Articulation rate filtering of CQCC features for automatic speaker verification

Massimiliano Todisco, Héctor Delgado and Nicholas Evans

Department of Digital Security, EURECOM, Sophia Antipolis, France
{todisco,delgado,evans}@eurecom.fr

Abstract

This paper introduces a new articulation rate filter and reports its combination with recently proposed constant Q cepstral coefficients (CQCCs) in their first application to automatic speaker verification (ASV). CQCC features are extracted with the constant Q transform (CQT), a perceptually-inspired alternative to Fourier-based approaches to time-frequency analysis. The CQT offers greater frequency resolution at lower frequencies and greater time resolution at higher frequencies. When coupled with cepstral analysis and the new articulation rate filter, the resulting CQCC features are readily modelled using conventional techniques. A comparative assessment of CQCCs and mel frequency cepstral coefficients (MFCC) for a short-duration speaker verification scenario shows that CQCCs generally outperform MFCCs and that the two feature representations are highly complementary; fusion experiments with the RSR2015 and RedDots databases show relative reductions in equal error rates of as much as 60% compared to an MFCC baseline.

Index Terms: Automatic speaker verification, constant Q cepstral coefficients, articulatory filter.

1. Introduction

Through our work to develop spoofing countermeasures to help protect automatic speaker verification (ASV) from circumvention, we recently investigated the application of constant Q transform analysis as a novel approach to spoofing detection [1]. The motivation for this work revolved around the potential benefit of using features for spoofing detection which are fundamentally different to those used for ASV.

The new features, termed constant Q cepstral coefficients (CQCCs), are based on the constant Q transform, which employs a variable time-frequency resolution. Compared to the discrete Fourier transform (DFT), the CQT frequency resolution is greater at lower frequencies whereas the time resolution is greater at higher frequencies. As a result, CQCCs tend to capture greater spectral detail at lower frequencies and greater temporal details at higher frequencies, detail which is generally lost through more traditional approaches to time-frequency analysis.

The results of the spoofing detection study were extremely encouraging; they were, at the time of writing, the best reported spoofing detection results produced using the standard ASVspoof 2015 database [2]. The performance improvement delivered through CQCCs for spoofing detection motivated our recent work to investigate their use for ASV itself.

Herein lies the contributions of this paper. It reports the first assessment of CQCCs for ASV and compares their performance to traditional MFCCs. Feature complementarity is also assessed with score fusion experiments. Motivated by the benefit of RASTA filtering, the second contribution of this paper is a new articulation rate filter tailored to CQCC analysis.

The rest of the paper is organised as follows. Section 2 describes the constant Q transform and CQCC extraction. The new articulatory filter is described in Section 3. Sections 4 and 5 describe the experimental setup and results before conclusions are presented in Section 6.

2. Constant Q Cepstral Coefficients

Constant Q cepstral coefficients (CQCCs) were introduced recently in the context of spoofing detection for ASV [1]. CQCC extraction draws upon the combination of the constant Q transform and cepstral analysis. CQCCs are an appealing alternative to traditional MFCCs; they offer a time-frequency resolution more closely related to that of human perception.

2.1. The constant Q transform

The constant Q transform (CQT) is a perceptually motivated approach to time-frequency analysis introduced by Youngberg and Boll [3] in 1978. The original algorithm has been refined over the last few decades, e.g. [4]. In contrast to Fourier-based approaches, the centre/bin frequencies of the CQT scale are geometrically distributed, thereby following the equal-tempered scale [5] of Western music. This is one reason why CQT has attracted significant attention in the field of music signal processing, e.g. [6, 7, 8, 9]. Compared to the short-time Fourier transform (STFT), the CQT gives a greater frequency resolution for lower frequencies and a greater temporal resolution for higher frequencies. The CQT of a discrete signal $x(n)$ is defined by:

$$X^{CQ}(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j-n+N_k/2) \quad (1)$$

where $k = 1, 2, ..., K$ is the frequency bin index, $a_k(n)$ are the basis functions, $*$ is the complex conjugate and N_k is a variable window length – full details are presented in [1]. The center frequencies f_k are defined according to $f_k = 2^{(k-1)/(B)} f_1$, where f_k is the center frequency of bin k, f_1 is the center frequency of the lowest frequency bin and B is the number of bins per octave. In practice, B determines the time-frequency resolution trade-off.

The Q-factor is a measure of the filter selectivity and reflects the ratio between the center frequency and the bandwidth:

$$Q = \frac{f_k}{f_{k+1} - f_k} = (2^{1/B} - 1)^{-1} \quad (2)$$

The CQT is similar to a wavelet transform with a relatively high Q factor (≃100 bins per octave). Wavelet techniques are, however, not well suited to this computation [10]. For example,

The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

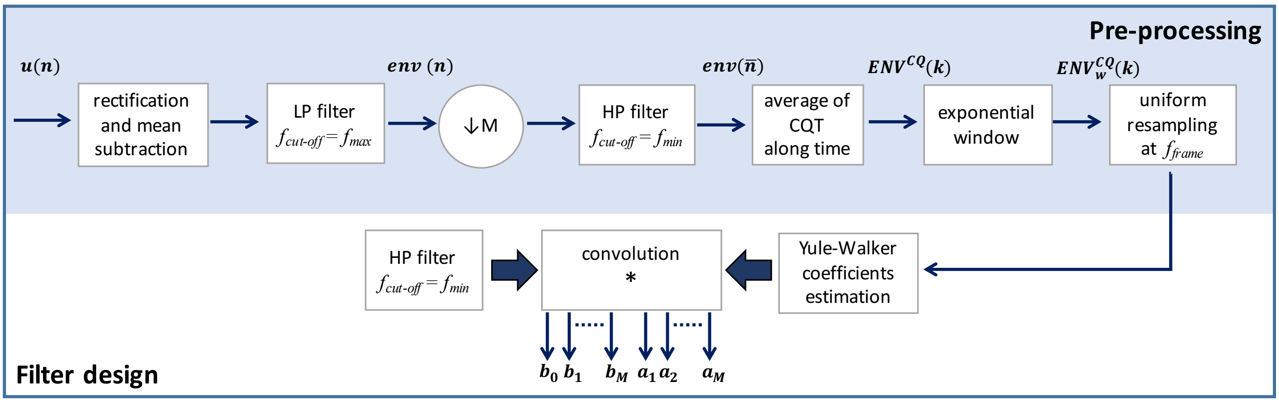
Pre-processing
$u(n)$
rectification and mean subtraction
LP filter $f_{cut\text{-}off} = f_{max}$
$env(n)$
↓M
$env(\bar{n})$
HP filter $f_{cut\text{-}off} = f_{min}$
average of CQT along time
$ENV^{CQ}(k)$
exponential window
$ENV_w^{CQ}(k)$
uniform resampling at f_{frame}
Yule-Walker coefficients estimation
convolution *
HP filter $f_{cut\text{-}off} = f_{min}$
b_0 b_1 b_M a_1 a_2 a_M
Filter design

Figure 1: *Block diagram of ARTE filter design.*

methods based on iterative filter banks would require the filtering of the input signal many hundreds of times [11]. Efficient computations of the CQT can be found in [12] and [13].

2.2. CQCC extraction

The cepstrum of a time sequence $x(n)$ is obtained from the inverse transformation of the logarithm of the squared-magnitude spectrum. The spectrum is usually obtained using the discrete Fourier transform (DFT) whereas the cepstrum is usually implemented with the discrete cosine transform (DCT). The cepstrum is an orthogonal decomposition of the spectrum. It maps N Fourier coefficients onto r independent, decorrelated cepstrum coefficients that characterise the speech signal:

$$CC(r) = \sum_{k=0}^{K-1} \log \left| X^{DFT}(k) \right|^2 \cos \left[\frac{r\left(k - \frac{1}{2}\right)\pi}{K} \right] \quad (3)$$

where $k = 0...K-1$ is the DFT index.

Since the CQT frequency scale is geometrically spaced, whereas the basis functions of the DCT are linearly spaced, cepstral analysis cannot be performed using (3) without modification. Instead, a spline interpolation method can first be applied in order to resample the geometric scale to a uniform, linear scale [1]. The cepstrum can then be obtained in the usual way by operating on the linearised CQT-derived spectrum $\bar{X}^{CQ}$. CQCCs are thus extracted according to:

$$CQCC(p) = \sum_{l=0}^{L-1} \log \left| \bar{X}^{CQ}(l) \right|^2 \cos \left[\frac{p\left(l - \frac{1}{2}\right)\pi}{L} \right] \quad (4)$$

where $p = 0...L-1$ and where l is the linear-scale index. The full CQCC extraction algorithm is described in [1].

3. Articulation rate (ARTE) filtering

This section presents an articulation rate filter specifically tailored to CQCC features.

3.1. Motivation

The motivation for this work stems from the success of relative spectral (RASTA) processing [14] used widely as a component of feature extraction. RASTA filtering applies a band-pass filter to the temporal trajectories of individual spectral feature components. The effect is to emphasize the components of a signal that are typical of natural speech, i.e. those which reflect a typical articulation rate. A number of extensions to RASTA filtering have been proposed in the literature, e.g. [15, 16].

With the conventional RASTA filtering of CQCCs giving somewhat disappointing results, this idea is extended here through the estimation of filter coefficients at the utterance level. Ordinarily, with typical articulation rates in the region of 1 to 16 Hz, this would be extremely challenging; 512 temporal samples at a sampling rate of 16 kHz would correspond to a spectral resolution of 31.25 Hz between two adjacent DFT samples. With greater resolution at lower frequencies, however, the CQT offers a natural solution to this problem.

The following thus reports a new approach to relative spectral processing based on the CQT. Its goal is identical to that of RASTA filtering, namely to emphasise the components of a speech signal that are indicative of the articulation rate. Unlike RASTA, however, the proposed filter is designed adaptively for each utterance.

3.2. Pre-processing

The pre-processing stage focuses on a frequency region of interest which is 2 octaves wider than the typical region of articulation, that is, from $f_{min} = 0.5$ Hz to $f_{max} = 32$ Hz. The approach is illustrated in Figure 1. First, the envelope $env(n)$ of an input utterance $u(n)$ is calculated through rectification and mean subtraction with low-pass filtering. The latter is a zero-phase 2^{nd} order Butterworth filter with cut-off frequency f_{max}. The envelope is then down-sampled by a factor $M = \frac{f_s}{10 f_{max}}$ (to improve efficiency while avoiding aliasing) before being processed with a high-pass zero-phase 1^{st} order Butterworth filter with cut-off frequency f_{min}. The result is a new band-passed envelope $env(\bar{n})$, which has a fixed sampling rate of $\bar{f}_s = 10 f_{max}$. $\bar{n}$ is the down-sampled discrete time index.

The time-frequency spectrogram $ENV^{CQ}(k, j)$ of the envelop $env(\bar{n})$ is then computed with the constant Q transform, where k is the frequency bin index and $j = 1...J$ is the temporal frame index, where J is the number of frames. In order to suppress temporal influence, the average spectrum is then determined and an exponential window $w(k)$ is used to remove the influence of the first and last octaves, thereby focusing on frequencies between 1 and 16 Hz:

$$ENV_w^{CQ}(k) = \frac{1}{J} \sum_{j=1}^{J} w(k) \cdot ENV^{CQ}(k, j) \quad (5)$$

The minimum and maximum frequency for CQT computation are also set to f_{min} and f_{max}, respectively. The number of

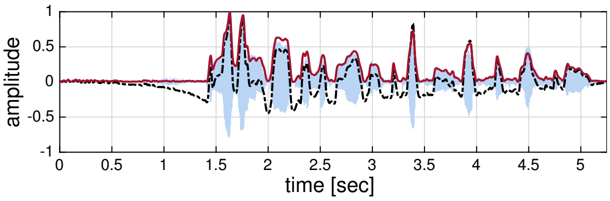

Figure 2: *Time representation of a speech utterance $u(n)$. The dotted line represents the envelope $env(n)$. The red solid line is the envelop $env(\bar{n})$ processed by the high-pass filter.*

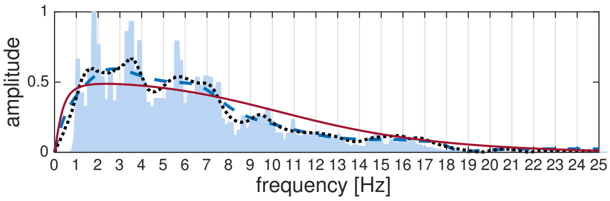

Figure 3: *The shaded area is the CQT spectrum of the utterance envelop, $ENV_w^{CQ}(k)$. The three lines represent the Yule-Walker filter response approximations for three different number of coefficients (solid=3, dashed=12 and dotted=24).*

bins per octave B is set to 96. The window $w(k)$ is defined as:

$$\begin{cases} e^{\frac{k-B}{\tau B}} & 1 \leq k < B \\ 1 & B+1 \leq k < B\,(oct-1) \\ e^{-\frac{k-[B(oct-1)+1]}{\tau B}} & B\,(oct-1)+1 \leq k < B \cdot oct \end{cases} \tag{6}$$

where τ is the time constant which is set to 10^{-1}, and *oct* is the number of octaves, which is set to 6.

CQCCs are computed according to Equation 4 for a frame-blocked signal with frame rate $f_{frame} = 100$ Hz. In order to cover the frequency range dictated by the frame rate, the frequency range of $ENV_w^{CQ}(k)$ is expanded from 0.5-32 Hz to 0-$\frac{f_{frame}}{2}$) Hz with an amplitude of zero, where $\frac{f_{frame}}{2}$ is the Nyquist frequency of the frame rate. This is achieved using the same approach as described in the original description of CQCC extraction in [1]: the geometrically spaced frequency scale of $ENV_w^{CQ}(k)$ is transformed to an extended linear frequency scale with a constant bin width of B/f_{min}.

3.3. Filter design and application to CQCCs

The ARTE filter coefficients may now be estimated. This is achieved with a modified Yule-Walker [17] autoregressive moving-average (ARMA) filter. The Yule-Walker equations provide a least-squares estimate of the parameters of a recursive infinite impulse response (IIR) digital filter for a specified frequency response. These coefficients are then convolved with those of a high pass 1^{st} order Butterworth filter with cut-off frequency f_{min}. The frequency response of the M-th order filter can be written in the usual way as:

$$H\left(z\right) = \frac{\sum_{m=0}^{M} b_m z^{-m}}{1 + \sum_{m=1}^{M} a_m z^{-m}} \tag{7}$$

where $H(z)$ is the transfer function in the z domain. Figures 2 and 3 illustrate the results of ARTE filtering for a real utterance and for different filter orders. All further work reported here was performed with an ARTE filter of order 3. Figure 4 illustrates the application of ARTE filter to the temporal trajectory of the 20-th CQCC coefficient of the utterance in Figure 2.

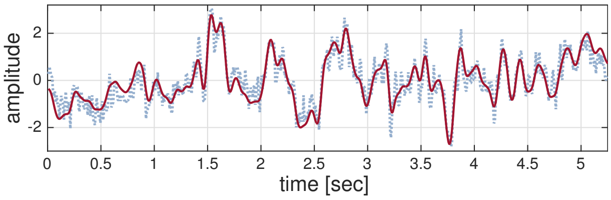

Figure 4: *20-th CQCC coefficient of the utterance in Figure 2 before (dotted blue line) and after (solid red line) ARTE filtering.*

Table 1: *Number of speakers (S), client models (Cl.) and trials for the RSR2015 part 1 and RedDots part 1 databases. Trial modes: Target-Correct (TC), Target-Wrong (TW), Impostor-Correct (IC) and Impostor-Wrong (IW). M=male, F=female.*

	RSR2015 DEV	RSR2015 EVAL	RedDots
#	M / F	M / F	M / F
S	157 / 143		35 / 6
Cl.	1492 / 1405	1708 / 1470	320 / 58
TC	8931 / 8419	10244 / 8810	3242 / 634
TW	259001 / 244123	297076 / 255490	29178 / 5706
IC	437631 / 387230	573664 / 422880	120086 / 4438
IW	6342019 / 5612176	8318132 / 6131760	1080774 / 39798

4. Experimental Setup

Presented in the following is an overview of the experimental setup including the databases, feature extraction and classifier.

4.1. Databases

Both RSR2015 [18] and RedDots [19] address short-duration, text-dependent ASV. Part 1 of both corpora, used in this evaluation, involve the use of fixed pass-phrases for authentication. There are four different types of trials, according to the speaker and text: target-correct (TC) refers to trials where both the speaker and text match. A target-wrong (TW) trial involves the target speaker but the incorrect pass-phrase. An impostor-correct (IC) trial involves an impostor speaker but the correct pass-phrase. Finally, in impostor-wrong (IW) trials, neither the speaker nor the text matches. Table 1 illustrates the number of speakers, client models and each type of trial for both the RSR2015 and RedDots databases. Clients are enrolled using 3 speech utterances and models are dependent on both speaker and pass-phrase.

4.2. MFCC extraction

Pre-emphasised speech signals are frame-blocked using a sliding window of 20 ms with a 10 ms shift. The discrete Fourier transform is applied to Hamming windowed frames to estimate the power spectrum before 19th order MFCCs (excluding the 0-th coefficient) are extracted using the discrete cosine transform (DCT) of 20 log-power, Mel-scaled filterbank outputs. RASTA filtering is then applied before delta and delta-delta coefficients are computed from the static parameters thereby resulting in feature vectors of dimension 57. Speech activity detection (SAD) based on energy modelling is applied to discard low-energy content. Finally, cepstral mean and variance normalization is applied to compensate for channel variation.

4.3. CQCC extraction and ARTE filtering

CQCC features are extracted as described in Section 2 and with a maximum frequency of $F_{max} = F_{NYQ}$, where F_{NYQ} is the Nyquist frequency of 8kHz. The minimum frequency is set to $F_{min} = F_{max}/2^9 \simeq 15$Hz (9 being the number of octaves).

Table 2: *Performance for RSR2015 and RedDots databases in terms of EER. R=RASTA filtering, A=ARTE filtering. Results illustrated independently for M=male and F=female trials.*

Condition	TW		IC		IW	
Gender	M	F	M	F	M	F
MFCC-R	2.42	0.51	2.66	1.39	0.26	0.07
CQCC	7.54	4.31	5.48	3.51	1.63	0.63
CQCC-R	3.56	1.22	3.21	1.77	0.65	0.13
CQCC-A	2.79	0.81	2.17	1.19	0.50	0.07
Fusion LR	**1.95**	**0.21**	**1.68**	**0.62**	**0.17**	**0.02**
LSVF [21]	2.33	0.64	2.64	1.61	0.31	0.06
HMM [22]	1.00	0.58	1.43	0.97	0.20	0.05

(a) *RSR2015 development set*

Condition	TW		IC		IW	
Gender	M	F	M	F	M	F
MFCC-R	0.98	0.41	1.54	1.42	0.13	0.09
CQCC-A	0.93	0.51	0.96	0.76	0.12	0.07
Fusion LR	**0.54**	**0.20**	**0.76**	**0.57**	**0.03**	**0.05**
LSVF [21]	0.97	0.49	1.58	1.72	0.12	0.05
HMM [22]	0.66	0.14	1.33	0.53	0.09	0.03

(b) *RSR2015 evaluation set*

Condition	TW		IC		IW	
Gender	M	F	M	F	M	F
MFCC-R	**5.12**	8.38	3.19	6.62	0.80	2.37
CQCC-A	8.36	7.41	3.98	5.09	1.49	2.05
Fusion LR	5.29	**6.48**	**2.35**	**4.42**	**0.52**	**1.42**

(c) *RedDots dataset.*

The number of bins per octave B is set to 96. These parameters result in a time shift or hop of 8ms. Static CQCC coefficients of order 29 are extracted and then processed with the ARTE filter described in Section 3. SAD is applied to remove low-energy content in identical fashion as for MFCCs before delta coefficients are appended, thereby resulting in feature vectors of dimension 58. Cepstral mean and variance normalization are again applied in the same way as for MFCCs. A Matlab implementation of CQCC-ARTE feature extraction is available online[1].

4.4. Classification and metrics

The classifier is based upon conventional Gaussian mixture models (GMMs) where speaker specific models are obtained from the maximum a posteriori (MAP) adaptation of a universal background model (UBM). Other experiments not reported here confirm that more advanced back-ends do not deliver superior performance in the case of short-duration training and testing [18], as is the case with the RSR and RedDots databases. A 512-component UBM is trained on the TIMIT database [20]. MAP adaptation is applied with a relevance factor of 10 and scores are the log-likelihood ratio given the target model and the UBM. Finally, performance is assessed in terms of equal error rate (EER).

5. Results

5.1. RSR2015

ASV results for the RSR2015 development and evaluation sets are illustrated in Tables 2a and 2b respectively. For the development set, the performance for MFCC features with RASTA filtering (MFCC-R) is compared to that for raw CQCC features, CQCC features with RASTA filtering (CQCC-R) and CQCC features with ARTE filtering (CQCC-A). CQCC-A shows the best performance among CQCC variants which is equivalent to that of MFCC-R features. Further experiments were thus conducted with only the MFCC-R and CQCC-A feature sets.

Logistic regression score fusion of MFCC-R and CQCC-A (Fusion LR) results obtained using the BOSARIS toolkit [23] are also illustrated in row 5 of Table 2a and row 3 of Table 2b. They show significant improvements in performance across both development and evaluation sets thereby illustrating the complementarity of MFCC-R and CQCC-A features. For female trials and the IC condition of the development set, a baseline EER of 1.42% drops to 0.57%, a relative reduction of 60%.

Comparative results from the literature are illustrated in rows 6 and 7 of Table 2a and rows 4 and 5 in Table 2b. Results reported in [21] were obtained using local spectral variability features (LSVF) and a similar standard GMM back-end. Results reported in [22] were obtained with the explicit modelling of time-sequence information by means of hidden Markov models (HMM). While MFCC and CQCC results are not dissimilar to those for LSVF, they are inferior to those for an HMM approach. Fusion results compare more favourably, in some cases even outperforming the HMM approach, even without the modelling of time-sequence information.

5.2. RedDots

Results for the RedDots database are illustrated in Table (2c). These results are produced with exactly the same configurations as those used for results obtained for the RSR2015 database. Once again results for MFCC and CQCC features illustrated in rows 1 and 2 show equivalent performance. Fusion results deliver almost universally consistent improvements in performance. This is an especially promising result given that no optimisation was performed for experiments with the RedDots database. For female trials and the IW condition, a baseline EER of 2.37% drops to 1.42%, a relative reduction of 40%.

6. Conclusions

This paper reports (i) the first application of constant Q cepstral coefficients (CQCCs) to automatic speaker recognition and (ii) a new articulation filter (ARTE). CQCC features are extracted with the constant Q transform, a perceptually-inspired alternative to Fourier-based approaches to time-frequency analysis. The ARTE filter performs an identical role to RASTA filtering which is commonly applied to Mel frequency cepstral coefficients (MFCCs).

On their own, CQCCs deliver equivalent performance to MFCCs for short-duration, text-dependent automatic speaker recognition. With different time and frequency resolutions, fusion experiments shows that CQCCs are complementary to MFCC features. These findings are demonstrated on both the RSR2015 and the more recent RedDots copora, the latter even without additional optimisation.

Further work could investigate the integration of CQCC features within alternative text-dependent approaches to ASV which explicitly model time-sequence information. It would also be interesting to determine whether or not the performance of CQCCs translates to longer duration trials such as those in the standard conditions of the NIST SRE datasets. Given longer duration training and testing, other research directions would also include the exploration of CQCC features in an i-vector framework. Finally, it is also of interest to determine whether ARTE filtering is beneficial to other features such as MFCCs.

[1] `http://audio.eurecom.fr/content/software`

7. References

[1] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop*, Bilbao, Spain, 2016.

[2] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilci, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *INTERSPEECH*, Dresden, Germany, 2015.

[3] J. Youngberg and S. Boll, "Constant-q signal analysis and synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, Apr 1978, pp. 375–378.

[4] J. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, January 1991.

[5] R. E. Radocy and J. D. Boyle, *Psychological foundations of musical behavior*. C. C. Thomas, 1979.

[6] E. Dorken and S. H. Nawab, "Improved musical pitch tracking using principal decomposition analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. ii, Apr 1994, pp. II/217–II/220 vol.2.

[7] W. J. Pielemeier, G. H. Wakefield, and M. H. Simoni, "Time-frequency analysis of musical signals," *Proceedings of the IEEE*, vol. 84, no. 9, pp. 1216–1230, Sep 1996.

[8] G. Costantini, R. Perfetti, and M. Todisco, "Event based transcription system for polyphonic piano music," *Signal Process.*, vol. 89, no. 9, pp. 1798–1811, Sep. 2009.

[9] G. Costantini, M. Todisco, R. Perfetti, R. Basili, and D. Casali, "Svm based transcription system with short-term memory oriented to polyphonic piano music," in *MELECON 2010 - 2010 15th IEEE Mediterranean Electrotechnical Conference*, April 2010, pp. 196–201.

[10] S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, 3rd ed. Academic Press, 2008.

[11] M. Vetterli and C. Herley, "Wavelets and filter banks: theory and design," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2207–2232, Sep 1992.

[12] G. A. Velasco, N. Holighaus, M. Dorfler, and T. Frill, "Constructing an invertible constant-Q transform with nonstationary Gabor frames," in *Proc. Digital Audio Effects (DAFx-11)*, 2011.

[13] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society (53rd Conference on Semantic Audio)*, G. Fazekas, Ed., AES (Vereinigte Staaten (USA)), 6 2014.

[14] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[15] J. de Veth and L. Boves, "Phase-corrected rasta for automatic speech recognition over the phone," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Apr 1997, pp. 1239–1242 vol.2.

[16] P. Yuan, M. Lin, K. Xiangli, L. Zhengqing, and W. Lei, "A study on echo feature extraction based on the modified relative spectra (rasta) and perception linear prediction (plp) auditory model," in *IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol. 2, Oct 2010, pp. 657–661.

[17] B. Friedlander and B. Porat, "The modified Yule-Walker method of ARMA spectral estimation," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-20, no. 2, pp. 158–173, March 1984.

[18] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication*, vol. 60, pp. 56 – 77, 2014.

[19] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brummer, D. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *INTERSPEECH*, 2015, pp. 2996–3000.

[20] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351 – 356, 1990.

[21] M. Sahidullah and T. Kinnunen, "Local spectral variability features for speaker verification," *Digital Signal Processing*, vol. 50, pp. 1 – 11, 2016.

[22] A. Larcher, K. Lee, P. L. S. Martinez, T. H. Nguyen, B. Ma, and H. Li, "Extended RSR2015 for text-dependent speaker verification over VHF channel," in *INTERSPEECH*, Singapore, 2014, pp. 1322–1326.

[23] N. Brümmer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," 2011.