k), \tag{1}$$

where $s(n)$ is the speech signal and a_k are the frame-dependent vocal tract filter coefficients of order p (the frame dependence on a_k is implicit for the remainder of the paper). The subscript d is used here to denote that $u_{\mathrm{d}}(n)$ represents the glottal flow derivative. This description of the vocal tract is beneficial because a) linear prediction methods [12] are readily available to model the vocal tract as an all-pole filter, b) they provide a compact and accurate representation that can be efficiently quantized, and c) inverse-filtering can be achieved by filtering with an FIR filter whose zeros cancel the poles of the vocal tract. By contrast, estimation of the parameters of a voice source model to reproduce an approximation to $u_{\mathrm{d}}(n)$ is less straightforward and is an area of ongoing research [2, 13]. Additionally, some existing models fail to capture all the degrees of freedom of the voice source, particularly features like the ripples caused by a nonlinear interaction between the glottis and vocal tract [14, 15].

The proposed approach differs from previously proposed models in that a set of amplitude- and scale-normalized 'prototype' voice source waveforms are generated from the decomposition of true voice source waveforms from a large database of real talkers. The approach uses principal component analysis (PCA) to decompose the speech and and Gaussian mixture modeling (GMM) to identify voice source prototypes. A previous method [16] used mel-frequency cepstrum for the GMM so that the prototype waveforms had to be derived explicitly from the data and the posterior probabilities of each vector under each mixture component. Here, the prototypes are implicit in the model as the mixture means can by transformed back to signal space. Re-synthesis depends on the application. For low-bitrate coding, the test cycle can be represented as mixture mean cycle of the closest class or for higher quality the voice source can be reconstructed as an appropriate linear combination of either mixture mean vectors or principal components. The result is a method for accurately and succinctly analysing and resynthesizing voice source waveforms, with potential uses in speech analysis, synthesis, coding, enhancement and recognition.

This paper is organized as follows. The voice source waveform is described in Sec. 2. In Sec. 3 the process of decomposing the voice source signal into principal component analysis is explained and in Sec. 4 the voice source prototypes are derived using Gaussian mixture models. Two analysis/synthesis

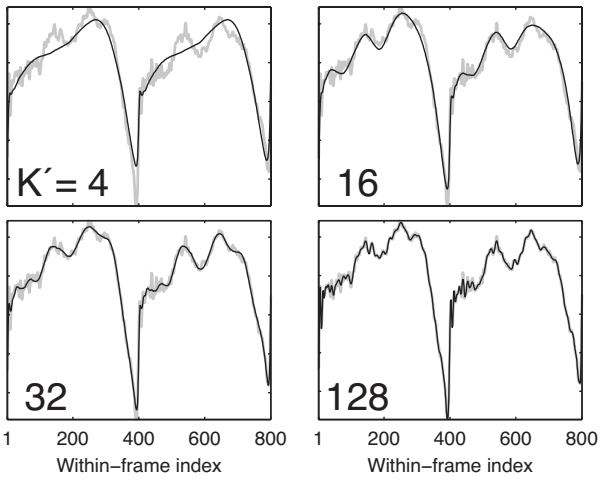

Figure 1: The gray lines show the voice source signal and the black lines the re-synthesized voice source. Using $K' = 4$ severely under-models the waveform whereas the $K' = 128$ captures the finest detail in the waveform.

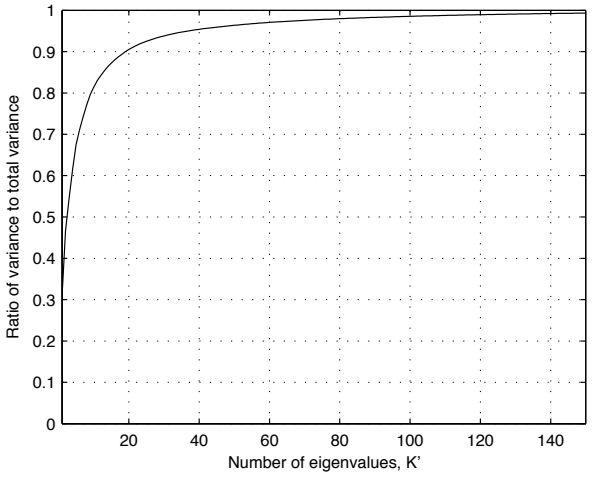

Figure 2: The variance represented in the first K' eigenvectors as a ratio to the total variance. The total number of eigenvalues is 800.

approaches are demonstrated in Sec. 5 and the paper is concluded in Sec. 6.

2. The Voice Source Signal

The voice source signal $u_{\rm d}(n)$ is obtained from (1) by inverse filtering the speech signal $s(n)$ using the vocal tract parameters a_k. Here the filter parameters model the vocal tract transfer function for every larynx cycle and are obtained by pre-emphasizing voiced segments of the speech so to correct for the spectral tilt caused by the glottal pulse [12].

The result of the inverse filtering, $u_{\rm d}(n)$, is first divided into scale- and amplitude-normalized overlapping two-cycle glottal-synchronous frames so that classification is based on waveform shape only,

$$\mathbf{u}_i = \updownarrow_{\alpha}^{\beta} \kappa u_{\rm d}(n), n \in \{n_i^c, \ldots, n_{i+2}^c - 1\}, \qquad (2)$$

where $\updownarrow_{\alpha}^{\beta}$ denotes a resampling operation of factor $\frac{\beta}{\alpha}$, $\beta = 2t_{max}f_s$, $\alpha = n_{i+2}^c - n_i^c$ and κ is a gain factor that normalizes A-weighted energy [17].

The APLAWD database [18] contains ten repetitions of five short sentences by five male and five female talkers. Sentence 2: *"Why are you early you owl?"* contains only voiced speech and provides all the data (approximately 22,000 glottal cycles) for model training. The speech is recorded at 20 kHz and contains contemporaneous EGG recordings. The SIGMA algorithm [19] was applied to these recordings to obtain the glottal closure instants (GCIs), n_i^c needed for the analysis.

The maximum period of voiced speech is t_{max}, set to 20 ms and f_s is the sampling frequency (20 kHz) resulting in the length of $\mathbf{u}_i$ of $\beta = 2t_{max}f_s = 800$ samples. Using two-cycle frames ensures that high-energy glottal closures occur in the centre of the window which aids the quality of resynthesis [20] and ensures that the excitation from glottal closure is not attenuated by windowing in the subsequent feature extraction. An example of a normalized, resampled voice source waveform is shown in Fig. 1 and its principal component approximation described next.

3. Principal Components Analysis

Principal component analysis (PCA) of the voice source waveform is obtained from the linear combination,

$$\mathbf{x}_i = (\mathbf{u}_i - \bar{\mathbf{u}}) = \sum_{k=1}^{K} z_{i,k}\mathbf{v}_k = \mathbf{V}\mathbf{z}_i \qquad (3)$$

where $\bar{\mathbf{u}}$ is the empirical mean of $\mathbf{u}$ and $\mathbf{v}_k$ are the eigenvectors of the covariance matrix $\mathbf{\Sigma}_x = E\{\mathbf{x}\mathbf{x}^T\}$ ($\mathbf{x}$ is zero mean by design). The coefficients $z_{i,k}$ are called the PCA spectra and represent the projection of $\mathbf{x}_i$ onto components $\mathbf{v}_k$. It is also assumed that the eigenvectors are ordered on the eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_K$. There are two reasons for applying PCA to the voice source waveform shapes. First, it provides a method for coding the voice source by representing it as the linear combination of the first (few) components. The second reason is to reduce the number of coefficients for the Gaussian mixture modeling which we describe in next section.

The voice source vectors are approximated such that the L^2 norm of the error $E\{(\hat{\mathbf{u}} - \mathbf{u})^T(\hat{\mathbf{u}} - \mathbf{u})\}$ is minimized by

$$\hat{\mathbf{u}}_i = \sum_{k=1}^{K'} z_{i,k}\mathbf{v}_k + \bar{\mathbf{u}} = \mathbf{V}'\mathbf{z}_i' + \bar{\mathbf{u}} \qquad (4)$$

where $K' < K$, $\mathbf{V}'$ is a $K \times K'$ matrix of the first K' eigenvectors and $\mathbf{z}_i'$ contains the first K' elements of $\mathbf{z}_i$. The number of eigenvectors K' is determined from the variance represented in the first K' eigenvectors as a ratio to the total variance $\sum_{i=1}^{K'}\lambda_i / \sum_{i=1}^{K}\lambda_i$. This is plotted in Fig. 2 where it can be seen that more than 90% of the variance is represented in the first 20 eigenvectors and suggests that the intrinsic dimensionality of the voice source is quite small ($\ll 800$).

The voice source waveform mean vector $\bar{\mathbf{u}}$ and the first four principle components $\mathbf{v_k}$ are shown in Fig. 3. All the components model the excitation with abrupt change at the glottal closure instants. The mean vector captures the average shape of the waveform whereas the first two components model the flatness in the closed phase and the steepness of the opening. Higher components contain higher frequencies for modeling finer details of the waveform.

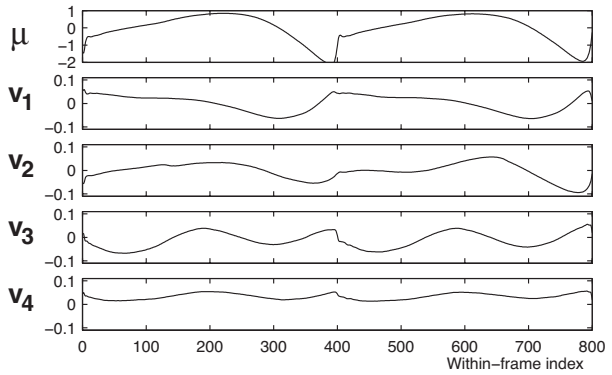

Figure 3: The mean voice source waveform and the first four principal components. They all model the excitation with the abrupt pulse. The principal component display increasingly higher frequency components.

Figure 1 shows how PCA approximates the voice source waveform for $K' = 4, 16, 32,$ and 128. The gray lines show the voice source signal and the black line the approximation. Choosing $K' = 4$ results in a bad approximation even to coarse features such as the duration of the return phase is not well captured. $K' = 16$ captures the coarse features but fails to model the kink apparent in the crest of the pulse. This is captured by both $K' = 32, 128$ but the approximation using $K' = 128$ also starts modeling the fine details in the waveform.

4. Gaussian Mixture Modeling

PCA can also be used to reduce the number of components of $\mathbf{u}_i$ to make Gaussian mixture modeling easier. The PCA spectra $\mathbf{z_i}$ can be modeled using GMM so that the total likelihood under the model is,

$$\begin{aligned} f(\mathbf{z}'_i) &= \sum_{m=1}^{M} p(\omega_m) f(\mathbf{z}'_i|\omega_m) \qquad (5) \\ &= \sum_{m=1}^{M} p(\omega_m) \frac{\exp(-\frac{1}{2}(\mathbf{z}'_i - \mu_m^{(z)})^T \mathbf{\Sigma}_m^{(z)-1}(\mathbf{z}'_i - \mu_m^{(z)}))}{\sqrt{(2\pi)^{K'}|\mathbf{\Sigma}_m^{(z)}|}} \end{aligned}$$

where $p(\omega_m)$, $\mu_m^{(z)}$ and $\mathbf{\Sigma}_m^{(z)}$ are the weight, mean vector and covariance matrix (diagonal) of the m-th mixture component ω_m. The number of principal components were chosen to be $K' = 64$ capturing more than 95% of the variance. The parameters are estimated using the EM-algorithm [21], terminating the iteration after 50 times or when the increase in log likelihood falls below 0.0001. The Fisher-ratio [22] did not increase significantly as the number of components were increased beyond $M = 16$ so this was chosen for the model.

The prototype voice source waveforms can be formed by transforming the mixture means,

$$\bar{\mathbf{u}}_m = \mathbf{V}' \mu_m^{(z)} + \bar{\mathbf{u}} \qquad (6)$$

Three prototype voice source waveforms are shown in Fig. 4. These prototypes exhibit interesting features captured by the mixture modeling. The basic shape parameter [4] varies and is very pronounced in Fig. 4(a) and Fig. 4(b) shows a very flat closed phase. Fig. 4(c) displays a clear fine-detail ripple. The remaining prototypes exhibit variation in all these parameters

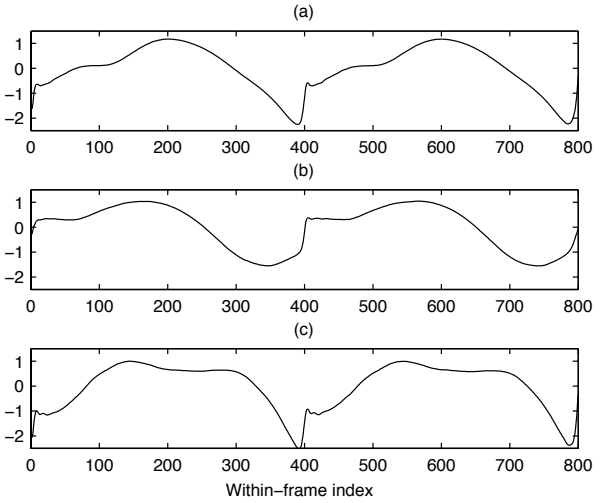

Figure 4: Selected mixture means. Components were numbered from 1 to 16 with weights in descending order. (a) Mixture 3, weight 0.10, (b) Mixture 9, weight 0.065, and (c) Mixture 12, weight 0.033.

and, additionally, provide an insight into interdependencies between them.

5. Analysis/Synthesis

Figure 1 shows how the principal components can be used in an analysis synthesis scheme. Here the voice source has been re-synthesized from K' components. The re-synthesized speech waveform is shown in Figure 5(a)-(d). The coefficients

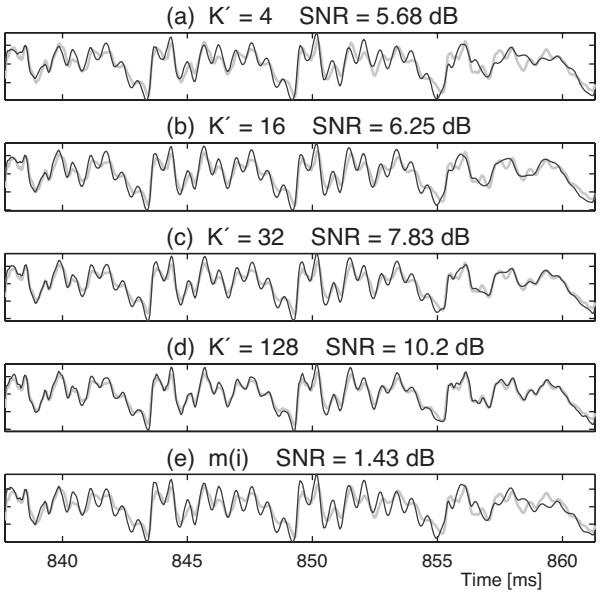

Figure 5: The gray lines show the speech signal and the black lines the result of the re-synthesis. Figures (a)-(d) show synthesized speech from K' PCA spectra and Figure (e) shows synthesized speech from prototype voice source waveforms.

needed to encode each pitch period are the K' PCA spectra $\mathbf{z}'_i$, the pitch period, and the energy and a further p vocal tract coefficients for a fixed rate of 10 ms. The signal-to-noise ratio $10\log_{10}\frac{\sum_n s^2(n)}{\sum_n [s(n)-\hat{s}(n)]^2}$ for the segment shown is 5.68, 6.25, 7.83, and 10.2 dB respectively.

Alternatively the voice source can be re-synthesized from a single GMM-derived prototype $\bar{\mathbf{u}}_{m(i)}$ where

$$m(i) = \arg\max_m p(\omega_m|\mathbf{z}'_i) = \arg\max_m \frac{p(\omega_m) f(\mathbf{z}'_i|\omega_m)}{f(\mathbf{z}'_i)}. \quad (7)$$

Figure 6 shows a speech signal, the hard classification $m(i)$ and the posterior probability $p(\omega_m|\mathbf{z}'_i)$. The vocal tract parameters, the pitch cycle energy and the period still need to be encoded but instead of K' PCA spectra $\mathbf{z}'_i$ only an integer $m(i) \in \{1, 2, \ldots 16\}$ represents the voice source waveform. The resulting signal-to-noise ratio is for the segment shown in Fig. 5(e) is 1.43 dB.

6. Conclusions

The paper presents a novel approach to modeling the voice source signal and shows how the proposed techniques can be used for an analysis/synthesis scheme applicable to coding, synthesis and voice morphing. The technique determines the PCA spectra of the voice source waveform vector and uses that to find the closest prototype voice source waveform. These prototypes display features such as the basic shape parameter and the closed-phase duration and which have interested researchers in the past.

7. Acknowledgements

This work was supported by the Royal Academy of Engineering through the The Global Research Award scheme.

8. References

[1] D. Y. Wong, J. D. Markel, and J. A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 350–355, Aug. 1979.

[2] 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 Audio Process.*, vol. 7, no. 5, pp. 569–576, Sept. 1999.

[3] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *Journal Acoust. Soc. of America*, vol. 49, pp. 583–590, Feb. 1971.

[4] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[5] D. H. Klatt and L. C. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," *Journal Acoust. Soc. of America*, vol. 87, no. 2, pp. 820–857, Feb. 1990.

[6] K. Ishizaka and J. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 51, pp. 1233–1268, 1972.

[7] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *Journal Acoust. Soc. of America*, vol. 97, pp. 1249–1260, 1994.

[8] J. N. Holmes, "The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoust.*, vol. 21, no. 3, pp. 298–305, 1973.

[9] P. Chytil and M. Pavel, "Variability of glottal pulse estimation using cepstral method," in *Proc. 7th Nordic Signal Processing Symposium (NORSIG)*, 2006, pp. 314–317.

[10] D. McElroy, B. P. Murray, and A. D. Fagan, "Wideband speech coding using multiple codebooks and glottal pulses," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 253–256.

[11] A. Bergstrom and P. Hedelin, "Code-book driven glottal pulse analysis," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1989, pp. 53–56.

[12] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[13] P. Alku and T. Backstrom, "Normalized amplitude quotient for parametrization of the glottal flow," *Journal Acoust. Soc. of America*, vol. 112, no. 2, pp. 701–710, Aug. 2002.

[14] T. V. Ananthapadmanabha and G. Fant, "Calculations of true glottal volume-velocity and its components," *Speech Communication*, vol. 1, pp. 167–184, 1982.

[15] D. G. Childers and C. F. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Trans. Biomed. Eng.*, vol. 41, pp. 663–671, July 1994.

[16] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Data-driven voice source waveform modelling," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, Apr. 2009.

[17] IEC, "IEC 61672:2003: Electroacoustics – sound level meters," IEC, Tech. Rep., 2003.

[18] G. Lindsey, A. Breen, and S. Nevard, "SPAR's archivable actual-word databases," University College London," Technical Report, June 1987.

[19] M. R. P. Thomas and P. A. Naylor, "The SIGMA algorithm for estimation of reference-quality glottal closure instants from electroglottograph signals," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[20] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Application of the DYPSA algorithm to segmented time-scale modification of speech," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Lausanne, Switzerland, Aug. 2008.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2001.

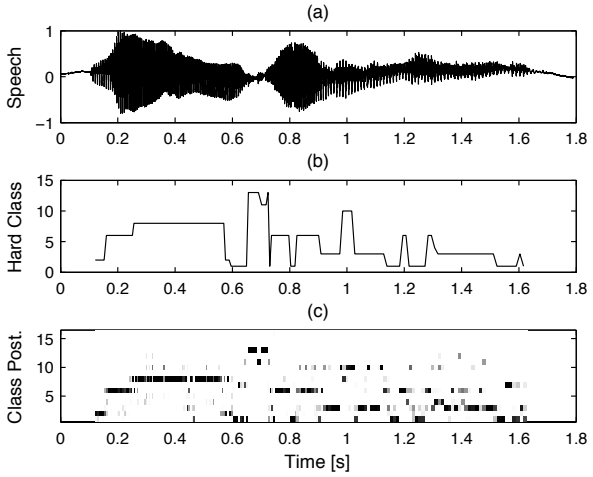

Figure 6: Speech signal analysis. a) Original speech signal, b) $\max p(\omega_m|\mathbf{z}'_i)$, mode-filtered with length 5, c) probability matrix $p(\omega_m|\mathbf{z}'_i)$, where black:= $(p(\omega_m|\mathbf{z}'_i) = 1)$.