DATA-DRIVEN VOICE SOURCE WAVEFORM MODELLING

Mark R. P. Thomas, Jon Gudnason, and Patrick A. Naylor

Imperial College London
Exhibition Road, London SW7 2AZ, UK
E-mail: {mrt102, jg, p.naylor}@imperial.ac.uk

ABSTRACT

This paper presents a data-driven approach to the modelling of voice source waveforms. The voice source is a signal that is estimated by inverse-filtering speech signals with an estimate of the vocal tract filter. It is used in speech analysis, synthesis, recognition and coding to decompose a speech signal into its source and vocal tract filter components. Existing approaches parameterize the voice source signal with physically- or mathematically-motivated models. Though the models 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 we present a data-driven approach to classify types of voice source waveforms based upon their mel-frequency cepstrum coefficients with Gaussian mixture modelling. A set of 'prototype' waveform classes is derived from a weighted average of voice source cycles from real data. An unknown speech signal is then decomposed into its prototype components and resynthesized. Results indicate that with sixteen voice source classes, low resynthesis errors can be achieved.

Index Terms— Voice source, inverse-filtering, closed-phase analysis, LPC

1. INTRODUCTION

A number of voice source models have been proposed that fall into two main categories: those motivated by a need for compact mathematical descriptions including Rosenberg [1], Fant [2], Klatt and Klatt [3] and Plumpe and Quatieri [4], and those motivated by physical modelling such as Ishizaka and Flanagan [5] and Story and Titze [6]. The proposed approach uses real voice source waveforms to create a codebook of 'prototype' signals, using Gaussian mixture modelling (GMM) [7] to classify cycles of the voice source based upon their mel-frequency cepstral coefficients (MFCCs) [8].

The motivation for modelling 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,

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

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). This description of the vocal tract is beneficial because a) linear prediction methods [9] 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 [4, 10]. 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 [11, 12]. By using a GMM to classify cycles of $u_{\mathrm{d}}(n)$ based upon their MFCCs, the proposed approach captures all significant features that are common to each class. While MFCC features may be suboptimal, they provide a good starting point owing to their extensive use in speech recognition.

The importance of accurately reproducing $u_{\mathrm{d}}(n)$ in speech synthesis is described in [13], 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 [14] 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 [15] and coding [16] with notable benefits over single-waveform models. Our method differs in that a set of amplitude- and scale-normalized prototype voice source waveforms are generated from a weighted average of true voice source waveforms from a large database of real talkers. Resynthesis involves calculating the probability that a test cycle is a member of each of the prototype classes, performing a weighted average of the prototype voice source waveforms, then scaling to reproduce the correct duration and amplitude. 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. In Section 2, the concept of a voice source 'prototype' is introduced and then derived from real data. Section 3 demonstrates the decomposition of an unknown voice source into its prototype components its subsequent resynthesis. Conclusions are drawn in Section 4.

2. MODEL TRAINING

2.1. Overview

The aim of voice source analysis is to characterize and model the voice source waveform, $u_{\mathrm{d}}(n)$, obtained by inverse filtering the speech signal, $s(n)$, with an estimate of the vocal tract filter. The APLAWD database [17] contains ten repetitions of five short sentences by five male and five female talkers and provides all the data (approximately 110,000 glottal cycles) for model training. LPC autoregressive (AR) coefficients [9], a_k, model the vocal tract filter for every larynx cycle of every utterance in the training corpus. The glottal closure instants (GCIs), n_i^c, are obtained using the SIGMA

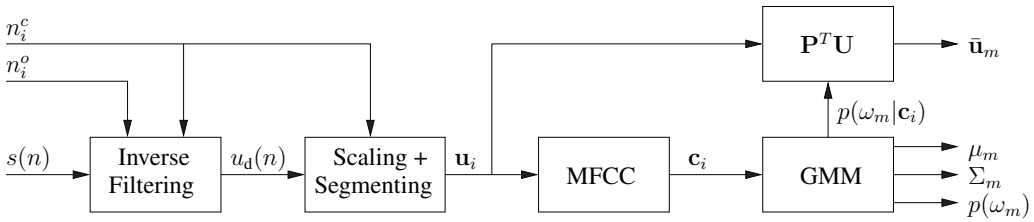

Fig. 1. System diagram for model training.

algorithm [18] applied to EGG signals recorded contemporaneously with the speech signal. Model training determines a set of classes, $\omega_m, m \in \{1, \ldots, M\}$ with associated 'prototype' waveforms that can describe voiced regions of $u_{\rm d}(n)$. Male and female data was mixed so as to provide a gender-independent representation of all voice source waveforms in the training set.

2.2. Segmentation

The result of the closed-phase 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,

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

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 [19]. The maximum period of voiced speech is t_{max}, set to 20 ms and f_s is the sampling frequency (Hz). 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.

$\mathbf{u}_i^T$ form the rows of an $(N \times L)$ data matrix,

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N]^T. \tag{3}$$

Converting the voice source waveform to the mel-cepstrum domain makes it suitable for clustering using diagonal covariance Gaussian mixture models. Classifying the rows of $\mathbf{U}$ begins by deriving C Mel-Frequency Cepstrum Coefficients (MFCCs) [8] for each glottal period i, represented by an $(N \times C)$ feature matrix $\mathbf{C}$ with rows $\mathbf{c}_i$. The cepstrum coefficients are computed using 29 mel filter banks, discarding the '0-th' and last 16 coefficents leading to the dimensionality C of $\mathbf{c}_i$ equal to 12.

The likelihood of feature vector $\mathbf{c}_i$ is computed as a weighted sum of Gaussians,

$$\begin{aligned} f(\mathbf{c}_i) &= \sum_{m=1}^{M} p(\omega_m) f(\mathbf{c}_i|\omega_m) \qquad (4) \\ &= \sum_{m=1}^{M} p(\omega_m) \frac{\exp(-\frac{1}{2}(\mathbf{c}_i - \mu_m)^T \mathbf{\Sigma}_m^{-1}(\mathbf{c}_i - \mu_m))}{\sqrt{(2\pi)^C |\mathbf{\Sigma}_m|}} \end{aligned}$$

where $p(\omega_m)$, μ_m and $\mathbf{\Sigma}_m$ are the weight, mean vector and covariance matrix (diagonal) of the m-th mixture component ω_m. The parameters are estimated using the EM-algorithm [7], terminating the iteration after 100 times or when the increase in log likelihood falls below 0.0001.

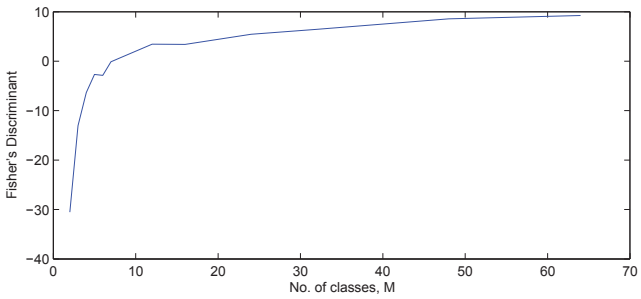

Fig. 2. Fisher's Discriminant as a function of M classes.

The probability that cluster m generates $\mathbf{c}_i$ is

$$p(\omega_m|\mathbf{c}_i) = \frac{p(\omega_m) f(\mathbf{c}_i|\omega_m)}{f(\mathbf{c}_i)}. \tag{5}$$

Let the $(N \times M)$ probability matrix $\mathbf{P}$ contain $p(\omega_m|\mathbf{c}_i) \forall m, i$,

$$\mathbf{P} = \begin{bmatrix} p(\omega_1|\mathbf{c}_1) & \cdots & p(\omega_M|\mathbf{c}_1) \\ \vdots & \ddots & \vdots \\ p(\omega_1|\mathbf{c}_N) & \cdots & p(\omega_M|\mathbf{c}_N) \end{bmatrix}. \tag{6}$$

2.3. Model Complexity

Fisher's Discriminant, F_m [21], measures the ratio of the intra- to inter-class variances; the higher the figure the more separated the classes. Randomly selecting half the speech samples as training data and the other half as test data, F_m was calculated, varying M from 2 to 64. Fig. 2 shows Fisher's Discriminant as a function of M. Asymptotic behaviour is seen beyond around $M = 16$ which is used from this point onwards. For the purposes of coding, for example, small M is desired to reduce data bandwidth.

2.4. Prototype Generation

The mean prototype voice source waveform, $\bar{\mathbf{u}}_m$, for class ω_m is calculated by soft classification of feature $\mathbf{c}_i$. Prototypes are derived from a weighted average of time-domain waveforms, $\mathbf{u}_i$, using the probabilities $p(\omega_m|\mathbf{c}_i)$ as weights.

$$\bar{\mathbf{u}}_m = \kappa_m \sum_i p(\omega_m|\mathbf{c}_i)\mathbf{u}_i, \tag{7}$$

where κ_m is a constant that normalizes A-weighted energy. The $\bar{\mathbf{u}}_m$ form the rows of the $(M \times L)$ prototype matrix,

$$\bar{\mathbf{U}} = [\bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2, \ldots, \bar{\mathbf{u}}_M]^T = \mathbf{P}^T\mathbf{U}. \tag{8}$$

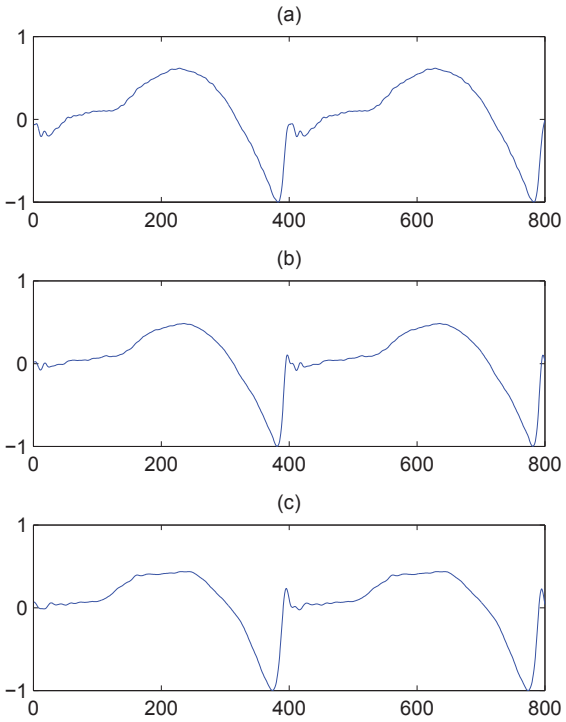

Fig. 3. Three of the set of sixteen prototypes, $\bar{\mathbf{u}}_m$, with varying ripple from nonlinear coupling between the glottis and vocal tract.

Many existing models of voice source waveforms include some scale-independent parameters, including the 'basic shape parameter' [2], defined as $\min(u_{\rm d}(n))/\max(u_{\rm d}(n))$, the open quotient (OQ) [3], and the duration of the return phase [2]. In [4], a polynomial 'fine' detail model describes the error between measured $u_{\rm d}(n)$ and the Fant model, attributed mainly to nonlinear interaction between the glottis and vocal tract.

Fig. 3 shows 3 of the 16 classes. The prototypes exhibit very low noise and very little 'overshoot' from LPC framing errors [22]. The basic shape parameter is seen varying from high in Fig. 3(a) to low in Fig. 3(c) and, at the same time, fine-detail ripple varies and is most pronounced in Fig. 3(c). The open phase and duration of the return phase in Fig. 3(c) are noticeably longer. The remaining prototypes exhibit variation in all these parameters and, additionally, provide an insight into interdependencies between them.

3. ANALYSIS/SYNTHESIS

A test utterance can now be decomposed into AR coefficients, a_k, and voice source prototypes, $\bar{\mathbf{u}}_m$. In a similar manner to the signal framing for prototype training in (2), the test utterance is split into N overlapping frames, $\mathbf{u}_i$, where frame i contains two cycles of voiced speeeh. The voice source decomposition for frame i is

$$\boldsymbol{\gamma}_i = [\gamma_{1,i}, \gamma_{2,i}, \ldots, \gamma_{M,i}] = [p(\omega_1|\mathbf{u}_i), \ldots, p(\omega_M|\mathbf{u}_i)]. \quad (9)$$

Figure 4 gives an example of voice source decomposition of the all-voiced utterance, "Why are you early you owl?," spoken by a male speaker not included in the training corpus, showing a) the speech signal, b) $\max_m p(\omega_m|\mathbf{c}_i)$, passed through a moving mode filter of length 5, and c) $p(\omega_m|\mathbf{c}_i)$ as a reverse gray-scale for each mixture

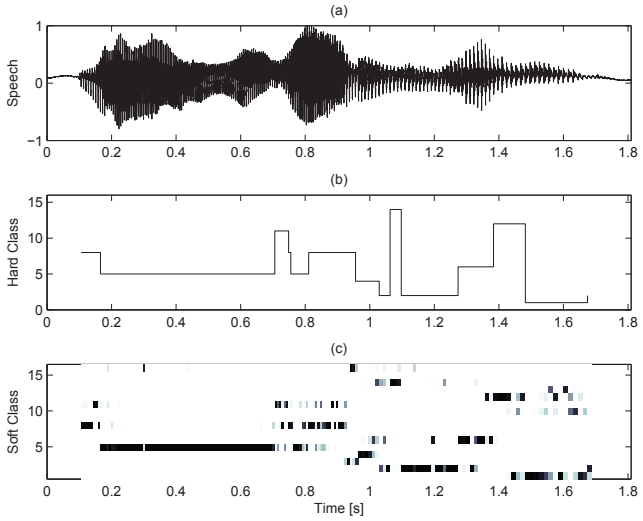

Fig. 4. Speech signal analysis. a) Original speech signal, b) $\max\, p(\omega_m|\mathbf{c}_i)$, mode-filtered with length 5, c) probability matrix $p(\omega_m|\mathbf{c}_i)$, where black:= $(p(\omega_m|\mathbf{c}_i) = 1)$.

component, black indicating a probability of one and white a probability of zero. The figure shows that most voiced cycles can be decomposed into a compact prototype set; for many cycles the probability of membership of a single class is close to 1 and inter-cycle class membership is piecewise-constant, suggesting significant potential run-length encoding coding gains. Informal tests have shown that this approach performs similarly for female voices.

We define a set $\mathcal{M}_i \subseteq \mathcal{M}_{all}$, $\mathcal{M}_{all} = \{1, \ldots, M\}$. $\mathcal{M}_i$ contains the class indices that produce the highest likelihood, reducing computational complexity. In Fig. 4(c), at time 0.8 s, a sensible choice is $|\mathcal{M}_i| = 3$ so that $\mathcal{M}_i = \{5, 8, 11\}$. A cycle of the voice source signal can then be resynthesized from the prototypes, $\bar{\mathbf{u}}_m$, with the decomposition terms,

$$\hat{\mathbf{u}}_i = \sum_{m \in \mathcal{M}_i} \gamma_{m,i} \left(\updownarrow_\alpha^\beta \kappa \bar{\mathbf{u}}_m \right), \quad (10)$$

where $\updownarrow_\alpha^\beta$ resamples $\bar{\mathbf{u}}_m$ to length $(n_{i+2}^c - n_i^c)$ for cycle i and κ is a gain factor to reproduce the same A-weighted energy as the source cycle. An approximation to the full $u_{\rm d}(n)$ is synthesized by windowing $\hat{\mathbf{u}}_i$ with a Hamming window, $\mathbf{w}_i$, shifting to centre the waveform on n_i^c and summing,

$$\hat{u}(n) = \sum_i (\hat{\mathbf{u}}_i \odot \mathbf{w}_i) * \delta_{n_i^c}, \quad (11)$$

where $\odot$ is a Hadamard (element-by-element) product, $*$ a linear convolution operator and $\delta_{n_i^c}$ is a unit delta placed at GCI n_i^c. The speech signal can be resynthesized in a similar manner to (1),

$$s(n) \simeq \hat{s}(n) = \hat{u}(n) + \sum_{k=0}^{p} a_k s(n-k). \quad (12)$$

Fig. 5 shows, for $\mathcal{M}_i = \mathcal{M}_{all}\ \forall i$, a) $s(n)$ and resynthesized $\hat{s}(n)$ and b) the corresponding segment of $u_d(n)$ and its resynthesized approximation, $\hat{u}_d(n)$. The signal-to-noise ratio for the segment shown, $SNR = 10 \log_{10} \frac{\sum_n s^2(n)}{\sum_n [s(n) - \hat{s}(n)]^2} = 11.75$ dB, shows a high correlation between the two sets of signals.

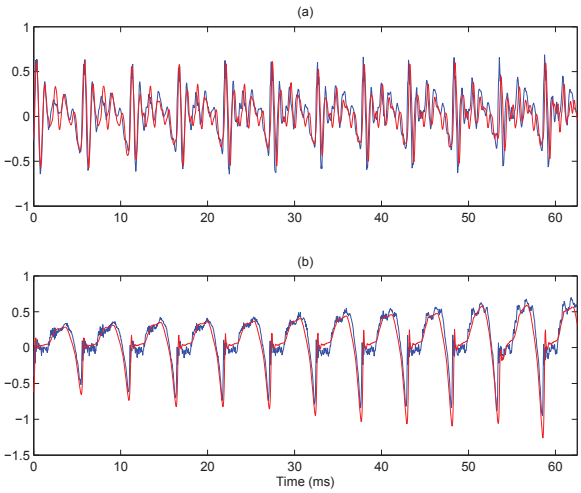

Fig. 5. Original and resynthesized signals. a) Speech signal, $s(n)$, (blue) and resynthesized, $\hat{s}(n)$, (red), b) Voice source waveform, $u_{\mathrm{d}}(n)$, (blue) and resynthsized, $\hat{u}_{\mathrm{d}}(n)$, (red).

The proposed method has been trained and tested on modal voiced speech only. An obvious extension is a generalization that processes noisy frames such as breathy, unvoiced and mixed voiced/unvoiced speech waveforms. Research into alternative approaches for creating prototypes is necessary due to the de-noising effect of the existing averaging procedure.

4. CONCLUSIONS

A novel, data-driven technique for the modelling of voice source waveforms has been presented. It has been shown that, by classifying cycles of inverse-filtered speech according to 12 mel-frequency cepstrum coefficients, 16 prototype waveforms with associated means, diagonal covariance matrices and weights, can be derived. This allows the decomposition of an unknown inverse-filtered modal voiced speech signal into a weighted prototype description which can be resynthesised with an SNR of up to $\sim$12 dB. The proposed technique has potential uses in speech analysis, synthesis, coding, enhancement and recognition by providing a compact and reliable description of glottal waveforms.

5. REFERENCES

[1] A. E. Rosenberg, "Effect of Glottal Pulse shape on the Quality of Natural Vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–590, Feb. 1971.

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

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

[4] 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 Processing*, vol. 7, no. 5, pp. 569–576, Sept. 1999.

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

[6] B. H. Story and I. R. Titze, "Voice Simulation with a Body-Cover Model of the Vocal Folds," *J. Acoust. Soc. Amer.*, vol. 97, pp. 1249–1260, 1994.

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

[8] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continously Spoken Sentences," *IEEE Trans Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.

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

[10] P. Alku and T. Backstrom, "Normalized Amplitude Quotient for Parametrization of the Glottal Flow," *J. Acoust. Soc. Amer.*, vol. 112, no. 2, pp. 701–710, August 2002.

[11] T. V. Ananthapadmanabha and G. Fant, "Calculations of True Glottal Volume-Velocity and its Components," *Speech Communication*, vol. 1, pp. 167–184, 1982.

[12] D. Childers and C. Wong, "Measuring and Modeling Vocal Source-Tract Interaction," *IEEE Transactions on Biomedical Engineering*, vol. 41, pp. 663–671, July 1994.

[13] J. 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.

[14] P. Chytil and M. Pavel, "Variability of Glottal Pulse Estimation Using Cepstral Method," in *Proc. 7th Nordic Signal Processing Symposium NORSIG 2006*, 2006, pp. 314–317.

[15] D. McElroy, B. Murray, and A. Fagan, "Wideband speech coding using multiple codebooks and glottal pulses," in *Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP-95*, vol. 1, 1995, pp. 253–256 vol.1.

[16] A. Bergstrom and P. Hedelin, "Code-book driven glottal pulse analysis," in *Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP-89*, 1989, pp. 53–56 vol.1.

[17] G. Lindsey, A. Breen, and S. Nevard, "SPAR's Archivable Actual-Word Databases," University College London," Technical Report, June 1987.

[18] 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*, Lausanne, Switzerland, August 2008.

[19] "IEC 61672:2003: Electroacoustics – Sound Level Meters," IEC, Tech. Rep., 2003.

[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*, Lausanne, Switzerland, August 2008.

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

[22] Y. Ting, D. Childers, and J. Principe, "Tracking spectral resonances," in *Proc. Fourth Annual ASSP Workshop on Spectrum Estimation and Modeling*, 1988, pp. 49–54.