Kadiri, Sudarsana Reddy; Alku, Paavo

Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/1.5131043

Published: 08/11/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Kadiri, S. R., & Alku, P. (2019). Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing. *Journal of the Acoustical Society of America*, *146*(5), EL418-EL423. https://doi.org/10.1121/1.5131043

Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing

Sudarsana Reddy Kadiri[1, a)] **and Paavo Alku**[1]

Department of Signal Processing and Acoustics, Aalto University, Espoo 12200, Finland

sudarsana.kadiri@aalto.fi,

paavo.alku@aalto.fi

(Dated: 4 July 2019)

Abstract: Voice source characteristics are known to vary in different phonation types due to tension of laryngeal muscles along with respiratory effort. In the present study, automatic classification of phonation types in singing is studied in four classes: modal/neutral, breathy, flow, and pressed. Existing studies in classification of phonation types in singing use voice source features and conventional mel-frequency cepstral coefficients (MFCCs) showing poor performance due to high pitch in singing. In this study, high-resolution spectra obtained using the zero-time windowing (ZTW) method is utilized to capture the effect of voice excitation. ZTW does not call for computing the source-filter decomposition which makes it robust to high pitch. For the classification of phonation types in singing, the study proposes extracting MFCCs from the ZTW spectrum. The results show that the proposed features give a clear improvement in classification accuracy compared to the existing voice source features and conventional MFCCs.

[a] Author to whom correspondence should be addressed.

1. Introduction

The human voice production mechanism is capable of affecting the type of phonation in voiced utterances by changing the vibration mode of the vocal folds. For speech signals, this gives rise to the coloring of speech with different voice qualities to signal, for example, vocal emotions (Gobl and Ní Chasaide, 2003). This capability is also manifested in singing affecting the timbre of the voice. Singer identity and the listener's feelings of the singer are expressed through modulations of voice quality. For example, a breathy voice has been reported to express sweetness, a pressed voice stronger expressions, and a flow voice very active singing (Sundberg, 1987, 1999).

According to Sundberg (1987, 1999), phonation types in singing can be categorized into four classes: modal (or neutral), breathy, flow (or resonant) and pressed (or tense). Sundberg expressed phonation types within a two-dimensional space spanned by subglottal pressure and glottal airflow (Sundberg, 1987, 1999). Sundberg's studies indicated that breathy and modal voices involve lower subglottal pressure levels than pressed and flow voices, while modal and pressed voices show reduced glottal airflow than breathy and flow voices.

Different phonation types primarily arise due to the adjustments in the larynx (Gobl and Ní Chasaide, 2003; Sundberg, 1987). In modal phonation, vocal folds vibrate fully along their entire length. In breathy phonation, there is a reduction in vocal fold abduction and minimal vocal fold contact area, which result in a high level of turbulent noise. The perceptual indicator of breathiness is the sensation of excessive laryngeal airflow (Childers and Lee, 1991; Grillo and Verdolini, 2008). Flow phonation corresponds to a vocal technique that is used exclusively in singing (Sundberg, 1995), and it is typically produced using a lowered larynx. Using flow phonation, higher levels of loudness can be achieved with less effort. Pressed phonation is associated with stronger muscular tension with an elevated larynx position, which influences the vocal tract shape.

Several studies (Airas and Alku, 2007; Proutskova *et al.*, 2013; Rouas and Ioannidis, 2016; Stoller and Dixon, 2016) have investigated phonation types in speech and singing with voice source features that have been derived from glottal flow estimates computed with glottal inverse filtering. The closing quotient of the glottal flow pulse and the difference between the amplitudes of the first two harmonics (H1-H2) in the voice source spectrum were shown to correlate with the amount of pressedness (Millgård *et al.*, 2016). The normalized amplitude quotient (NAQ), which measures the relative length of the glottal closing phase, was shown to be a more robust parameter than the conventional closing quotient for discriminating breathy, neutral, and pressed vowels (Airas and Alku, 2007; Kane and Gobl, 2013; Rouas and Ioannidis, 2016). The harmonics-to-noise ratio (HNR) as well as jitter and shimmer features were investigated by Wakasa *et al.* (2017) to discriminate pressed and neutral voices. The effects of the phonation type on the voice source were studied by Alku and Vilkman (1996) in female and male voices, and it was found that the maximum flow amplitude and the maximum flow declination rate (i.e., the negative peak of the glottal flow derivative) discriminated phonation types.

In conventional studies of phonation types, glottal features are typically first extracted from the source waveforms after which statistical tests are conducted to the computed features to compare their performance to discriminate phonation types. A different and more modern approach is to use machine learning to build an automatic classifier that is trained in a data-driven manner using voice source features. The first study in automatic classification of phonation types in singing was carried out by Proutskova *et al.* (2013) using voice source features derived with inverse filtering. Classification accuracy of phonation types (breathy, modal, flow, and tense) varied from 55% to 70% for various vowels and it was concluded that the voice source features alone are not sufficient for classification. This is mainly due to reduced accuracy of inverse filtering for singing voices, as singing voices are typically of high pitch and they show strong source-filter coupling. Several features including harmonic amplitudes, formant frequencies, formant bandwidths and amplitudes, HNR, and different voice source features were studied by Rouas and Ioannidis (2016). Their study showed that there are confusions between breathy and modal voices, and between voices produced in flow and pressed phonation. A large number of spectral statistics such as spectral centroid, spectral flux, spectral energies in different bands along with various voice source features and MFCCs were investigated by Stoller and Dixon (2016) for classification of phonation types in singing. Recently Kadiri and Yegnanarayana (2018a) proposed using cepstral features derived from the single frequency filtering (SFF) method that provides higher spectro-temporal resolution.

In the current study, high-resolution spectra obtained using the zero-time windowing (ZTW) method (Yegnanarayana and Gowda, 2013) are used to capture the effect of the voice excitation. For automatic classification of phonation types in singing, MFCCs computed using the ZTW spectrum are proposed as features. These novel features are used to train a support vector machine (SVM) classifier to conduct the automatic classification of the phonation type in singing.

2. Zero-time windowing (ZTW) and extraction of MFCCs using the ZTW spectrum

This section describes the signal processing methods used for deriving high-resolution spectrum with the ZTW method (Yegnanarayana and Gowda, 2013). In addition, the extraction of MFCCs using the ZTW spectrum is described. It is to be noted that the ZTW method does not assume the source-filter model of speech production.

2.1 ZTW

The objective of the ZTW method (Yegnanarayana and Gowda, 2013) is to derive the instantaneous spectrum so that the time-varying voice production characteristics can be captured. In this method, a voice signal to be analyzed is windowed with a heavily decaying window that provides high emphasis on the samples near the starting (zeroth) sampling instant, and hence the name zero-time windowing (ZTW). This happens for every instant of time and hence the technique provides high temporal resolution. Spectral characteristics are estimated using group delay, which provides good spectral resolution. Hence, the method provides higher temporal resolution while simultaneously maintaining spectral resolution.

Previous speech analysis studies have shown that the ZTW spectrum captures various characteristics of the excitation effectively, such as glottal opening and open phase, and also the vocal tract system, such as formants (Prasad and Yegnanarayana, 2016; Yegnanarayana and Gowda, 2013). Steps involved in extracting the instantaneous spectral characteristics using the ZTW method are as follows (Yegnanarayana and Gowda, 2013).

- The voice signal ($s[n]$) is first pre-emphasized in order to reduce the effects of low-frequency trend in the signal.

- Voice segment of $L\ ms$ (number of samples: $M = Lf_s/1000$) is considered at each instant, i.e., $s[n]$ is defined for $n = 0, 1, \ldots, M - 1$. The segment is multiplied with a window $w_1^2[n]$, where

$$\begin{aligned} w_1[n] &= 0, \qquad\qquad\qquad n = 0, \\ &= \frac{1}{4\sin^2(\pi n/2N)}, n = 1, 2, \ldots, N - 1. \end{aligned} \tag{1}$$

  N is the number of samples used in the computation of Discrete Fourier Transform (DFT) ($N >> M$). Multiplying the signal with the window $w_1^2[n]$ is approximately equivalent to integration in the frequency domain (Yegnanarayana and Gowda, 2013). Here, L=5 ms and N=1024 are used.

- Truncation of the signal at the instant $n = M - 1$ may result in a ripple effect in the frequency domain. The ripple effect is reduced by using another window ($w_2[n]$) for $n = 0, 1, \ldots, M-1$, defined as

$$w_2[n] = 2(1 + \cos(\pi\ n/M)) =\ 4\ \cos^2(\pi n/2M). \tag{2}$$

- The spectrum of the windowed signal (i.e., $x[n] = w_1^2[n]w_2[n]s[n]$) is estimated using the numerator of the group delay (NGD) function ($g_n[k]$) given by

$$g_n[k] = X_R[k]Y_R[k] + X_I[k]Y_I[k], \qquad k = 0, 1, 2, \ldots, N-1. \tag{3}$$

  where $X_R[k]$ and $X_I[k]$ are the real and imaginary parts of the N-point DFT $X[k]$ of $x[n]$. Likewise, $Y_R[k]$ and $Y_I[k]$ are the real and imaginary parts of the N-point DFT $Y[k]$ of $y[n] = nx[n]$.

- In order to highlight the hidden spectral characteristics due to heavily decaying window, the NGD function is differentiated twice. The spectral features, i.e., peaks in the spectrum correspond to the resonances of the vocal tract system.

- The Hilbert envelope (Yegnanarayana and Gowda, 2013) of the double-differentiated NGD is computed and is referred to as the ZTW spectrum, denoted by $S_n[k]$.

The steps involved in the ZTW method are shown in the schematic block diagram in Fig. 1. Figure 2 gives an illustration of ZTW spectrograms for soprano voices of different phonation types (breathy, modal, flow, and tense). It can be clearly seen that there exist remarkable spectral variations due to the voice excitation effects on the system characteristics.

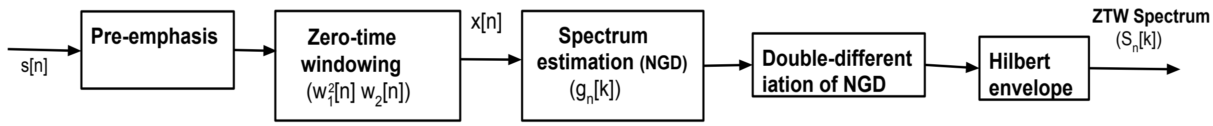

Fig. 1. Schematic block diagram describing the steps in the ZTW method.

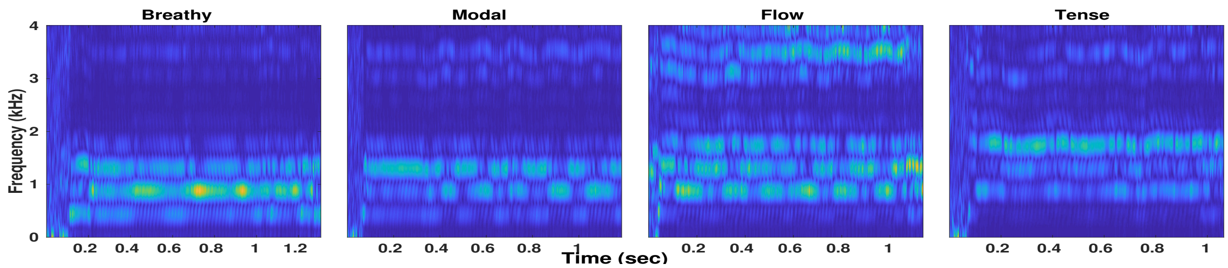

Fig. 2. (color online) An illustration of ZTW spectrograms for different phonation types (breathy, modal, flow, and tense) in soprano voices (vowel /A/).

2.2 Extraction of MFCCs using the ZTW spectrum

The schematic block diagram describing the steps involved in the extraction of MFCCs using the ZTW spectrum is shown in Fig. 3. The method performs the mel-filter bank analysis on the spectrum given by the ZTW method ($S_n[k]$), followed by logarithm and discrete cosine transform (DCT) operations, and can be expressed as follows

$$C_n[k] = DCT(\log(Mel(\mid S_n[k] \mid^2))), \tag{4}$$

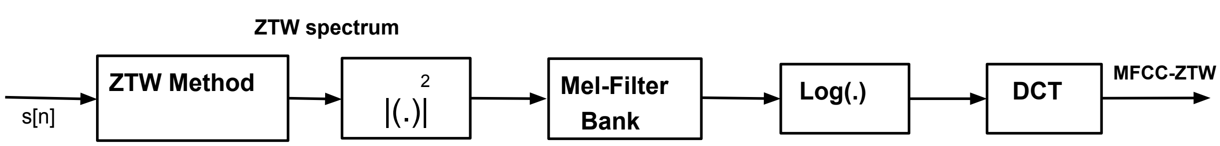

Fig. 3. Extraction of MFCCs using the ZTW spectrum.

where $c_n[k]$ denotes the mel-cepstrum. The resulting cepstral coefficients are referred to as MFCC-ZTW, and they represent compactly the effect of the excitation on vocal tract system characteristics. The MFCC-ZTW can in principle be obtained at each time instant. In this study, however, MFCC-ZTW is computed at glottal closure instants (GCIs). From the mel-cepstrum, the first 13 cepstral coefficients (including the zeroth coefficient) are considered for each frame. Delta and double-delta coefficients are also computed from the static coefficients.

3. Experimental protocol

This section describes the singing voice databases, the reference features, and the classifier.

3.1 Singing voice databases

Two databases (the soprano database and the baritone database) consisting of singing voices of different phonation types are used. The soprano database contains sustained vowels sung by a professional female Russian singer (Proutskova *et al.*, 2013). The phonation types correspond to Sundberg's definitions of breathy, modal, flow, and pressed voice (Sundberg, 1987). The database consists of 763 recordings in nine different vowels: /A, AE, I, O, U, UE, Y, OE, and E/. Pitch ranges from A3 to G5. The baritone database contains sustained vowels sung by a professional male Greek singer (Rouas and Ioannidis, 2016). The database consists of 487 singing voice samples of five vowels /A, O, E, I, U/. Pitch ranges from A2 to G4. Both of the databases were recorded at a sampling frequency of 44.1 kHz. More details of the databases are described by Proutskova *et al.* (2013) and Rouas and Ioannidis (2016).

3.2 Reference features

Five sets of reference features are considered for comparison based on recent studies on discrimination of phonation types (Kadiri and Yegnanarayana, 2018a,b; Kane and Gobl, 2013). These reference feature sets are: (1) conventional MFCCs, (2) voice quality (VQ) features, (3) excitation features derived from the modified zero frequency filtering (ZFF) method, (4) zero time windowing cepstral coefficients (ZTWCCs), and (5) single frequency filtering cepstral coefficients (SFFCCs). Conventional MFCCs are computed using Hamming-windowed frames of 25 ms with a shift of 5 ms. The VQ features are derived from glottal flow waveforms estimated by the iterative adaptive inverse filtering method (Alku, 2011). The VQ features consist of the normalized amplitude quotient (NAQ) (Alku *et al.*, 2002), the quasi-open quotient (QOQ) (Airas and Alku, 2007), H1-H2 (Airas and Alku, 2007; Mehta *et al.*, 2019), the parabolic spectral parameter (PSP) (Airas and Alku, 2007), the harmonic richness factor (HRF) (Airas and Alku, 2007), and the maximum dispersion quotient (MDQ) (Kane and Gobl, 2013). The excitation features are derived from approximate source waveforms computed using the modified ZFF method (Murty and Yegnanarayana, 2008). These features consist of the strength of excitation (SoE), the energy of excitation (EoE), the loudness measure, and the ZFF signal energy (Kadiri and Yegnanarayana, 2018a). Cepstral features are derived from the ZTW spectrum (Kadiri and Yegnanarayana, 2018b) and the SFF spectrum (Kadiri and Yegnanarayana, 2018a), and they are referred to as ZTWCCs and SFFCCs, respectively. For MFCCs, ZTWCCs, and SFFCCs, first 13 cepstral coefficients (static coefficients), delta, and double-delta coefficients are computed, which results in a 39-dimensional feature vector.

3.3 Classifier

Support vector machine (SVM) with a radial basis function (RBF) kernel is used as a classifier (Chang and Lin, 2011). Classification experiments are conducted using 10-fold cross-validation. One fold is held out to be used for testing, with the remaining nine folds used for training.

4. Classification experiments and results

Table 1 shows the results of the 10-fold cross validation experiments in terms of mean and standard deviation of the classification accuracy for the soprano and baritone voices. From the table, it can be seen that the proposed MFCC-ZTW features provide the highest average classification accuracy for both databases compared to the reference features. It is to be noted that the VQ features are not able to discriminate phonation types in singing as well as in speech (Kadiri and Yegnanarayana, 2018a; Kane and Gobl, 2013). This is because the VQ features (except MDQ) are derived from glottal flow waveforms estimated by inverse filtering whose accuracy is known to deteriorate for high-pitched voices (Alku, 2011). The performance of the existing MFCCs and SFFCCs is similar in both databases.

Tables 2 and 3 show the confusion matrices using the proposed MFCC-ZTW features for the soprano and baritone voices, respectively. From the tables, it can be observed that there is clear confusion between breathy and modal voices, and between flow and tense voices. It can also be observed that flow phonation shows confusion with voices of modal phonation.

Table 1. Mean and standard deviation of classification accuracy after 10-fold cross validation with different feature vectors for the soprano and baritone datasets.

Features	**Soprano** ($\mu \pm \sigma$)[%]	**Baritone** ($\mu \pm \sigma$)[%]
VQ	47.18 ±7.02	57.19 ±7.04
MFCCs	66.97 ±6.67	76.35 ±6.53
Excitation	52.11 ±5.92	47.11 ±6.12
ZTWCCs	66.06 ±6.13	74.62 ±4.38
SFFCCs	67.83 ±3.99	75.24 ±4.44
MFCC-ZTW	**80.32** ±3.74	**85.17** ±4.24

Table 2. Confusion matrix with 10-fold cross validation for the soprano database using the proposed MFCC-ZTW features.

	Breathy [%]	Modal [%]	Flow [%]	Tense [%]
Breathy	87.79	10.98	0	1.23
Modal	16.01	72.66	5.33	6.00
Flow	0.65	11.76	76.47	11.11
Tense	2.81	4.28	7.99	84.91

Table 3. Confusion matrix with 10-fold cross validation for the baritone database using the proposed MFCC-ZTW features.

	Breathy [%]	Modal [%]	Flow [%]	Tense [%]
Breathy	86.75	7.64	5.61	0
Modal	6.90	87.25	4.73	1.11
Flow	08.66	6.56	75.40	9.38
Tense	0	1.89	8.16	89.95

These observations are also in line with the results reported by Proutskova *et al.* (2013), by

Stoller and Dixon (2016), and by Rouas and Ioannidis (2016). Even though the proposed features show a remarkable improvement in accuracy, there is still confusion between modal and breathy, and between flow and tense voices. Hence, there is a need for exploring features that can capture changes in voice production characteristics, especially for the discrimination between breathy and modal voices, and between flow and tense voices.

5. Summary and conclusions

In this study, MFCCs derived using the ZTW spectrum were proposed for automatic classification of phonation types in singing. The ZTW method provides high-resolution spectra and captures the effect of excitation on the vocal tract characteristics. From the experimental results, it was shown that the proposed MFCC-ZTW features provide a better discrimination of phonation types in singing voices compared to several known reference features. The voice source features derived using glottal inverse filtering show less accuracy due to poor performance of inverse filtering in the estimation of the glottal flow for high-pitched voices in singing. The proposed features, however, do not suffer from this problem because they do not use source-filter separation for deriving features. This suggests that the proposed features can be useful for analyzing spontaneous, continuous speech in addition to singing voices. Despite the large improvement in classification accuracy achieved by the proposed MFCC-ZTW features in the current study, new voice production feature extraction techniques are needed to better discriminate breathy and modal voices, as well as voices in flow and tense phonation.

6. Acknowledgements

This study was partly funded by the Academy of Finland (project no. 312490).

References and links

Airas, M., and Alku, P. (**2007**). "Comparison of multiple voice source parameters in different phonation types," in *Proc. INTERSPEECH*, pp. 1410–1413.

Alku, P. (**2011**). "Glottal inverse filtering analysis of human voice production-a review of estimation and parameterization methods of the glottal excitation and their applications," Sadhana **36**(5), 623–650.

Alku, P., Backstrom, T., and Vilkman, E. (**2002**). "Normalized amplitude quotient for parametrization of the glottal flow," J. Acoust. Soc. Am. **112**(2), 701–710.

Alku, P., and Vilkman, E. (**1996**). "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," Folia Phoniatr Logop. **48**, 240 – 254.

Chang, C.-C., and Lin, C.-J. (**2011**). "Libsvm: A library for support vector machines," ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27.

Childers, D. G., and Lee, C. (**1991**). "Vocal quality factors: Analysis, synthesis, and perception," J. Acoust. Soc. Am. **90**(5), 2394–2410.

Gobl, C., and Ní Chasaide, A. (**2003**). "The role of voice quality in communicating emotion, mood and attitude," Speech Commun. **40**(1-2), 189–212.

Grillo, E. U., and Verdolini, K. (**2008**). "Evidence for distinguishing pressed, normal, resonant, and breathy voice qualities by laryngeal resistance and vocal efficiency in vocally trained subjects," J. Voice **22**(5), 546–552.

Kadiri, S. R., and Yegnanarayana, B. (**2018**a). "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC)," in *Proc. INTERSPEECH*, pp. 441–445.

Kadiri, S. R., and Yegnanarayana, B. (**2018**b). "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ztwccs)," in *Proc. INTERSPEECH*, pp. 232–236.

Kane, J., and Gobl, C. (**2013**). "Wavelet maxima dispersion for breathy to tense voice discrimination," IEEE Trans. Audio, Speech & Lang. Process. **21**(6), 1170–1179.

Mehta, D. D., Espinoza, V. M., Van Stan, J. H., Zañartu, M., and Hillman, R. E. (**2019**). "The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation," J. Acoust. Soc. Am. **145**(5), EL386–EL392.

Millgård, M., Fors, T., and Sundberg, J. (**2016**). "Flow glottogram characteristics and perceived degree of phonatory pressedness," J. Voice **30**(3), 287–292.

Murty, K. S. R., and Yegnanarayana, B. (**2008**). "Epoch extraction from speech signals," IEEE Trans. Audio, Speech, & Lang. Process. **16**(8), 1602–1613.

Prasad, R. S., and Yegnanarayana, B. (**2016**). "Determination of glottal open regions by exploiting changes in the vocal tract system characteristics," J. Acoust. Soc. Am. **140**(1), 666–677.

Proutskova, P., Rhodes, C., Crawford, T., and Wiggins, G. (**2013**). "Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing," J. New Music Res. **42**(2), 171–186.

Rouas, J.-L., and Ioannidis, L. (**2016**). "Automatic classification of phonation modes in singing voice: Towards singing style characterisation and application to ethnomusicological recordings," in *Proc. INTERSPEECH*, pp. 150 – 154.

Stoller, D., and Dixon, S. (**2016**). "Analysis and classification of phonation modes in singing," in *Proc. International Society for Music Information Retrieval Conference.*

Sundberg, J. (**1987**). *The Science of the Singing Voice* (Illinois University Press).

Sundberg, J. (**1995**). "Vocal fold vibration patterns and modes of phonation," Folia Phoniatrica et Logopaedica **47**(4), 218–228.

Sundberg, J. (**1999**). "The perception of singing," in *The Psychology of Music (Second Edition)* (Elsevier), pp. 171–214.

Wakasa, K., Matsubara, M., Hiraga, Y., and Terasawa, H. (**2017**). "Acoustic characteristics of pressed and normal phonations in choir singing by male singers," in *Proc. International Symposium on Musical Acoustics*, pp. 136–139.

Yegnanarayana, B., and Gowda, D. (**2013**). "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," Speech Commun. **55**(6), 782–795.