Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 31 (2015) 1–27

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

# A perceptually-motivated low-complexity instantaneous linear channel normalization technique applied to speaker verification

Victor Poblete[a,b], Felipe Espic[a,1], Simon King[c], Richard M. Stern[d], Fernando Huenupán[e], Josué Fredes[a], Nestor Becerra Yoma[a,*]

[a] *Speech Processing and Transmission Laboratory, Electrical Engineering Department, University of Chile, Santiago, Chile*
[b] *Institute of Acoustics, Universidad Austral de Chile, Valdivia, Chile*
[c] *Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK*
[d] *Department of Electrical and Computer Engineering and Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA*
[e] *Departamento de Ingeniería Eléctrica, Universidad de la Frontera, Temuco, Chile*

Received 7 February 2014; received in revised form 21 October 2014; accepted 30 October 2014
Available online 7 November 2014

## Abstract

This paper proposes a new set of speech features called Locally-Normalized Cepstral Coefficients (LNCC) that are based on Seneff's Generalized Synchrony Detector (GSD). First, an analysis of the GSD frequency response is provided to show that it generates spurious peaks at harmonics of the detected frequency. Then, the GSD frequency response is modeled as a quotient of two filters centered at the detected frequency. The numerator is a triangular band pass filter centered around a particular frequency similar to the ordinary Mel filters. The denominator term is a filter that responds maximally to frequency components on either side of the numerator filter. As a result, a local normalization is performed without the spurious peaks of the original GSD. Speaker verification results demonstrate that the proposed LNCC features are of low computational complexity and far more effectively compensate for spectral tilt than ordinary MFCC coefficients. LNCC features do not require the computation and storage of a moving average of the feature values, and they provide relative reductions in Equal Error Rate (EER) as high as 47.7%, 34.0% or 25.8% when compared with MFCC, MFCC + CMN, or MFCC + RASTA in one case of variable spectral tilt, respectively.

*Keywords:* Channel robust feature extraction; Auditorymodels; Spectral local normalization; Synchrony detection

## 1. Introduction

### 1.1. Motivation

The use of perceptually-motivated features is widespread across spoken language technology, with non-linear frequency scales and compression of the dynamic range of the spectral energy (*e.g.*, by taking the logarithm or cube

[*] Corresponding author. Tel.: +56 2 29784205.
*E-mail addresses:* vpoblete@ing.uchile.cl (V. Poblete), felipe.espic@ed.ac.uk (F. Espic), Simon.King@ed.ac.uk (S. King), rms@cs.cmu.edu (R.M. Stern), fhuenu@ufro.cl (F. Huenupán), jfredes@ing.uchile.cl (J. Fredes), nbecerra@ing.uchile.cl (N.B. Yoma).
[1] Now with the Centre for Speech Technology Research,University of Edinburgh, Edinburgh, UK.

http://dx.doi.org/10.1016/j.csl.2014.10.006
0885-2308/© 2014 Elsevier Ltd. All rights reserved.

root of filterbank outputs) being ubiquitous. In automatic speech recognition (Gales, 1998; Hermansky et al., 2013), speaker diarization (Tranter and Reynolds, 2006) and speaker verification (Reynolds and Rose, 1995; Campbell, 1997), Mel Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) or Perceptual Linear Prediction coefficients (PLPs) (Hermansky, 1990) are popular features, and in statistical parametric speech synthesis Mel-scaled features are also common (Tokuda et al., 2000).

Of course, the human auditory system performs operations far more sophisticated than warping the frequency scale and compressing dynamic range, but these are much less frequently found in speech processing applications. In this paper, we exploit an oft-neglected property of many auditory models: their ability to produce representations that are relatively invariant to changes in the channel. Our starting point is Seneff's auditory model (Seneff, 1988) and its two non-interacting parallel representations in the auditory nerve. One of these is the instantaneous mean rate of firing of neurons in individual nerve fibres (Tchorz and Kollmeier, 1999; Qin et al., 2008), whose counterpart is the usual spectral envelope employed in typical speech features such as MFCCs, where it is implemented as a filterbank (Bimbot et al., 2004). The other representation captures *synchrony* and is thought to be less variant in the presence of noise (Young and Sachs, 1979; Ali et al., 2002; Kim et al., 2006; Young, 2008; Kayser et al., 2009), and possibly also changes in the transmission channel (Rosen, 1992; Watkins and Makin, 1996; Tchorz et al., 1996).

In Section 2 we show the development of our idea. Taking inspiration from both the theoretical properties and the empirically-observed behavior of Seneff's Generalized Synchrony Detector (GSD) (Seneff, 1984, 1988), we propose a kind of local spectral energy normalization to compensate for variations in channel frequency response. We identify, and offer a solution for, a potential problem with the behavior of Seneff's model which may explain why some previous attempts to use the model directly in speech recognition applications (summarized in Section 2.2.1) have demonstrated only limited improvement in accuracy (Jankowski and Lippmann, 1992; Ohshima and Stern, 1994; Jankowski et al., 1995; Ali et al., 2000, 2002; Kim et al., 2006; Stern and Morgan, 2012a).

We then show in Section 3 how our proposed feature extraction method can be very simply realized within one stage of a typical frame-based procedure, with very little computational cost. The proposed feature extraction is memoryless and involves no time delays or look-ahead, and therefore does not add any latency to the system. The resulting features do not necessitate, in principle, any alterations to statistical models learned from them. To demonstrate the effectiveness of these features, we present results in Section 4 for a speaker verification task in which we observe that the proposed "self-normalizing" features are able to compensate for variations in the channel frequency response more effectively and conveniently than standard features (MFCCs). Additionally, their performance is competitive with MFCCs used in combination with conventional channel compensation techniques, such as Cepstral Mean Normalization (CMN) (Liu et al., 1993) and Relative Spectra (RASTA) (Hermansky and Morgan, 1994). However, whereas CMN requires a reliable estimate of the cepstral mean in the neighborhood of each frame being normalized, and RASTA requires computation of several frames in order to be stable, the proposed method LNCC performs normalization instantaneously within each frame without any external reference. Removing the requirement that the local cepstral mean must be estimated is advantageous in applications where the channel may be rapidly varying. This is because choosing the sliding window size over which the mean is estimated becomes difficult (*e.g.*, Hsu and Lee, 2009). It also leads to a simple and convenient frame-by-frame implementation that may be attractive in some situations. We provide experimental results to demonstrate that the proposed method is generally at least as good as CMN in all scenarios tested, and superior in the case of rapidly changing channels.

### 1.2. *The need for robust speech features*

In the current work, we restrict ourselves to dealing with (mostly linear) channels whose frequency response may differ between training and testing data, that may vary from one test utterance to the next or indeed within an utterance, and that is unknown at test time. We aim to extract features from the speech signal that are robust – by which we mean invariant – to changes in the channel (*e.g.*, Togneri and Pullella, 2011; Chen et al., 2003). Specifically, we target variations in the frequency response of acoustic channels which are a consequence of the relative position of the speaker with respect to the microphone.

#### 1.2.1. *Perceptual stability of speech sounds and airborne sound transmission-loss curves*

*Perceptual stability of speech sounds*: In this paper, one of the physiological motivations for choosing time-varying spectral tilts (either constant tilt, or varying tilt), to induce spectral contrasts in the acoustic channel of transmission is based on the fact that human speech perception is highly resilient to acoustic distortions related to the listening environment (Miettinen et al., 2012). In a normal auditory system, the cochlea performs a frequency decomposition of the speech sounds transmitted; the basilar membrane resonates with higher frequencies towards the basal entrance, and with lower frequencies progressively towards the apex. Thus the speech spectrum is organized spatially and it is transmitted through electrical potentials in auditory nerve fibres, representing speech sounds in a place code that is tonotopic (Miller et al., 1999). Physiological evidence in normal hearing suggests that in order to be effective, this internal sensorineural representation of the speech spectrum must maintain perceptual stability to changes in the global spectral energy distribution of the surrounding environment (such as those characterized by spectral tilts), and be able to normalize the effects of the surroundings on the speech spectrum (Stilp et al., 2010). In many ways this effect is quite similar to visual color constancy (Nassau, 1983). Nevertheless, it remains unclear how the human brain solves the problem of recognizing speech clearly while artificial speaker recognition systems struggle with this task (Kriegstein et al., 2010).

Distortions in the channel can arise, for example, due to changes of relative distance and orientation between the speaker and the microphone in locations such as the inside of a meeting room or in the corridor of a classroom. This is a problematic real-life source of variability that causes a non-stationary acoustic channel mismatched condition (Darwin et al., 1989; Hasan and Hansen, 2011). Moreover, in situations when speakers walk from room to room or where the microphone itself is moving, the distortion can be rapidly varying or slowly varying (Wang et al., 2007a,b, 2011). The existence of these rapid or slow transmission channel variations could distort the representation of the speech spectrum, altering its spectral envelope and impairing the corresponding automatic speech recognition or speaker verification performance (Soong and Rosenberg, 1988).

*Airborne sound transmission-loss curves*: The different representations of the variabilities in the acoustic channel proposed to model the channel distortions that may occur, especially in non-stationary environments, are also inspired by the airborne sound transmission-loss curves between spaces separated by some common building materials (*e.g.*, Crocker, 2007). The sound transmission loss (STL) characterizes the sound energy transmitted through a surface of a wall, door, or other building element (Vér and Beranek, 2006). It is defined as the logarithmic ratio of the incident sound energy relative to the transmitted sound energy (STL in dB) and can be estimated from theoretical considerations or determined from laboratory measurements (Fahy and Walker, 1998). Airborne sound transmission is measured over a range of audio frequencies between 100 and 5000 Hz, according to standardized tests (see *e.g.*, series of test standards ISO10140-2 (2010) and ISO140-III (1995)). The STL is widely used to be indicative of the perceived disturbance caused by various types of transmitted sounds (Park et al., 2008; Park and Bradley, 2009). The building materials could include common walls or single panels like doors or windows between classrooms, or between offices and meeting rooms which are of considerable importance *e.g.*, in acoustic design of spaces (Sato and Bradley, 2008). Here we propose to use varying spectral tilt to mimic the acoustic channel variability caused by transmitted sounds through partitions separating spaces.

There is evidence from laboratory tests that the STL varies considerably as a function of frequency (Crocker, 2007). For example, the typical STL behavior of a single uniform panel depends on sound frequency and the type of material (*e.g.*, wood, gypsum board, glass, steel, and concrete). In addition, the surface density, stiffness, and damping, shows characteristic frequency ranges and specific bands within which the STL is highly dependent on the incident sound frequency (Bies and Hansen, 2009).

It is possible to distinguish five characteristic frequency ranges (see *e.g.*, Long (2006), his Fig. 9.15; Norton and Karczub (2003), their Fig. 3.20). First, at very low frequencies the transmission loss is controlled by the stiffness of the panel which dominates the sound transmission characteristics. In this range the STL decreases with the frequency at the rate of −6 dB per octave. This consideration becomes a negative aspect in low frequency sound transmission problems particularly in lightweight panel construction (Tadeu and António, 2002). Second, at the frequency of the first panel resonance, the transmission of sound is almost complete and, as a consequence, the transmission loss passes through a minimum value determined by the damping of the panel. Third, at frequencies above the first panel resonance, there is a broad frequency range in which the sound reduction is primarily a function of the surface density of the panel (i.e., its mass). This frequency range is referred to as the mass law range, due to the approximately linear dependence on the mass of the panel (Fahy, 1987; Fahy and Gardonio, 2007). In this range the transmission loss increases with frequency at the rate of 6 dB per octave (Hansen, 2005). Fourth, in the region of coincidence at higher frequencies there is a sharp drop in the transmission loss, and damping controls the depth of the notch. Finally, at very high frequencies the transmission loss rises again at the rate of 9–10 dB per octave (Bies and Hansen, 2009; Norton and Karczub, 2003).

### *1.2.2. Time-varying channels*

A great number of approaches have been described in the literature to enhance the robustness of automatic speech and speaker recognition systems with respect to changes in the channel. We do not attempt a survey of these methods here, but refer the reader to, for example Seltzer and Bradley (2004), Buchner et al. (2005), Morales et al. (2009), Meyer and Kollmeier (2011), Lu et al. (2011). Typically, such methods attempt to improve recognition accuracy for cases where the training and testing data have been acquired under different acoustic conditions – for example, in order to enable the systems to cope with changes in microphone. Some methods aim to extract invariant features, while others attempt to adjust the statistical model. Our proposed method is of the former type, but could in principle be combined with model compensation techniques.

### *1.2.3. Application scenarios*

In numerous real applications, the channel between the speaker and the automatic speech recognition (or speaker verification, speaker diarization, etc.) system may vary over time. A few examples of such applications are mentioned below.

*Meeting transcription*: The task of richly transcribing human–human interactions has received considerable attention over the last decade (Hori et al., 2012; Yokoyama et al., 2013), particularly for the scenario of small business meetings with around 4 participants (Hain et al., 2006, 2007, 2012; Renals et al., 2007). A key problem in this domain is dealing with distantly-positioned microphones, such as those on the table-top, in randomly-positioned portable devices, or comprising microphone arrays. Tasks that are performed on speech captured in this way range from speech detection, speaker diarization, and transcription of the words, to higher-level analysis such as content-linking (Sangwan et al., 2013; Malionek et al., 2013).

The use of microphone arrays is widespread in this task domain, because of their ability to beamform, and thus to somewhat isolate the signal of a target speaker from other speakers or noise sources. Nevertheless, the properties of the physical acoustic channel between speaker and microphone (or microphone array) are still highly variable and are a cause of degradation in, for example, accuracy of transcriptions. Sources of variability in this channel include distance between speaker and microphone, beamforming arrays which low-pass filter off-axis signals (Brandstein and Ward, 2010 their Fig. 1.1), occlusion of the direct path by intervening objects such as laptop screens (Wölfel and McDonough, 2009 their Fig. 1.1), and so on.

We address one component of this complex puzzle, and – as will be justified in Section 4.5.1 – we will model the situation as an unknown and potentially time-varying spectral tilt imposed on the test recordings.

*Lecture transcription*: Another task that has received a growing level of attention recently is that of transcribing lectures (Trancoso et al., 2006; Bell et al., 2013). Typically, this task is performed from recordings made with lapel-microphones, which are used because they are relatively discrete compared to close-talking headsets. Unfortunately, this leads to frequent and rapid changes in the acoustic channel between speaker and microphone, due to head turning. While acceptably low error rates are possible in good conditions, when acoustic conditions degrade, the Word Error Rate (WER) can increase to 40–45% (Leeuwis et al., 2003; Park et al., 2005; Hsu and Glass, 2006; Glass et al., 2007). The alternative to lapel microphones is the use of distant microphones or arrays, but these are subject to similar problems, as described above.

*Human–machine interaction*: Speaking and hearing take place in situations where the acoustic environment is not constant and where speakers are affected by auditory input from the environment, other speakers, and feedback of their own speech (Cooke et al., 2013b). Background noise causes speakers to adjust their speech in a variety of ways (*see*, Cooke et al., 2014, 2013a, for a comprehensive review) including so-called 'Lombard' speech (Cooke and Lecumberri, 2012), in which one of the principal changes in addition to increased intensity is a reduction in the spectral tilt, leading to an overall flatter spectrum compared to normal. Often, speakers and listeners are also mobile, with each making continuous adjustments to speaking style and head position to compensate for the changing channel.

Machines that listen, whether they are socially-interactive mobile robots operating in public spaces such as supermarkets, museums, and expositions (Jensen et al., 2005; Ishi et al., 2008) or static systems using beamforming microphone arrays (Wölfel and McDonough, 2009) are faced with the same challenges of varying channel and speaking style – for example, the spectral tilt of the speaker's speech will varying with their speaking effort, which will change with the physical distance between speaker and 'listener' (robot or microphone array); the frequency response of a directional microphone will vary (typically with increased spectral tilt due to low-pass filtering) when the speaker is off-axis, compared to being on-axis (see Section 4.5.1 for experimental verification of this effect).

### *1.3. Scope of our work*

The auditory model-inspired features that we will introduce in Section 3 are specifically designed to be inherently and instantaneously robust to unknown and potentially time-varying channel frequency response, as present in the applications described in Section 1.2.3. Therefore, we limit the experimental investigation reported in Section 4 to such a scenario and do not address other aspects of robustness, such as additive noise or reverberation.

## 2. Development of the proposed approach from an auditory model

Models of the auditory system attempt to capture various behaviors of the natural system that they are mimicking (Stern and Morgan, 2012b). Some of these behaviors may be useful for speech feature extraction, so in this section we motivate our proposed features by starting from auditory models. We will identify a behavior which acts as a localized and instantaneous normalization, and which is not currently part of typical perceptually-motivated speech features used in pattern recognition applications.

One problem with such typical features, such as MFCCs or PLPs, is that they capture not only important speech features such as the frequencies of formants, but also channel properties too such as overall spectral tilt (Hansen and Varadarajan, 2009). Of course, a vast array of noise-robustness techniques is available to be applied either to these features, or to models learned from them. The features we propose are inherently less variant to channel differences than MFCCs.

### *2.1. Auditory modelling*

In speech technology, the most widely-used features are (usually decorrelated) representations of the envelope of the power spectrum (Wölfel, 2009a,b). On the other hand, auditory modelling has long known that the auditory system makes use not only of the spectral envelope but also information related to the synchrony between the responses in different nerve fibres (Johnson, 1980; Sachs, 1984; Eggermont, 1998; Dreyer and Delgutte, 2006). This synchrony-related information is more invariant to signal level differences than the rate-place representation of the spectral energy, is able to capture periodic signals even in the presence of noise, and so is thought (Smith et al., 2002; Moore, 2008; Heinz and Swaminathan, 2009) to be one of the reasons for the auditory system's incredible robustness to a wide variety of listening conditions, such as additive noise or channel distortion (Ghitza, 1994; Shao et al., 2010; Anderson et al., 2010).

#### *2.1.1. Mean rate representations and the spectral envelope*

Most conventional feature extraction schemes (such as MFCC and PLP coefficients) are based on short-time energy in a set of frequency bands, which is more directly related to mean-rate than temporal synchrony in the physiological responses of the auditory system (Davis and Mermelstein, 1980; Hermansky, 1990; Hermansky and Morgan, 1994; Dimitriadis et al., 2011). For example, the Mel-scaled filterbank, from which MFCCs are derived, captures the spectral envelope only (Kumaresan and Rao, 1999). The spectral envelope obviously carries information about both the speech signal and the transmission channel and any additive noise (Kuwabara and Sagisaka, 1995; Watkins and Makin, 1996; Zilovic et al., 1998; Parikh and Loizou, 2005; Miettinen et al., 2011). Separating these out *after* feature extraction is a blind separation problem and therefore only solvable by making some assumptions. A typical assumption would be that the channel changes more slowly than the speech spectrum (Stockham et al., 1975; Hermansky and Morgan, 1994; Gaubitch et al., 2013); this leads to a method in which a relatively long-term average is subtracted in the cepstral domain – Cepstral Mean Normalization (CMN) (Atal, 1974; Furui, 1981; Schwartz et al., 1993; Liu et al., 1993; Hermansky and Morgan, 1994). The disadvantage of this type of normalization is that it requires the estimation of the average cepstrum over some window (*e.g.*, all frames of the current utterance, or the previous *N* frames) (Soong and Rosenberg, 1988; Rose and Reynolds, 1990); if too short a window is used, then the estimated mean will contain some speech information, not just channel information. If the assumption about the channel changing slowly relative to the selected window/batch size is not correct, then the estimated mean will not accurately reflect the channel response and the normalization will be less effective (Bořil and Hansen, 2010; Nakano et al., 2010; Wang et al., 2011).

### 2.1.2. *Average localized synchrony rate (ALSR)*

In addition to mean rate representations, the auditory system is known to make use of another representation which captures temporal information, although precisely how the two are combined in the brain remains an open question (Moore, 2014). While temporal coding is clearly important for binaural sound localization (Stern et al., 2006; Joris and Yin, 2007), it may also play a role in the robust interpretation of signals from individual ears as well (Young, 2008).

For example, Young and Sachs (1979) demonstrated that the average localized synchrony rate (ALSR) that is derived from auditory nerve firing is much more robust to changes in intensity of vowel-like sounds than the corresponding mean-rate of response as a function of characteristic frequency (CF). The ALSR describes the extent to which the neural response at a given CF is synchronized to the nearest harmonic of the fundamental frequency of the vowel. These results suggest that the timing information associated with the response to low-frequency components of a signal can be substantially more robust to variations in intensity (and potentially various other types of signal variability and/or degradation such as varying channel or additive noise) than the mean-rate of the neural response.

A vast array of auditory models which include synchrony detection have been proposed (*e.g.*, Jankowski and Lippmann, 1992; Jankowski et al., 1995; Ali et al., 2002; Kim et al., 2006, for helpful reviews), and so we do not offer a survey of them here. Instead, we focus on the particular model that was the inspiration for the features we propose.

### 2.1.3. *From ALSR to Generalized Synchrony Detector (GSD)*

Seneff's auditory model (Seneff, 1988) consists of 40 recursive linear filters implemented in cascade form which cover a frequency range from 130 to 6400 Hz. The bandwidth of the channels is 0.5 Bark (Seneff, 1988). These filters mimic the nominal auditory-nerve frequency responses as described by Kiang et al. (1965) and other contemporary physiologists (Liberman, 1978; Young and Sachs, 1979; Sachs and Young, 1979; Sinex and Geisler, 1983; Delgutte and Kiang, 1984; Pickles, 2008). Seneff's model employs an "inner hair cell model" that includes four stages: (1) nonlinear half-wave rectification using an inverse tangent function for positive inputs and an exponential function for negative inputs, (2) short-term adaptation that models the release of transmitter in the synapse, (3) a lowpass filter with cutoff frequency of approximately 1 kHz to suppress synchronous response at higher input frequencies, and (4) a rapid automatic gain control (AGC) stage to maintain an approximately-constant response rate at higher input intensities when an auditory-nerve fibre is nominally in saturation.

Reflecting the fact that the auditory system makes use of two representations, Seneff proposed two non-interacting parallel modules that operate on the hair-cell model outputs. The first of these was an envelope detector, which produced a statistic intended to model the instantaneous mean-rate of response of a given fibre. The second operation was called a GSD, motivated by the ALSR measure of Young and Sachs (1979) and each channel $i$ is modelled (Seneff, 1985; Ali et al., 2002) as in Eq. (1), where $y[n]$ is the speech waveform value at sample $n$.

$$\mathrm{GSD}_i(y) = A_s \arctan\left[\frac{1}{A_s}\left(\frac{\langle|y[n]+y[n-n_i]|\rangle-\delta}{\langle|y[n]-\beta^{n_i}y[n-n_i]|\rangle}\right)\right] \tag{1}$$

The hair-cell output for this channel $i$ is compared to itself delayed by the reciprocal of the centre frequency $f_i^c$ of the filter in each channel ($n_i$ in Eq. (1)), and the short-time averages (i.e., envelope detection, denoted by $\langle\ldots\rangle$ in Eq. (1)) of the magnitudes (denoted by $|\ldots|$ in Eq. (1)) of the sums and differences of these two quantities are divided by one another. A threshold $\delta$ is introduced to suppress response to low-intensity signals and the resulting quotient is passed through a saturating half-wave rectifier (arctan [...] in Eq. (1)) to limit the magnitude (Seneff, 1985). A value slightly less than 1 is used for the constant $\beta$ in the denominator while the constant $\delta$ in the numerator has a rather small value (Seneff, 1985). The parameter $A_s$ represents a control in the linear range for the input speech waveform (Seneff, 1985; Ali et al., 2002).

With the limited computational resources available at that time, Seneff could only compare the mean-rate and GSD response visually for selected inputs. The GSD was a useful representation of the spectral components, including in noise (see *e.g.*, Seneff, 1985; Chigier and Leung, 1992; Jankowski and Lippmann, 1992; Ohshima and Stern, 1994). Newer and more sophisticated models than Seneff's have of course been proposed in more recent times (see Moore, 2003, 2014; Pickles, 2008; Stern and Morgan, 2012a, for a comprehensive review). Nevertheless, these newer models are not relevant to the work described in this paper because we are using a particular property of Seneff's model as the *inspiration* for our proposed method, rather than implementing the complete model.

## 2.2. The potential of GSD-like features for speech recognition

### 2.2.1. Previous attempts to use this model

The generalized synchrony detector model proposed by Seneff (1985) corresponds to one of the first attempts for developing a spectral representation from the temporal coding that occurs in the auditory nerve fibres (instead of their rate codes) for use as front ends to automatic speech recognition systems (Seneff, 1986b; Stern and Morgan, 2012a). Seneff reported strong evidences that auditory based representations are interesting and worthy of study in speech analysis systems. According to Seneff (1988) preliminary results of the two distinct spectral representations for the speech signal, one based on the discharge rate (rate coding) of the auditory nerve fibres and the other based on the synchronous response of the fibres (synchrony coding), indicated that the rate response outputs are successful for locating acoustic boundaries. Similarly, the synchrony outputs applied to speaker-independent vowel recognition in continuous speech showed superior performance (Seneff, 1987). However, there was no explanation on the neural interaction mechanisms between rate versus synchrony coding and how the auditory system uses some of the information of theses two representations in real communication situations (Smith et al., 2002; Moore, 2008).

Although Seneff's GSD has been used as a feature extraction method for speech recognition, such as the detection of formant frequencies (Seneff, 1984, 1986a; Kim et al., 1999), its performance compared to conventional mean-rate inspired features such as MFCCs (Jankowski et al., 1995; Ali et al., 2002) has been mixed. In general, in clean speech, GSD features provide recognition accuracies that are no better than what is provided by conventional MFCC or PLP features (and in some cases their performance is worse), but in additive noise, they can be helpful (*e.g.*, Chiu and Stern, 2008). An extension of the Seneff GSD was proposed by Ali et al. (2002). This approach, known as Average Localized Synchrony Detection, also produces a synchrony spectrum and provides better recognition results under noise conditions than the Seneff's original GSD detector.

Furthermore, the GSDs must be perfectly tuned to the formant frequencies in order to obtain a clean output (Seneff, 1988). This was a major problem of the original GSD algorithm (Ali et al., 2002).

### 2.2.2. A frequency-domain analysis of GSD

Seneff's original GSD is defined in the time domain by Eq. (1). Because it is more convenient to perform feature extraction for speech/speaker recognition in the frequency domain, we perform a frequency-domain analysis of Seneff's GSD by passing pure tones (sinusoids) at different frequencies sweeping the entire spectrum, using the time-domain filter of Eq. (1), which is effectively a form of frequency analysis. Phase is neglected, since it is generally believed that the human auditory system is phase insensitive (Meddis and Hewitt, 1991a,b). Nevertheless, there is some evidence from recent years that both phase and amplitude-envelope information may be relevant and potentially useful for improving both speech processing systems and human speech perception especially in speech-in-noise conditions, competing speakers, and in reverberant environments (Paliwal et al., 2011; Kleinschmidt et al., 2011; Chen et al., 2009; Shi et al., 2006). However, here we use only the magnitude response of the filters for simplicity, in the absence of compelling evidence that other combinations of phase and envelope information would provide better performance.

Eq. (1) is the ratio of two terms, numerator and denominator, which can be analyzed separately. Eqs. (2) and (3) give these terms, which are computed for each channel $i$ at each analysis frame.

$$\text{Numerator}_{\text{GSD}} = \langle |y[n] + y[n - n_i]| \rangle - \delta \tag{2}$$

$$\text{Denominator}_{\text{GSD}} = \langle |y[n] - \beta^{n_i} y[n - n_i]| \rangle \tag{3}$$

Consider, as an example, the response of the numerator and denominator terms of one GSD channel (after a band-pass filter) tuned to a center frequency $f_i^c$ of 692 Hz, to 1024 pure tones spanning the frequency range 60–3500 Hz, as shown in Fig. 1.

### 2.2.3. Spurious responses of the GSD

The frequency responses of numerator and denominator shown in Fig. 1 initially look promising, being centered at the tuned frequency of 692 Hz as expected, and with the denominator bandwidth being slightly wider than that of the numerator. Nevertheless, if we examine the final GSD response – the numerator divided by the denominator – as plotted in Fig. 2, we observe additional peaks at higher frequencies, along with the desired peak at 692 Hz. Seneff herself describes this limitation of the GSD (Eq. (1)), stating that it produces spurious peaks at harmonics of the

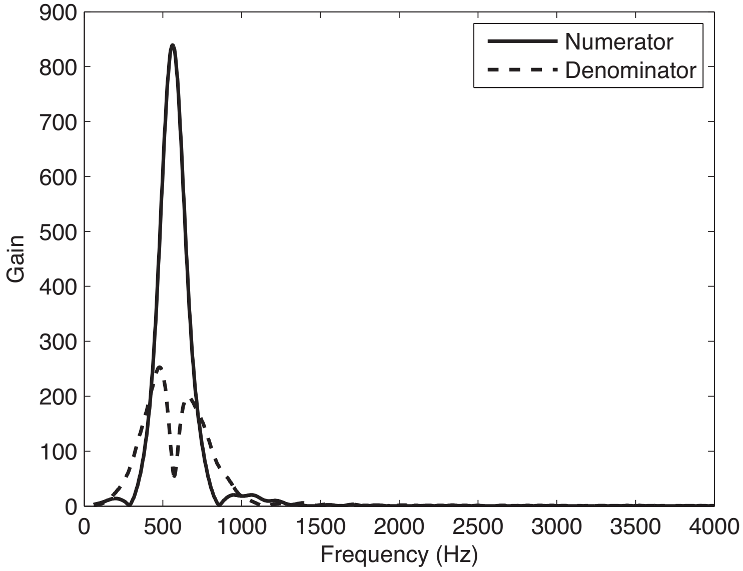

Fig. 1. Frequency response of the numerator and denominator of a GSD tuned at 692 Hz. The numerator in Eq. (2) is shown as a solid line and the denominator in Eq. (3) is shown as a dashed line. The values used for the constants are $\delta = 1 \times 10^{-5}$ and $\beta = 0.999$.

detected frequency. These observations notwithstanding, the behavior of each GSD channel in the region around its center frequency still has desirable properties, and we will exploit these in our proposed features described in Section 3 below.

Examining the numerator and denominator responses plotted on a logarithmic scale as in Fig. 3 reveals the cause of this behavior. In the figure, the denominator is plotted as its reciprocal to understand more clearly its relationship with the numerator. In the next section, we construct a GSD-like channel that preserves the desirable normalization behavior provided by the denominator term, but that does not produce spurious responses outside its nominal “passband”.

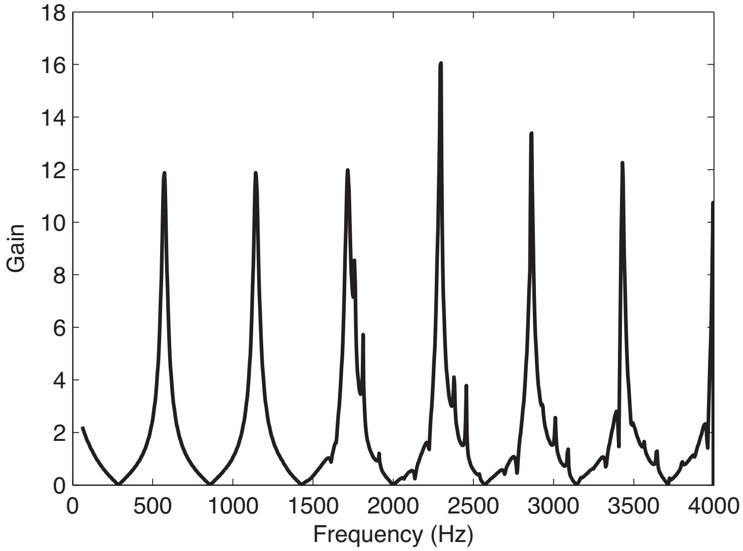

Fig. 2. Frequency response of a single GSD channel at $f_i^c$ = 692 Hz.

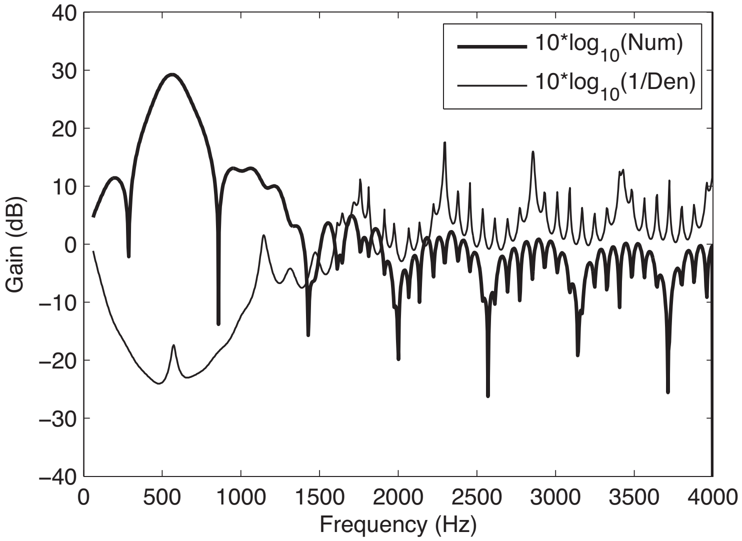

Fig. 3. Log magnitude of the frequency response of the numerator and denominator of the GSD tuned at $f_i^c = 692$ Hz. The numerator (Eq. (2)) is shown with a thick line and the reciprocal of the denominator (Eq. (3)) with a thin line.

## 3. The proposed features

Our target applications, described in Section 1.2.3, involve channels with time-varying frequency responses, including situations in which the physical arrangement of speaker and microphone may vary. We therefore seek features that are relatively invariant to changes in the channel frequency response. The proposed features achieve this by using a form of local normalization inspired by the ratio between the numerator and denominator terms of Seneff's GSD, given in Eqs. (2) and (3). We refer to these features as LNCC, for Locally-Normalized Cepstral Coefficients.

### *3.1. From the GSD to a frequency-domain model suitable for speech technology*

By examining the behavior of the GSD, we can identify some desirable attributes that are not found in typical features such as MFCCs. We note from the form of Eq. (1) and from Fig. 3 (ignoring for the moment the spurious higher-frequency responses)that the numerator part acts as a band-pass filter centered around a particular frequency, and that its output is divided by (i.e. normalized by) a denominator term which is a filter that responds to energy on either side of the numerator filter. In other words, a local normalization is being performed: the output of a GSD channel relates to the amount of energy in a particular frequency band *relative* to the energy in neighboring (lower and higher frequency) regions. With an appropriately-selected filter bandwidth, the effect is one of preserving spectral peaks (which are speech-related) while being relatively invariant to overall spectral tilt, for example. We note that the concept of a response in a localized central region being inhibited or suppressed by a response over a broader range of space or frequency is commonly encountered in vision (*e.g.*, Werblin et al., 1996) and audition (*e.g.*, Sachs and Kiang, 1968; Houtgast, 1972), and Wang and Shamma (1994), among others, have commented on the utility of this type of mechanism for speech recognition. We can achieve a similar behavior directly in the frequency domain, by designing simple filters for the numerator and denominator respectively (Fig. 5). Such a pair of filters will perform, in the frequency domain, a similar local normalization to that performed in the time-domain by GSD (Eq. (1)). By working in the frequency domain (just as in conventional MFCCs), the filters can easily be designed so as to only respond within the main passband, eliminating the spurious higher-frequency peaks seen in Fig. 2.

The pair of filters for one such channel were designed through informal experimentation and are show in Fig. 4. Responses of the pair of actual filters configured at a center frequency $f_i^c$ of 515 Hz are shown in Fig. 5. The numerator filter is essentially the same as the triangular filter commonly employed in the filterbank used to derive MFCCs (Davis

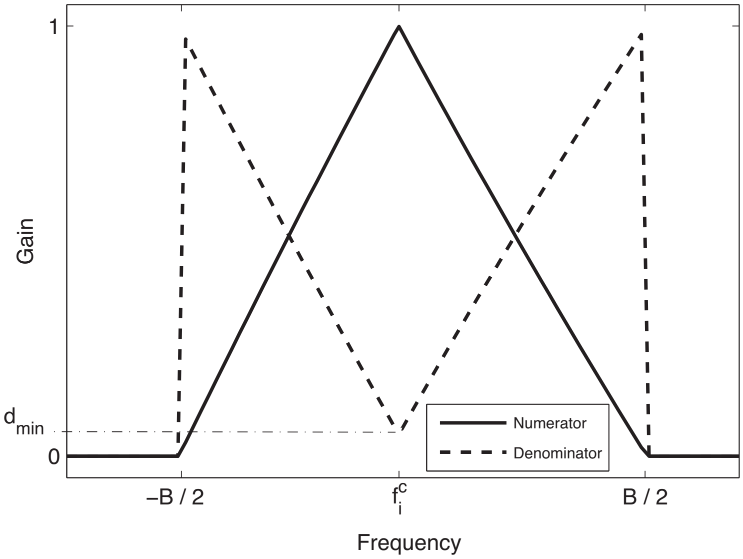

Fig. 4. The shapes of the magnitude of the numerator (solid line) and denominator (dashed line) filters, for a single channel of the proposed self-normalizing filterbank. $f_c$ is the center frequency of the channel, $d_{\min}$ the minimum centered value of the denominator, and $B$ its bandwidth. In our work, these frequencies are defined on the Bark scale (Zwicker, 1961).

and Mermelstein, 1980) and is described in the frequency domain by Eq. (4) for each channel $i$ with center frequency $f_i^c$. The denominator filter captures energy on either side of the numerator filter; it is described by Eq. (5).

$$\text{Numerator}_{\text{LNCC}}(f) = \begin{cases} -\dfrac{2}{B}|f - f_i^c| + 1 & \text{when } |f - f_i^c| \leq \dfrac{B}{2} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$\text{Denominator}_{\text{LNCC}}(f) = \begin{cases} \dfrac{2}{B}\,(1 - d_{\min})\,|f - f_i^c| + d_{\min} & \text{when } |f - f_i^c| \leq \dfrac{B}{2} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

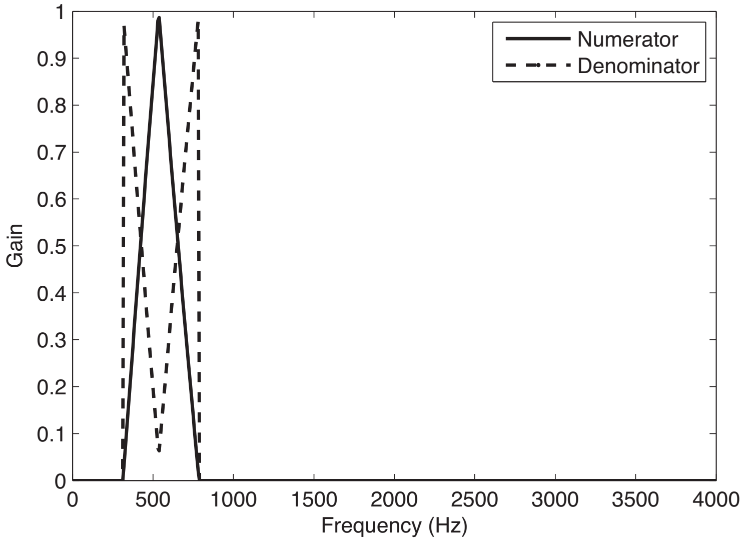

Fig. 5. Frequency response of the numerator and denominator tuned at $f_i^c = 515\,\text{Hz}$.

While both filters Numerator$_{\text{LNCC}}$($f$) and Denominator$_{\text{LNCC}}$($f$) have a nonzero response only for frequencies in the range of $-\frac{2}{B} \leq |f - f_i^c| \leq \frac{2}{B}$, it is easily seen that the response of Numerator$_{\text{LNCC}}$($f$) is greatest for a narrow range of frequencies about $f = f_i^c$, while Denominator$_{\text{LNCC}}$($f$) is responsive to activity in the surrounding frequency regions. By assembling a bank of such filter pairs, we can extract a locally-normalized filterbank representation of the signal, which can then be used subsequently to compute cepstral features, following the same steps as for deriving MFCCs from conventional filterbank outputs (Davis and Mermelstein, 1980). In all experiments presented in this paper, the filters are constructed on a Bark scale.

It is trivial to replace the filterbank normally used in MFCC feature extraction with this bank of self-normalizing filter pairs. By spacing the filters on a perceptual scale (such as the Bark scale used in our work) followed by logarithmic compression and a truncated cosine transform, we derive speech features that will have very similar properties to conventional MFCCs (*e.g.*, they are statistically decorrelated), but with the addition of the local normalization during the filterbank stage. The overall effect combines filtering with a filterbank (which removes fine detail from the spectrum such as harmonics of the fundamental frequency $F_0$) and local normalization (which removes very coarse variations in the spectral shape, such as overall tilt, which we assume arise mostly from channel variability).

In other words, the proposed features can be used as a straightforward “drop-in” replacement for MFCCs without any changes to the statistical model, for example. Fig. 6 describes the complete sequence of steps required to extract LNCC features, and shows the corresponding steps for conventional MFCC feature extraction for comparison.

### *3.1.1. Frequency response of the proposed self-normalizing filter pairs*

The frequency responses of the proposed numerator and denominator filters defined in Eqs. (4) and (5) are illustrated in Fig. 7 which plots the individual responses of one pair of numerator and denominator filters, on a logarithmic scale. The combined response of the numerator divided by the denominator is plotted in Fig. 8. This plot reveals that, when combined, the pair of filters in LNCC exhibits a sharper response than the triangular filters in a typical MFCC filterbank.

While Fig. 8 shows the response to pure tones, it is more useful to examine the response to a broadband signal (i.e. a vowel) in order to observe the normalization effect. Fig. 9 plots the spectral envelopes estimated by the proposed normalized filterbank, and compares this to the corresponding response of a conventional filterbank such as the filters typically used to derive MFCCs.

### *3.1.2. Robustness to channel mismatch*

In Fig. 10 we observe the response of the LNCC filterbank when speech is filtered by a channel with a non-flat frequency response, in this case, a spectral tilt of −6 dB/octave. The classical filterbank preserves the channel response in its output, whereas the normalized filterbank exhibits a response that is almost invariant to the channel response, while preserving key speech-related properties such as the spectral peaks. In the experiments presented in this paper, we compare the proposed features with conventional MFCCs which are optionally normalized using Cepstral Mean Normalization (Atal, 1974; Furui, 1981; Wang et al., 2007b). Additionally, other feature-based techniques are compared against the proposed features. In particular, we use combinations of MFCC and RASTA, and LNCC and RASTA filtering, respectively (Hermansky et al., 1991a,b; Hermansky and Morgan, 1994). Moreover, we compare the proposed approach with channel normalization techniques at the model level: joint factor analysis (JFA) (Kenny et al., 2008), and i-Vectors (Dehak et al., 2011).

It is important to note that almost all these other schemes require information outside the current frame being processed, and so are less effective for rapidly-varying channels (Leus and Moonen, 2003; Leus, 2004). For example: CMN requires an accurate estimate of the cepstral mean, which may be hard to obtain reliably in some cases (Qi Li et al., 2002); RASTA makes an equivalent assumption, that the channel changes substantially more slowly than the speech spectral envelope (Hermansky and Morgan, 1994).

## 4. Speaker verification experiments

To investigate the ability of the proposed features to normalize for varying channels, we conducted a sequence of speaker verification experiments on speech degraded by various channels. These involve simulated channels imposing spectral tilt which mimics the effect of off-axis or occluded microphones (Section 4.5) as well as spectral tilt characteristics that vary within a utterance (Section 4.6). For reasons of experimental control and repeatability, channel responses

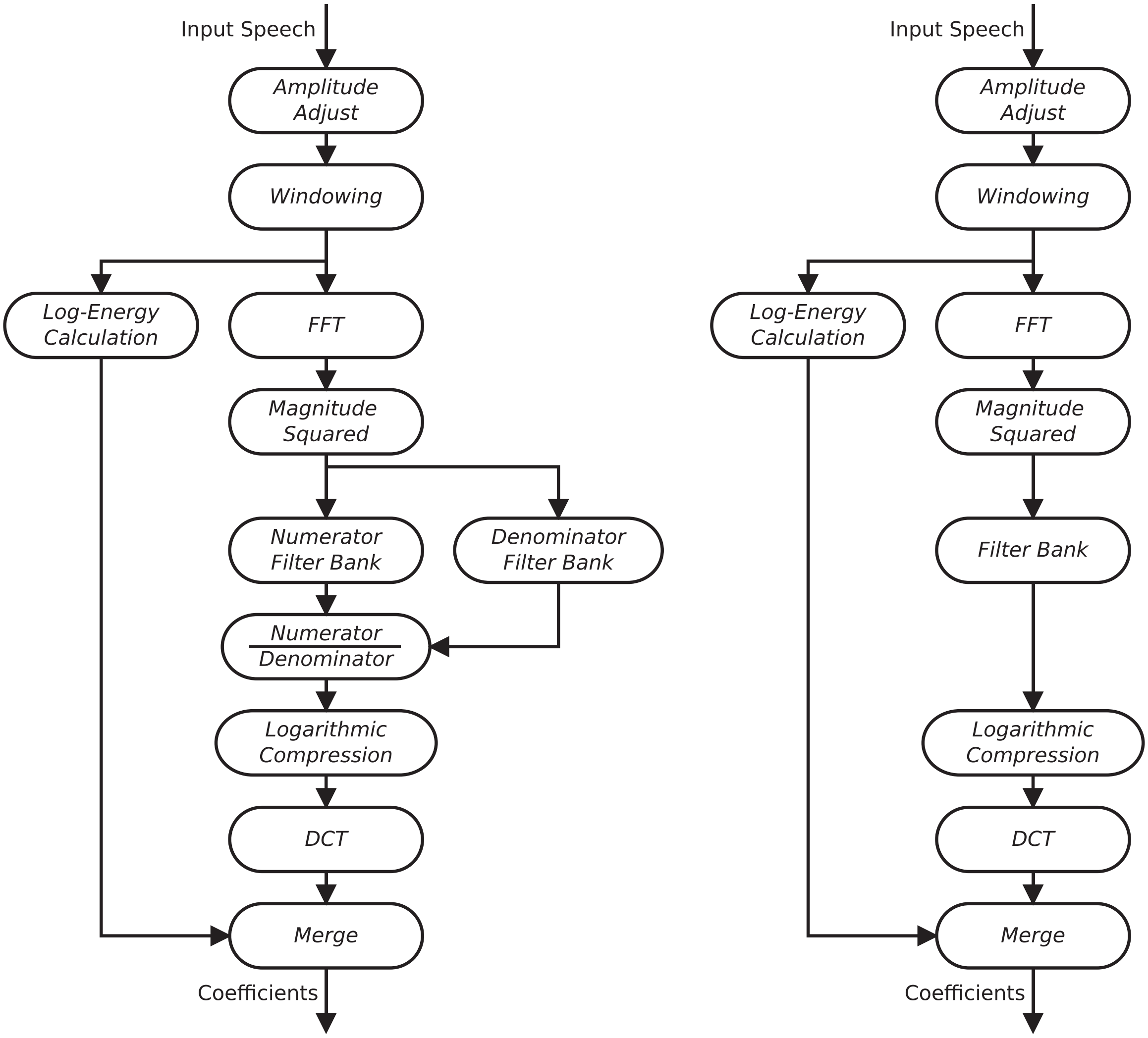

Fig. 6. Flowcharts for LNCC (left) and MFCC (right) feature extraction: note the similarity between the two, with the only difference being the normalization of the filterbank outputs in LNCC. It is common to append delta and delta–delta co-efficients; this is not shown in the diagrams.

were simulated. In all experiments, the system was trained using only clean speech. Test speech was degraded with respect to the training data by imposing static and time varying spectral tilt.

It is possible to apply noise suppression techniques to improve the quality of the estimated features. The simplest approach to feature normalization with MFCCs is Cepstral Mean Normalization (CMN) (Furui, 1981). Similarly, RASTA filtering (Hermansky and Morgan, 1994) applies a bandpass filter in the log-spectral or cepstral domain. This filter suppresses modulation frequencies that are not in the range of modulation frequencies typically associated with speech utterances (*e.g.*, slowly-varying convolutive channel variability would produce a low-frequency component of the modulation spectrum). CMN and RASTA filtering do not explicitly use any channel information (Kinnunen and Li, 2010).

The current trend in state-of-the-art in speaker verification systems is to model the feature vectors with a GMM-UBM, using-utterance dependent adapted GMM mean supervectors (*e.g.*, the concatenation of the mean vectors of the universal background model UBM, obtained by the MAP adaptation, can be interpretated as a supervector) as the features representing the speech segments, and model these supervectors employing factor analysis techniques (Hasan and Hansen, 2013). The technique of joint factor analysis (JFA) is used for this purpose (Kenny et al., 2007a,b; Yin et al., 2007). The JFA model considers the variability of a Gaussian supervector as a linear combination of the speaker and channel components (Kenny et al., 2008; Kenny and Dumouchel, 2004).

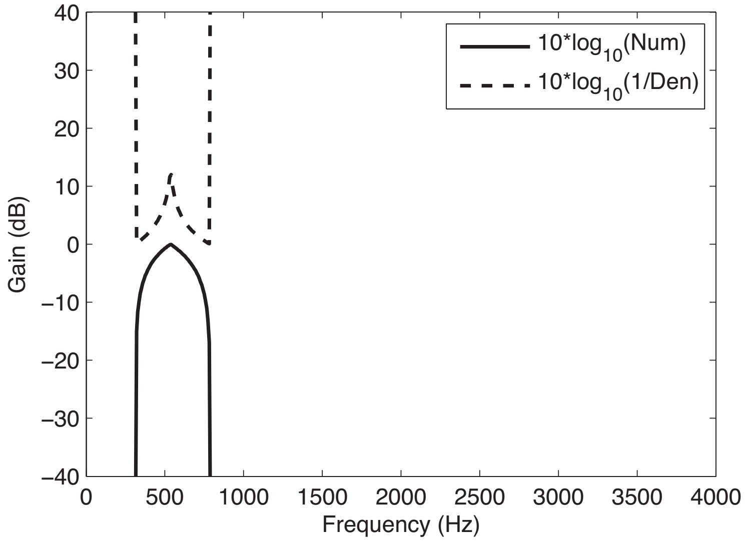

Fig. 7. Frequency response of the numerator and denominator separately, both tuned at $f_i^c = 515$ Hz, on a logarithmic scale.

The recently proposed i-Vector method (Dehak et al., 2011) utilizes a factor analysis framework to perform dimensionality reduction on the super-vectors while retaining important speaker discriminant information. In a standard i-Vector representation, a single model is used to represent the speaker and channel variability (Hasan and Hansen, 2013). It is worth emphasizing that JFA and i-Vector models have practical deficiencies. As discussed by Kinnunen and Li (2010) one of these practical deficiencies is sensitivity to training and test lengths, especially for short utterances (10–20 s). For that reason they may not be particularly accurate for short-duration utterances to the point where they may be outperformed even by a more traditional GMM-UBM approach (Hanilçi and Ertaş, 2013; Hautamäki et al., 2013).

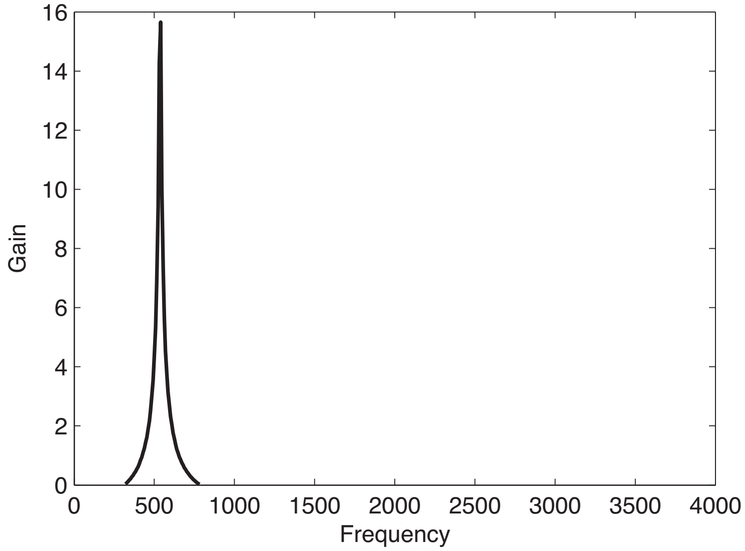

Fig. 8. Frequency response of the numerator divided by the denominator, both tuned at $f_i^c = 515$ Hz, on a logarithmic scale.

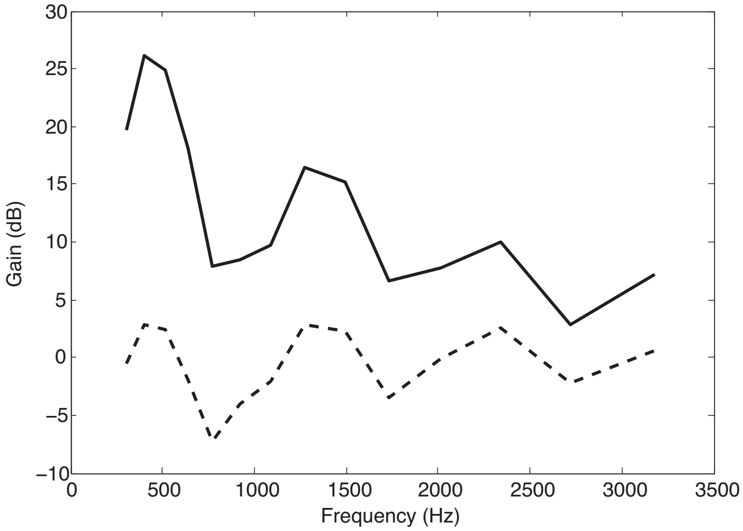

Fig. 9. Spectral envelopes (i.e. filterbank outputs plotted immediately after the logarithmic compression step in Fig. 6) for a single frame of voiced speech using a conventional Mel-scale filterbank (solid line), and for the proposed LNCC filterbank (dashed line). To aid readability, the solid line has been shifted by +15 dB. Observe that the proposed self-normalizing filterbank preserves important spectral shape information, such as the spectral peaks, but removes overall spectral tilt.

### 4.1. Speaker verification system

The experiments on text-independent speaker verification were carried out with ALIZE – Open Source Toolkit for state-of-the-art speaker recognition (Bonastre et al., 2008; Larcher et al., 2013). All experiments were based on the LIA-SpkDet toolkit (Bonastre et al., 2008), SPro (speech signal processing toolkit) (Bonastre et al., 2005), and the ALIZE library (Bonastre et al., 2004), and are derived from the work of Fauve et al. (2007). This software is based on a classical Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification system (Reynolds et al., 2000; Bimbot et al., 2004).

The Universal Background Model (UBM) is trained using background impostor speakers, with 256 Gaussian components using diagonal covariance matrices. A speaker-dependent Gaussian Mixture Model (GMM) is generated for each speaker by employing *maximum a posteriori* (MAP) adaptation (Reynolds et al., 2000). By doing so, the correspondence of the Gaussians within each speaker-dependent GMM with those in the background GMM is preserved (Reynolds et al., 2000).

Given a verification attempt where the identity of the speaker $s$ is claimed, $O$ denotes the observation sequence corresponding to the claimant's utterance. The output score of the system is a cohort-normalized log likelihood, $\log L(O)$:

$$\log\ L(O) = \log\ L(O/\lambda_s) - \overline{\log\ L(O/\lambda_{\bar{s}})} \quad (6)$$

where $\log L(O/\lambda_s)$ is the log likelihood of the client hypothesis and $\lambda_s$ is the speaker $s$ model, and $\overline{\log\ L(O/\lambda_{\bar{s}})}$ is the averaged log likelihood of the cohort of impostor models.

We include the use of channel normalization techniques at the model level, specifically, JFA (Kenny et al., 2008) and i-Vectors (Dehak et al., 2011). We employ a GMM-UBM with Joint Factor Analysis (JFA) model using subspace dimensions equal to 100 speaker factors and 10 session factor (Kenny et al., 2007a,b, 2008; Vogt et al., 2009). The configuration of JFA is chosen empirically. Similarly, we use a GMM-UBM with i-Vector model using a total variability subspace of dimension 300 (Kanagasundaram et al., 2011; Dehak et al., 2011). The configuration of i-Vector is also chosen empirically. As described by Yoma and Villar (2002), frames with higher local segmental SNR provide more reliable information than those with low segmental SNR. Also, voiced sounds (*e.g.*, vowels) show much higher speaker discrimination ability than fricative sounds. Accordingly, all the frames whose normalized energy with respect to the maximum utterance frame energy was lower than a given threshold are discarded.

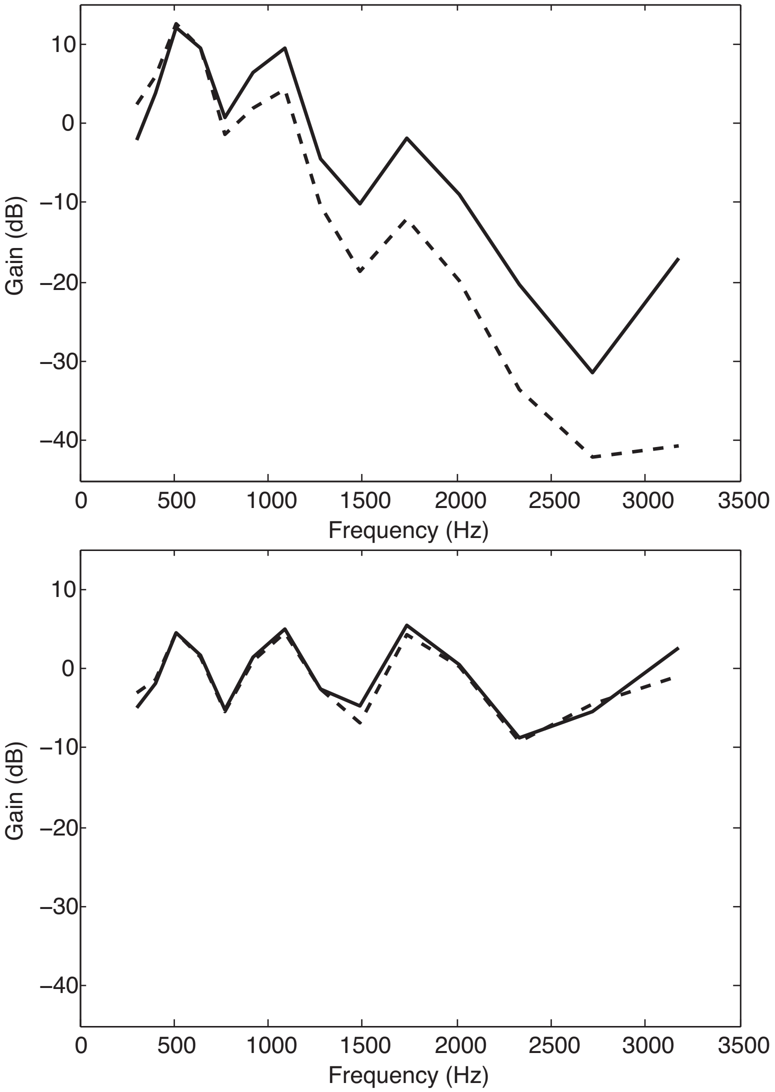

Fig. 10. Spectral envelopes (i.e., filterbank outputs plotted just after the logarithmic compression steps in Fig. 6) for a single frame of voiced speech using a conventional Mel-scale filterbank (upper figure), and using the proposed LNCC filterbank (lower figure). The responses to unmodified speech are shown in solid lines and the responses to speech filtered through a channel that imposes a −6 dB/octave spectral tilt are shown in dashed lines. Observe that the proposed features are invariant to the channel's spectral tilt, whereas the conventional filterbank outputs are highly sensitive to it.

### *4.2. Feature extraction*

Features were extracted using LNCC and MFCC processing, as described by Fig. 6. The frame duration in all cases was 25 ms with a 50% overlap. Frame selection was used to remove frames that do not carry useful information. This was performed by using frame based log energy detection with threshold equal to 2.5 dB below the maximum frame energy within the utterance. A frequency range from 200 to 3860 Hz was covered by 14 triangular filters uniformly arranged on a Bark scale, in the case of MFCCs, or 28 pairs (unless otherwise noted) of numerator and denominator filters uniformly arranged on a Bark scale in the case of the proposed LNCC features. If an LNCC channel goes beyond the range 0 Hz to Nyquist frequency, it is simply truncated. The DCT was truncated at 11 coefficients in both cases, then the first coefficient was replaced by the log frame energy. Finally, the resulting 11 coefficients are augmented with deltas and delta-delta to make up the final feature vector of dimension 33 for each frame.

### 4.3. Task

All experiments used the entire YOHO Speaker Verification Corpus, which comprises high quality recorded speech at 8 kHz sampling rate (Campbell and Higgings, 1994). YOHO supports the development, training, and testing of speaker verification systems with a vocabulary comprising two-digit numbers spoken continuously in sets of three (*e.g.*, "62-31-53" pronounced as "sixty-two thirty-one fifty-three"). The database was divided into enrollment and verification portions. The experiments were performed using 138 speakers (106 males and 32 females), four enrollment sessions per speaker with 24 utterances per session, and ten verification sessions per speaker with four utterances per session. These speakers were divided as follows: 40 background impostor speakers to train the background models and 98 test client speakers for use in verification attempts. For each speaker, one 96-utterance enrollment session was used. False rejection curves were estimated with 98 speakers × 40 verification signals per client = 3920 utterances. False acceptance curves were obtained with 98 speakers × 97 impostors × 40 verification signals per impostor = 380,240 experiments.

### 4.4. Initial experiments: sensitivity to parameter settings

Preliminary experiments were performed to determine how sensitive the proposed features are to the various parameters which must be chosen: the bandwidth of the filters (all filters have the same bandwidth on a Bark scale), the number of channels (the number of filters also determines their spacing, as a bandwidth that is too narrow would leave a "gap" in the filterbank's overall response), and the parameter $d_{\text{min}}$ which prevents division by zero at the centre frequency of each pair of numerator and denominator filters. As we described in Section 3.1.1, the LNCC filters exhibit a sharper response than the triangular filters in the MFCC filterbank. Therefore, we typically obtain best performance with a larger number of filters (*e.g.*, 28) than in the MFCC filterbank (which comprises 14 filters). All experiments regarding parameter sensitivity were performed with clean speech, and with speech processed through a channel with a −6 dB/octave spectrally-tilted frequency response.

#### 4.4.1. Number of LNCC channels and filter bandwidth

As can be seen in Fig. 11, performance on clean speech is relatively unaffected by the bandwidth until it becomes too narrow – this is presumably because at narrow bandwidths with a constant number of channels, "gaps" start to appear between filters and speech information is then missed. For spectrally-tilted speech, the same effect is seen with narrow bandwidths, but we also observe a worsening of performance at wider bandwidths. This is presumed to be a consequence of the local normalization becoming "less local" and therefore less effective.

Also in Fig. 11 it is observed that 28 LNCC channels leads to lower EER than 14 filters. Although not presented in this paper, further experiments were carried out with 20 and 56 LNCC channels. However, those configurations did not lead to significant improvements in Equal Error Rates when compared to 28 channels.

#### 4.4.2. Denominator minimum centre value ($d_{min}$)

Fig. 12 describes EER as a function of $d_{\text{min}}$ for clean speech and speech corrupted by −6 dB/octave spectral tilt. The LNCC coefficients were computed using 28 channels, and a bandwidth $B = 3.5$ Barks. According to Fig. 12 there is a wide range of values for $d_{\text{min}}$ ($0 \leq d_{\text{min}} \leq 0.01$) for which EER shows little variation.

### 4.5. Experiment 1: simulated distant microphone and acoustic obstacles

In this first experiment, we consider channels with *mismatched* spectral tilts. The spectral tilt of the channel through which the test utterances have passed is different (except for the 'clean' test condition) from that of the enrollment data, which were always clean.

#### 4.5.1. Simulating the frequency response of a distant microphone

As mentioned in Section 1.2.3 one of the consequences of using a distant, off-axis, or occluded microphone or microphone array to capture speech is that some unknown spectral shaping will be imposed on the speech by the channel. Speech produced with increased vocal effort may also vary the spectral tilt with respect to clean speech. Overall the effect is one where the test speech has a different average spectral shape to the clean training speech.

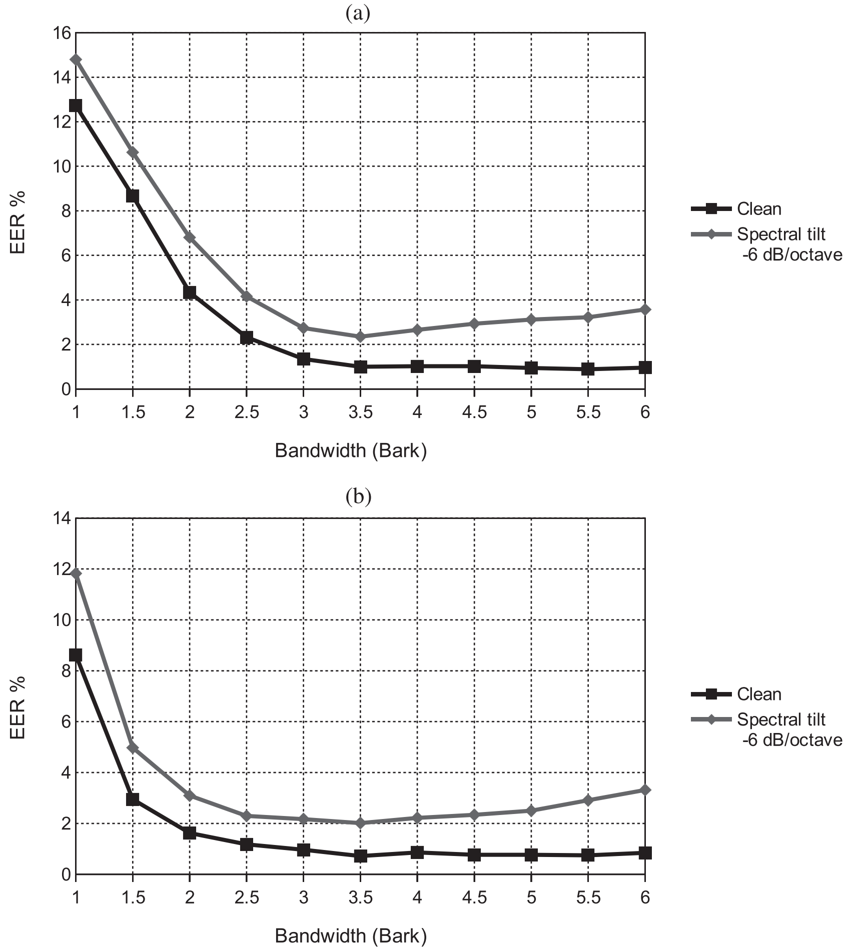

Fig. 11. Sensitivity to the filter bandwidth. Both with $d_{min} = 0.001$. (a) 14 channels, (b) 28 channels.

We simulated this using a simple filter that imposes −3 dB/octave, −6 dB/octave or −9 dB/octave spectral tilt. These particular values were motivated by the characteristics of the sound transmission loss curves mentioned in Section 1.2.1, and verified in informal experiments in which we re-recorded speech reproduced over a loudspeaker with the microphone set off-axis, or speech recorded with occlusions placed between the loudspeaker and microphone.

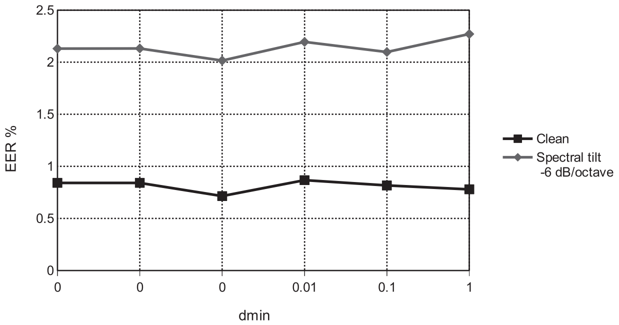

Fig. 12. Sensitivity to the $d_{\min}$ parameter. LNCC with 28 channels, $B = 3.5$ Barks.

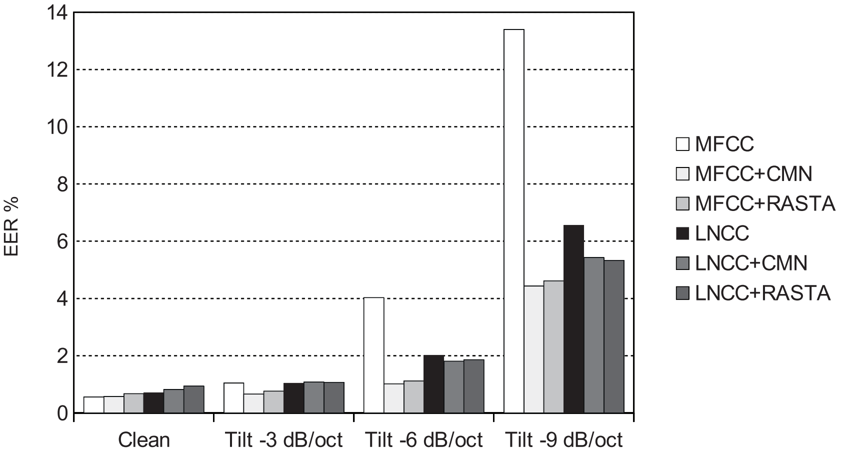

Fig. 13. Performance for clean and constant spectral tilt conditions. LNCC features are computed using 28 channels, $d_{\text{min}} = 0.001$ and $B = 3.5$ Barks.

### *4.5.2. Results*

Fig. 13 presents Equal Error Rates (EERs) for speaker verification obtained using both the proposed LNCC features and standard MFCC features, in conditions that are clean or have a constant spectral tilt of −3 dB/octave, −6 dB/octave or −9 dB/octave. Results for both MFCC and LNCC in combination with CMN or RASTA are also shown. The exact EERs are provided, along with those for Experiment 2, in Table 1. Note that, because neither the i-Vector or the JFA techniques offer good performance, we do not plot these results in the figures, but we do include them in Table 1 for completeness. The poor performance of i-Vectors or JFA is consistent with results from the literature (*e.g.*, Dehak et al., 2011; Kanagasundaram et al., 2011; Mandasari et al., 2013; Hautamäki et al., 2013; Kenny et al., 2013; Hasan and Hansen, 2013; Larcher et al., 2014); i.e., they only work well for relatively long utterances, which is not the case in our experiments.

LNCC features provide substantial and statistically significant relative reductions in EER as high as 49.9% ($p < .001$) and 51.0% ($p < .001$) compared to the MFCC baseline, at constant spectral tilts of −6 dB/octave and −9 dB/octave, respectively. These results suggest that LNCC is far more robust than MFCC to constant spectral tilt.

Table 1
Summary of results. LNCC features are computed using 28 channels, $d_{\text{min}} = 0.001$ and $B = 3.5$ Barks.

| Test data | Equal Error Rate (EER) % | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | MFCC | MFCC CMN | MFCC RASTA | MFCC i-Vectors | MFCC JFA | LNCC | LNCC CMN | LNCC RASTA |
| Clean | 0.561 | 0.579 | 0.673 | 2.05 | 3.60 | 0.714 | 0.819 | 0.939 |
| Constant −3 dB/octave | 1.05 | 0.66 | 0.765 | 2.42 | 3.44 | 1.05 | 1.08 | 1.07 |
| Constant −6 dB/octave | 4.03 | 1.02 | 1.12 | 5.71 | 7.83 | 2.02 | 1.81 | 1.86 |
| Constant −9 dB/octave | 13.4 | 4.43 | 4.62 | 15.7 | 18.6 | 6.56 | 5.43 | 5.33 |
| 0 to −6 dB/octave STEP | 1.68 | 1.07 | 1.07 | 3.04 | 4.13 | 1.48 | 1.43 | 1.27 |
| 0 to −6 dB/octave | 1.35 | 0.79 | 0.94 | 2.70 | 3.98 | 1.24 | 1.17 | 1.10 |
| 0 to −6 to 0 dB/octave STEP | 1.40 | 1.34 | 0.887 | 3.37 | 3.88 | 1.38 | 1.47 | 1.30 |
| 0 to −6 to 0 dB/octave | 1.15 | 1.01 | 0.84 | 2.63 | 3.52 | 1.19 | 1.28 | 1.22 |
| 0 to −6 to 0 to −6 dB/octave STEP | 1.63 | 0.94 | 1.05 | 3.40 | 4.06 | 1.56 | 1.35 | 1.35 |
| 0 to −6 to 0 to −6 dB/octave | 1.24 | 0.75 | 0.92 | 2.65 | 3.65 | 1.25 | 1.15 | 1.19 |
| 0 to −9 dB/octave STEP | 12.9 | 9.68 | 8.56 | 9.87 | 11.4 | 6.48 | 4.68 | 4.43 |
| 0 to −9 dB/octave | 5.48 | 3.17 | 2.87 | 5.59 | 7.91 | 2.73 | 2.23 | 1.91 |
| 0 to −9 to 0 dB/octave STEP | 6.15 | 5.49 | 3.57 | 9.85 | 8.78 | 3.80 | 3.93 | 3.67 |
| 0 to −9 to 0 dB/octave | 2.49 | 2.02 | 1.12 | 4.63 | 5.29 | 1.73 | 2.07 | 1.79 |
| 0 to −9 to 0 to −9 dB/octave STEP | 13.0 | 10.3 | 9.16 | 11.0 | 12.8 | 6.80 | 5.15 | 4.75 |
| 0 to −9 to 0 to −9 dB/octave | 5.30 | 2.91 | 2.83 | 5.56 | 7.58 | 2.65 | 2.18 | 2.02 |

When CMN is applied to LNCC, further significant relative reductions in EER of 10.4% ($p < .001$) and 17.2% ($p < .001$) are achieved over LNCC alone, for the constant spectral tilt conditions of −6 dB/octave and −9 dB/octave, respectively. When RASTA is applied instead of CMN, these relative reductions in EER become 7.82% ($p < .001$) and 18.7% ($p < .001$) at constant spectral tilts of −6 dB/octave and −9 dB/octave, respectively. However, LNCC does not benefit as much from these additional normalization techniques as much as MFCC does.

In summary, the results under conditions of constant spectral tilt demonstrate that:

- LNCC is more robust to spectral tilt that MFCC.
- CMN or RASTA can improve the performance of LNCC, although less effectively than for MFCC.
- LNCC alone provides an EER that is much closer to the performance of MFCC+CMN or MFCC+RASTA than MFCC alone, but with a simpler implementation than CMN or RASTA processing schemes.

### 4.6. *Experiment 2: rapidly-varying channels*

We now extend the scenarios covered in Experiment 1 to include *time-varying* channels. We simulate a number of different channels, with varying amounts of spectral tilt, and varying rates of change over time.

#### 4.6.1. *Signal processing to simulate a time-varying channel response*

A dynamic filter was designed in order to modify the spectral tilt over time. Specifically, a 1024-point FIR filter with the desired spectral slope was applied on a frame-by-frame basis to the incoming signal. The slope of the target tilt linearly varies between 0 dB/octave and a lower bound expressed in dB per octave, within each utterance. In our experiments, the two values of −6 dB/octave and −9 dB/octave were used as lower bounds. After applying the spectral tilt, the energy of each frame was normalized to compensate for the attenuation produced by the spectral tilt filter. The filter was applied only to the speech portions of each test file, as found by end-pointing. Three types of time-varying spectral tilts were constructed using the filter:

- Slow_tilt_1: the spectral tilt changes over time from 0 dB/octave at the start of the speech portion, to the lower bound (e.g., -6 dB/octave) by the end of the speech portion.
- Slow_tilt_2: the spectral tilt changes over time from 0 dB/octave at the start of the speech portion, to the lower bound at the middle of the speech portion, then back to 0 dB/octave by the end of the speech portion.
- Slow_tilt_3: the spectral tilt changes over time, from 0 dB/octave at the start of the speech portion, to the lower bound, back to 0 dB/octave and, finally back to the lower bound by the end of the utterance.

Three further time-varying patterns were also constructed:

- Step_tilt_1: spectral tilt at a constant value (the lower bound) is applied only in the 2nd half of the speech portion.
- Step_tilt_2: as above, but the spectral tilt is applied only in the 2nd and 3rd quarters of the speech portion.
- Step_tilt_3: as above, but the spectral tilt is applied only in the 2nd, 3rd and 6th sixths of the speech portion.

which gives us a total of six patterns, each of which can be applied at a lower bound spectral tilt of either −6 dB/octave or −9 dB/octave.

#### 4.6.2. *Results*

Figs. 14 and 15 summarize the performance of the proposed LNCC features under the six patterns of time-varying channel conditions, at the two distinct values of the lower bound spectral tilt. As with Experiment 1, the exact EERs are provided in Table 1.

LNCC provides relative reductions in EER, over standard MFCC, of between 1.29% ($p < .001$) and 49.61% ($p < .001$) for the abruptly-changing channels (Step_tilt_1 to Step_tilt_3, at −6 dB/octave or −9 dB/octave). For the gradually time-varying channels, LNCC provides relative reductions in EER as high as 50.2% ($p < .001$) (for the Slow_tilt_1 channel at −9 dB/octave). We can conclude that, under the vast majority of time-varying channel conditions, LNCCs offer better performance than MFCCs.

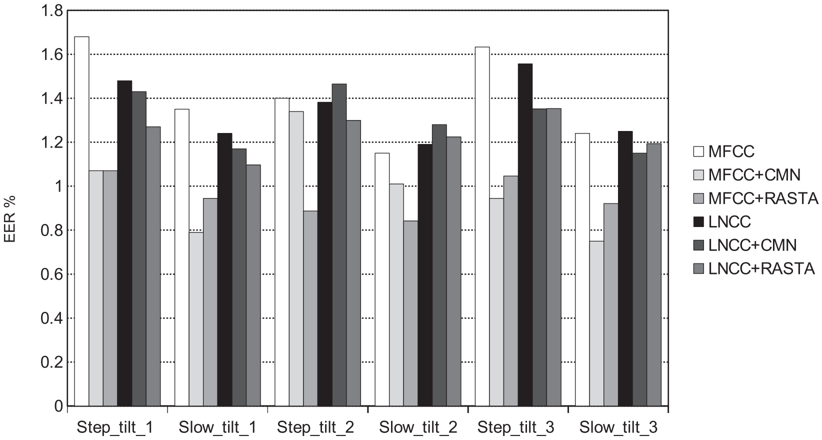

Fig. 14. Performance for varying tilt with lower bound −6 dB/octave. LNCC features are computed using 28 channels, $d_{\min} = 0.001$ and $B = 3.5$ Barks.

When the spectral tilt has a lower bound of -6dB/octave, CMN is sometimes highly effective in combination with MFCC, providing the best performance under four of the six patterns of changing tilt (Fig. 14). However, it is inconsistent, and the performance of CMN is less impressive under the other two conditions (Step_tilt_2 and Slow_tilt_2). Despite this fact, MFCC + CMN and MFCC + RASTA outperform LNCC, LNCC + CMN and LNCC + RASTA. Observe that CMN/RASTA is not as helpful for LNCC as for MFCC here for reasons that we believe are related to an inadequate model for the variability due to the effects of unknown linear filtering. This problem will be the subject of future research.

LNCC provides better performance in the context of more severe spectral tilt, as can be seen in Fig. 15. Here, neither CMN or RASTA perform as well, and the LNCC features – even without the benefit of CMN or RASTA – often outperform all of MFCC, MFCC + CMN and MFCC + RASTA. The combination of LNCC with either CMN or RASTA usually provides further reductions in EER.

### 4.7. Summary of results

We now collate the results of Experiments 1 and 2, and consider the overall performance of LNCC compared to MFCC, with and without CMN or RASTA processing. Fig. 16 presents these collated results, which cover a wide variety of channel conditions including clean, constant spectral tilt, gradually-varying spectral tilt and abruptly-varying spectral tilt. It is clear that the proposed LNCC features are superior to MFCCs both in absolute terms (mean EER) and in their consistency across different channels (standard deviation of EER). LNCC also outperforms the MFCC baseline

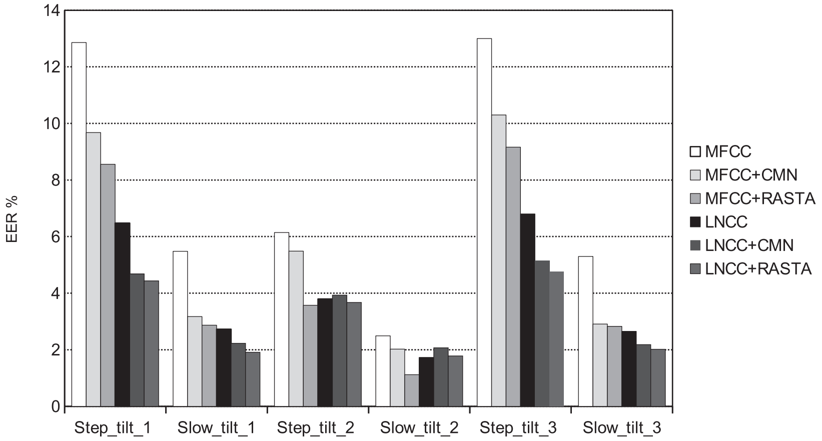

Fig. 15. Performance for varying tilt with lower bound −9 dB/octave. LNCC features are computed using 28 channels, $d_{\min} = 0.001$ and $B = 3.5$ Barks.

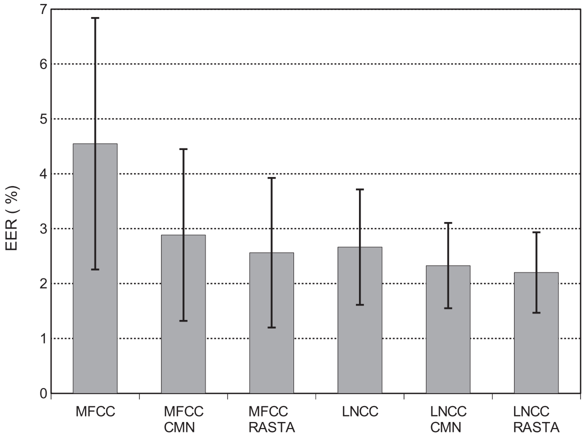

Fig. 16. Overall performance for all tested conditions. The average EER is in bars and the standard deviation in whiskers. LNCC features are computed using 28 channels, $d_{\min} = 0.001$ and $B = 3.5$ Barks.

for the majority of channels with constant or varying spectral tilts (Table 1). Besides, LNCC provides a lower average EER than MFCC + CMN. However, MFCC + RASTA outperforms LNCC, but the difference is small, as shown in Fig. 16.

LNCC alone offers the advantages of simplicity and not needing to access any information outside the current frame in order to carry out its local normalization. These could be advantageous in certain applications.

The decision on whether CMN provides significant improvements to MFCC depends on channel conditions, whereas the proposed LNCC features provide consistently good performance across all conditions and never suffer from the extremely high error rates which we observe in some cases for MFCCs. This can be seen in the very large error bars (representing the standard deviation) for MFCC in Fig. 16. The performance of LNCC is far more consistent: the error bars for LNCC are smaller than those for MFCC + CMN or MFCC + RASTA. This effect is even more pronounced when LNCC are combined with CMN or RASTA: the error bars for LNCC + CMN and LNCC + RASTA are the smallest of all.

Overall, CMN and RASTA improve the performance of LNCC in most cases, achieving relative reductions in EER of up to 30% in some conditions. However, these improvements are not as great as for MFCC features. This is likely to be because the model for convolutional distortion with LNCC is more complex than the low-frequency additive constant in MFCC, which both CMN and RASTA aim to compensate for.

As already mentioned, neither i-Vector or JFA systems perform well in this scenario (Table 1). Our experiments, which use short speech utterances (< 5s in YOHO database), confirm a known weakness of JFA and i-Vector systems, and this poor performance is also consistent with the results of other recent studies in state-of-the-art speaker verification that have been already cited. Therefore, we can also conclude that LNCC enables more robust speaker verification than i-Vector or JFA approaches, under the conditions used in this paper. LNCC is also, of course, dramatically simpler and easier to implement than JFA or i-Vectors.

## 5. Conclusions

In this paper, a perceptually-motivated and extremely simple but effective way to normalize speech features *instantaneously* is proposed. The effectiveness of the proposed features has been demonstrated for a speaker verification task across a wide variety of linear channel conditions.

With a constant −9 dB/octave spectral tilt, the proposed Locally-Normalized Cepstral Coefficients give a dramatic reduction in EER as high as 51.0% when compared with ordinary MFCCs, and the reductions were as high as 47.7%, 34.0% or 25.8% when compared with MFCC, MFCC + CMN or MFCC + RASTA under one of the variable spectral tilt conditions (STEP_tilt_3, −9 dB/octave).

While CMN and RASTA do further improve the performance of LNCC in most cases, achieving relative reductions of 30% in some conditions, they do not provide improvements that are quite as dramatic as when they are used in conjunction with MFCC coefficients.

We conclude that the proposed LNCC features are an attractive alternative to MFCC or MFCC + CMN in any situation where it is difficult to estimate the cepstral means accurately. Other application scenarios might include those where very low latency or low complexity is desired, in which computing and storing the moving average required by CMN may be inconvenient. Because all processing in LNCC is performed independently within each frame and no information needs to be exchanged between frames, it is also amenable to simple parallel implementations.

In future work we plan to evaluate the proposed features for an automatic speech recognition (ASR) task, although it is possible that the self-normalizing filterbank may remove a small amount of phonetic information along with the channel information, so some modifications may be necessary to limit the amount of normalization that is performed. Another obvious line of investigation would be to combine LNCCs with MFCCs or PLPs using either feature combination or system combination. The effect of unknown linear filtering associated with LNCC is more complex than the low frequency additive constant that is addressed by CMN and RASTA for MFCC, and needs to be modeled; alternatively, multi-frame normalization schemes may also need to be developed for LNCC. Finally, we plan to compare the proposed features with the baselines using even shorter utterances for speaker verification.

## Acknowledgements

The research leading to these results was funded by CONICYT-ANILLO project ACT 1120 and CONICYT-FONDECYT project 1100195. S. King was partly funded by EPSRC grant EP/I031022/1 (Natural Speech Technology). Richard Stern was partially funded by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch.

## References

Ali, A.M., Van Der Spiegel, J., Mueller, P., 2000. Auditory-based speech processing based on the average localized synchrony detection. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Instanbul, pp. 1623–1626.

Ali, A.M., Van Der Spiegel, J., Mueller, P., 2002. Robust auditory-based speech processing using the average localized synchrony detection. IEEE Trans. Speech Audio Process. 10, 279–292.

Anderson, S., Skoe, E., Chandrasekaran, B., Kraus, N., 2010. Neural timing is linked to speech perception in noise. J. Neurosci. 30, 4922–4926.

Atal, B., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Am. 55, 1304–1312.

Bell, P., Yumamoto, H., Swietojanski, P., Wu, Y., McInnes, F., Hori, C., Renals, S., 2013. A lecture transcription system combining neural network acoustic and language models. In: Proceedings of Interspeech 2013, Lyon, pp. 3087–3091.

Bies, D.A., Hansen, C.H., 2009. Engineering Noise Control: Theory and Practice, fourth Ed. Taylor & Francis Group, New York.

Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I.M., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., Reynolds, D.A., 2004. A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process. 52, 430–451.

Bonastre, J.F., Scheffer, N., Fredouille, C., Matrouf, D., 2004. Nist'04 speaker recognition evaluation campaign: New lia speaker detection platform based on alize toolkit. In: NIST SRE 2004 Workshop: Speaker Detection Evaluation Campaign, Toledo, pp. 1–10.

Bonastre, J.F., Scheffer, N., Matrouf, D., Fredouille, C., Larcher, A., Preti, A., Pouchoulin, G., Evans, N., Fauve, B., Mason, J., 2008. Alize/spkdet: a state-of-the-art open source software for speaker recognition. In: Proceedings IEEE Odyssey, ISCA Speaker Recognition Workshop, Stellenbosch, pp. 1–8.

Bonastre, J.F., Wils, F., Meignier, S., 2005. Alize, a free toolkit for speaker recognition. In: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, pp. 737–740.

Bořil, H., Hansen, J.H.L., 2010. Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments. IEEE Trans. Audio Speech Language Process. 18, 1379–1393.

Brandstein, M., Ward, D., 2010. Microphone Arrays: Signal Processing Techniques and Applications. In: Digital Signal Processing. Springer.

Buchner, H., Benesty, J., Kellermann, W., 2005. Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication. Signal Process. 85, 549–570.

Campbell, J.P., 1997. Speaker recognition: a tutorial. Proc. IEEE 85, 1437–1462.

Campbell, J.P., Higgings, A., 1994. YOHO Speaker Verification. Linguistic Data Consortium, Philadelphia, PA.

Chen, J., Paliwal, K.K., Nakamura, S., 2003. Cepstrum derived from differentiated power spectrum for robust speech recognition. Speech Commun. 41, 469–484.

Chen, J., Wu, X., Li, L., Chi, H., 2009. Simulated phase-locking stimulation: an improved speech processing strategy for cochlear implants. ORL - J. Oto-Rhino-Laryngol. Relat. Specialit. 71, 221–227.

Chigier, B., Leung, H.C., 1992. The effects of signal representations, phonetic classification techniques, and the telephone network. In: Proceedings of the Second International Conference on Spoken Language Processing, Banff, Alberta, pp. 97–100.

Chiu, Y.-H., Stern, R.M., 2008. Analysis of physiologically-motivated signal processing for robust speech recognition. In: Proceedings of Interspeech 2008, Brisbane, pp. 1000–1003.

Cooke, M., King, S., Garnier, M., Aubanel, V., 2014. The listener talker: a review of human and algorithmic context-induced modifications of speech. Comp. Speech Language 28, 543–571.

Cooke, M., Lecumberri, M.L., 2012. The intelligibility of lombard speech for non-native listeners. J. Acoust. Soc. Am. 132, 1120–1129.

Cooke, M., Mayo, C., Valentini-Botinhao, C., 2013a. Intelligibility-enhancing speech modifications: the hurricane challenge. In: Proceedings of Interspeech 2013, Lyon, pp. 3552–3556.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Sauert, B., Tang, Y., 2013b. Evaluating the intelligibility benefit of speech modifications in known noise conditions. Speech Commun. 55, 572–585.

Crocker, M.J., 2006. In: Crocker, M.J. (Ed.), Handbook of Noise and Vibration Control. John Wiley & Sons, Inc., Hoboken, New Jersey.

Darwin, C.J., McKeown, J.D., Kirby, D., 1989. Perceptual compensation for transmission channel and speaker effects on vowel quality. Speech Commun. 8, 221–223.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28, 357–366.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouelle, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Language Process. 19, 788–798.

Delgutte, B., Kiang, N.Y.S., 1984. Speech coding in the auditory nerve: I. vowels-like sounds. J. Acoust. Soc. Am. 75, 866–876.

Dimitriadis, D., Maragos, P., Potamianos, A., 2011. On the effects of filterbank design and energy computation on robust speech recognition. IEEE Trans. Audio Speech Language Process. 19, 1504–1516.

Dreyer, A., Delgutte, B., 2006. Phase locking of auditory-nerve fibers to the envelopes of high frequency sounds: Implications for sound localization. J. Neurophysiol. 96, 2327–2341.

Eggermont, J.J., 1998. Is there a neural code? Neurosci. Biobehav. Rev. 22, 355–370.

Fahy, F.J., 1987. Sound and Structural Vibration Radiation, Transmission and Response. Academic Press, London.

Fahy, F.J., Gardonio, P., 2007. Sound and Structural Vibration. Academic Press, London.

Fahy, F.J., Walker, J.G., 1998. Fundamentals of Noise and Vibration. Taylor and Francis, London.

Fauve, B.G.B., Matrouf, D., Scheffer, N., Bonastre, J.F., Mason, J.S.D., 2007. State-of-the-art performance in text-independent speaker verification through open-source software. IEEE Trans. Audio Speech Language Process. 15, 1960–1968.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. 29, 254–272.

Gales, M.J.F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Comp. Speech Language 12, 75–98.

Gaubitch, N.D., Brookes, M., Naylor, P.A., 2013. Blind channel magnitude response estimation in speech using spectrum classification. IEEE Trans. Audio Speech Language Process. 21, 2162–2171.

Ghitza, O., 1994. Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Trans. Speech Audio Process. 2, 115–132.

Glass, J., Hazen, T., Cyphers, S., Malioutov, I., Huynh, D., Barzilay, R., 2007. Recent progress in the MIT spoken lecture processing project. In: Proceedings of Interspeech 2007, Antwerp, pp. 2553–2556.

Hain, T., Burget, L., Dines, J., Garau, G., Karafit, M., Lincoln, M., Wan, V., 2006. The AMI meeting transcription system. In: Proceedings of the NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation Workshop.

Hain, T., Burget, L., Dines, J., Garau, G., Wan, V., Karafiat, M., Vepa, J., Lincoln, M., 2007. The AMI system for the transcription of speech in meetings. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007, Honolulu, pp. 357–360.

Hain, T., Burget, L., Dines, J., Garner, P., Grezl, F., Hannani, A., Huijbregts, M., Karafiat, M., Lincoln, M., Wan, V., 2012. Transcribing meetings with the AMIDA system. IEEE Trans. Audio Speech Language Process. 20, 486–498.

Hanilçi, C., Ertaş, F., 2013. Investigation of the effect of data duration and speaker gender on text-independent speaker recognition. Comp. Elect. Eng. 39, 441–452.

Hansen, C., 2005. Noise Control: From Concept to Application. Taylor and Francis Group, New York.

Hansen, J.H.L., Varadarajan, V., 2009. Analysis and compensation of Lombard speech acroos noise type and levels with application to In-Set/Out-of-Set speaker recognition. IEEE Trans. Audio Speech Language Process. 17, 366–378.

Hasan, T., Hansen, J.H.L., 2011. Robust speaker recognition in non-stationary room environments based on empirical mode decomposition. In: Proceedings of Interspeech 2011, Florence, pp. 2722–2736.

Hasan, T., Hansen, J.H.L., 2013. Acoustic factor analysis for robust speaker verification. IEEE Trans. Audio Speech Language Process. 21, 842–853.

Hautamäki, V., Cheng, Y., Rajan, P., Lee, C., 2013. Minimax i-vector extractor for short duration speaker verification. In: Proceedings of Interspeech 2013, Lyon, pp. 3708–3712.

Heinz, M.G., Swaminathan, J., 2009. Quantifying envelope and fine-structure coding in auditory-nerve responses to chimaeric speech. J. Assoc. Res. Otolaryngol. 10, 407–423.

Hermansky, H., 1990. Perceptual linear predictive PLP analysis of speech. J. Acoust. Soc. Am. 87, 1738–1752.

Hermansky, H., Cohen, J.R., Stern, R.M., 2013. Perceptual properties of current speech recognition technology. Proc. IEEE 101, 1968–1985.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2, 578–589.

Hermansky, H., Morgan, N., Bayya, A., Khon, P., 1991a. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In: Proceedings of Eurospeech, Genova, pp. 1367–1370.

Hermansky, H., Morgan, N., Bayya, A., Khon, P., 1991b. (RASTA-PLP) speech analysis technique. In: Proceedings International Conference on Acoustics, Speech, and Signal Processing, San Francisco, pp. 121–124.

Hori, T., Fujimoto, M., Ogawa, A., Kinoshita, K., Nakamura, A., 2012. Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera. IEEE Trans. Audio Speech Language Process. 20, 499–513.

Houtgast, T., 1972. Psychophysical evidence for lateral inhibition in hearing. J. Acoust. Soc. Am. 51, 1885–1894.

Hsu, B.J., Glass, J., 2006. Style and topic language model adaptation using HMM-LDA. In: Proceedings of the Conference on Empirical Methods in Natural Processing, Sydney, pp. 373–381.

Hsu, C.W., Lee, L.S., 2009. Higher order cepstral moment normalization for improved robust speech recognition. IEEE Trans. Audio Speech Language Process. 17, 205–220.

Ishi, C.T., Matsuda, S., Kanda, T., Jitsuhiro, T.H.I.S.N., Hagita, N., 2008. A robust speech recognition system for communication robots in noisy environments. IEEE Trans. Robot. 24, 759–763.

ISO10140-2:2010. 2010. Acoustics - Laboratory measurement of sound insulation of building elements – Part 2: Measurement of airborne sound insulation.

ISO140-III. 1995. Acoustics - Measurement of sound insulation in buildings and of building elements – Part 3: Laboratory measurement of airborne sound insulation of building elements.

Jankowski, C.R., Lippmann, R.P., 1992. Comparison of auditory models for robust speech recognition. In: Proceedings of the DARPA Speech and Natural Language Workshop, New York, pp. 453–454.

Jankowski, C.R., Vo, H.D., Lippmann, R.P., 1995. A comparison of signal processing front ends for automatic word recognition. IEEE Trans. Speech Audio Process. 3, 286–293.

Jensen, B., Tomatis, N., Drygajlo, A., Siegwart, R., 2005. Robots meet human interaction in public spaces. IEEE Trans. Indus. Electron. 52, 1530–1546.

Johnson, D., 1980. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. J. Acoust. Soc. Am. 68, 1115–1122.

Joris, P., Yin, T., 2007. A matter of time: Internal delays in binaural processing. Trends Neurosci. 30, 70–78.

Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., Mason, M., 2011. I-vector based speaker recognition on short utterances. In: In Proceedings of Interspeech 2007, Florence, pp. 2341–2344.

Kayser, C., Montemurro, M.A., Logothetis, N.K., Panzeri, S., 2009. Spike-phase coding boost and stabilizes information carried by spatial and temporal spike patterns. Neuron 61, 597–608.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007a. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio Speech Language Process. 15, 1435–1447.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007b. Speaker and session variability in gmm-based speaker verification. IEEE Trans. Audio Speech Language Process. 15, 1448–1460.

Kenny, P., Dumouchel, P., 2004. Disentangling speaker and channel effects in speaker verification. In: Proceedings International Conference on Acoustics, Speech, and Signal Processing, Montreal, pp. 37–40.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., 2008. A study of interspeaker variability in speaker verification. IEEE Trans. Audio Speech Language Process. 16, 980–988.

Kenny, P., Stafylakis, T., Ouellet, P., Alam, J., Dumouchel, P., 2013. PLDA for speaker verification with utterances of arbitrary duration. In: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2013), Vancouver, pp. 7649–7653.

Kiang, N.Y.S., Watanabe, T., Thomas, E.C., Clark, L.F., 1965. Discharge Patterns of Single Fibers in the Cat's Auditory Nerve. MIT Press, Cambridge. MA.

Kim, C., Chiu, Y.B., Stern, R.M., 2006. Physiologically-motivated synchrony-based processing for robust automatic speech recognition. In: Proceedings of Interspeech 2006, Pittsburgh, pp. 1483–1486.

Kim, D.S., Lee, S.Y., Kil, R.M., 1999. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Trans. Speech Audio Process. 7, 55–69.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Commun. 52, 12–40.

Kleinschmidt, T., Sridharan, S., Mason, M., 2011. The use of phase in complex spectrum subtraction for robust speech recognition. Comp. Speech Language 25, 585–600.

Kriegstein, K., Smith, D.R., Patterson, R.D., Kiebel, S.J., Griffiths, T.D., 2010. How the human brain recognizes speech in the context of changing speakers. J. Neurosci. 30, 629–638.

Kumaresan, R., Rao, A., 1999. Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications. J. Acoust. Soc. Am. 105, 1912–1924.

Kuwabara, H., Sagisaka, Y., 1995. Acoustics characteristics of speaker individuality: control and conversion. Speech Commun. 16, 165–173.

Larcher, A., Bonastre, J.F., Fauve, B., Lee, K.A., Levy, C., Li, H., Mason, J.S., Parfait, J.Y., 2013. Alize 3. 0" open source toolkit for state-of-the-art speaker recognition. In: Proceedings of Interspeech 2013, Lyon, pp. 2768–2773.

Larcher, A., Lee, K.A., Ma, B., Li, H., 2014. Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Commun. 60, 56–77.

Leeuwis, E., Federico, M., Cettolo, M., 2003. Language modeling and transcription of the TED corpus lecture. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006, Toulouse, pp. 232–235.

Leus, G., 2004. On the estimation of rapidly time-varying channels. In: Proceedings of European Signal Processing Conference, Vienna, pp. 2227–2230.

Leus, G., Moonen, M., 2003. Deterministic subspace based blind channel estimation for doubly-selective channels. In: Proceedings of the 4th IEEE Workshop on Signal Processing Advances in Wireless Communications, pp. 210–214.

Liberman, M.C., 1978. Auditory nerve response from cats raised in a low noise chamber. J. Acoust. Soc. Am. 63, 442–455.

Liu, F., Stern, R.M., Huang, X., Acero, A., 1993. Efficient cepstral normalization for robust speech recognition. In: Proceedings DARPA Speech and Natural Language Workshop, Cambridge, pp. 69–74.

Long, M., 2006. Architectural Acoustics: Applications of Modern Acoustics. Elsevier Academic Press, UK, London.

Lu, X., Unoki, M., Nakamura, S., 2011. Sub-band temporal modulation envelopes and their normalization for automatic speech recognition in reverberant environments. Comp. Speech Language 25, 571–584.

Malionek, J., Oard, D.W., Sangwan, A., Hansen, J.H.L., 2013. Linking transcribed conversational speech. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, pp. 961–964.

Mandasari, M.I., Saeidi, R., McLaren, M., van Leeuwen, D., 2013. Quality measure functions for calibration of speaker recognition systems in various duration conditions. IEEE Trans. Audio Speech Language Process. 21, 2425–2438.

Meddis, R., Hewitt, M.J., 1991a. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. i: Pitch identification. J. Acoust. Soc. Am. 89, 2866–2882.

Meddis, R., Hewitt, M.J., 1991b. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. ii: Phase sensitivity. J. Acoust. Soc. Am. 89, 2883–2894.

Meyer, B., Kollmeier, B., 2011. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. Speech Commun. 53, 753–767.

Miettinen, I., Alku, P., Salminen, N., May, P.J.C., Tiitinen, H., 2011. Responsiveness of the human auditory cortex to degraded speech sounds: Reduction of amplitude resolution vs. additive noise. Brain Res. 1367, 298–309.

Miettinen, I., Alku, P., Yrttiaho, S., May, P.J.C., Tiitinena, H., 2012. Cortical processing of degraded speech sounds: Effects of distortion type and continuity. NeuroImage 60, 1036–1045.

Miller, R.L., Calhoun, B.M., Young, E.D., 1999. Contrast enhancement improves the representation of //-like vowels in the hearing-impaired auditory nerve. J. Acoust. Soc. Am. 106, 2693–2708.

Moore, B.C.J., 2003. An Introduction to the Psychology of Hearing. Elsevier Science, Academic Press.

Moore, B.C.J., 2008. The rol of temporal fine structure processing in pitch perception, masking, and speech perception for normal hearing and hearing-impaired people. J. Assoc. Res. Otolaryngol. 9, 399–406.

Moore, B.C.J., 2014. Auditory Processing of Temporal Fine Structure: Effects of Age and Hearing Loss, Audiology and Otology. World Scientific Publishing CO PTE LTD.

Morales, N., Toledano, D., Hansen, J.H.L., Garrido, J., 2009. Feature compensation techniques for ASR on band-limited speech. IEEE Trans. Audio Speech Language Process. 17, 758–774.

Nakano, A.Y., Nakagawa, S., Yamamoto, K., 2010. Distant speech recognition using a microphone array network. IEICE Trans. Inform. Syst. E93.D, 2451–2462.

Nassau, K., 1983. The Physics and Chemistry of Color: The Fifteen Causes of Color. Wiley, New York.

Norton, M., Karczub, D., 2003. Fundamentals of Noise and Vibration Analysis for Engineers. Cambridge University Press, Cambridge, UK.

Ohshima, Y., Stern, R.M., 1994. Environmental robustness in automatic speech recognition using physiologically-motivated signal processing. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994, Adelaide, pp. 1–4.

Paliwal, K., Wojcicki, K., Shannon, B., 2011. The importance of phase in speech enhancement. Speech Commun. 53, 2883–2894.

Parikh, G., Loizou, P.C., 2005. The influence of noise of vowel and consonant cues. J. Acoust. Soc. Am. 118, 3874–3888.

Park, A., Hazen, T., Glass, J., 2005. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, Philadelphia, pp. 497–500.

Park, H.K., Bradley, J.S., 2009. Evaluating signal-to-noise ratios, loudness, and related measures as indicators of airborne sound insulation. J. Acoust. Soc. Am. 126, 208–219.

Park, H.K., Bradley, J.S., Gover, B.N., 2008. Evaluating airborne sound insulation in terms of speech intelligibility. J. Acoust. Soc. Am. 123, 1458–1471.

Pickles, J.O., 2008. An Introduction to the Physiology of Hearing. Emerald Group Publishing Limited.

Qi Li, P., Zheng, J., Tsai, A., Zhou, Q., 2002. Robust end-point detection and energy normalization for real-time speech and speaker recognition. IEEE Trans. Speech Audio Process. 10, 146–157.

Qin, L., Wang, J.Y., Sato, Y., 2008. Representations of cat meows and human vowels in the primary auditory cortex of awake cats. J. Neurophysiol. 99, 2305–2319.

Renals, S., Hain, T., Bourlard, H., 2007. Recognition and understanding of meetings: The AMI and AMIDA projects. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2007, Kyoto, pp. 238–247.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian Mixture Models. Dig. Signal Process. 10, 19–41.

Reynolds, D.A., Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3, 72–83.

Rose, R.C., Reynolds, D.A., 1990. Text-independent speaker identification using automatic acoustic segmentation. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990, Albuquerque, pp. 293–296.

Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. B 336, 367–373.

Sachs, M.B., 1984. Neural coding of complex sounds: Speech. Annu. Rev. Physiol. 46, 261–273.

Sachs, M.B., Kiang, N.Y.-S., 1968. Two-tone inhibition in auditory-nerve fibers. J. Acoust. Soc. Am. 43, 1120–1128.

Sachs, M.B., Young, E.D., 1979. Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. J. Acoust. Soc. Am. 66, 470–479.

Sangwan, A., Kaushik, L., Yu, C., Hansen, J.H.L., Oard, D.W., 2013. Houston, we have a solution: using NASA Apollo Program to advance speech and language procesing technology. In: Proceedings of Interspeech 2013, Lyon, pp. 1135–1139.

Sato, H., Bradley, J.S., 2008. Evaluation of acoustical conditions for speech communication in working elementary school classrooms. J. Acoust. Soc. Am. 123, 2064–2077.

Schwartz, R., Anastasakos, T., Kubala, F., Makhoul, J., Nguyen, L., Zavaliagkos, G., 1993. Comparative experiments on large vocabulary speech recognition. In: Proceedings of the Workshop on Human Language Technology, Princeton, pp. 75–80.

Seltzer, M.L., Stern, R.M., 2004. Likelihood-maximizing beamforming for robust hands-free speech recognition. IEEE Trans. Speech Audio Process. 12, 489–498.

Seneff, S., 1984. Pitch and spectral estimation of speech based on an auditory synchrony model. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1984, San Diego, pp. 1–4.

Seneff, S. 1985. Pitch and spectral analysis of speech based on an auditory synchrony model, PhD. Dissertation. Massachusetts Institute of Technology, Cambridge.

Seneff, S., 1986a. Characterizing formants through straight line approximations without explicit formant tracking. In: Proceedings of the First Montreal Symposium on Speech Recognition, Montreal, pp. 21–27.

Seneff, S., 1986b. A computational model for the peripheral auditory system: application to speech recognition research. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1986, Tokyo, pp. 1983–1986.

Seneff, S., 1987. Vowel recognition based on line-formants derived from an auditory-based spectral. In: Proceedings of the 11th International Congress of Phonetic Sciences, Tallin.

Seneff, S., 1988. A joint synchrony/mean-rate model of auditory speech processing. J. Phonet. 16, 55–76.

Shao, Y., Srinivasan, S., Jin, Z., Wang, D., 2010. A computational auditory scene analysis system for speech segregation and robust speech recognition. Comp. Speech Language 24, 77–93.

Shi, G., Shanechi, M.M., Aarabi, P., 2006. On the importance of phase in human speech recognition. IEEE Trans. Audio Speech Language Process. 14, 1867–1874.

Sinex, D.G., Geisler, D., 1983. Responses of primary auditory fibers to consonant-vowel syllables. J. Acoust. Soc. Am. 73, 602–615.

Smith, Z.M., Delgutte, B., Oxenham, A.J., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature 416, 87–90.

Soong, F.K., Rosenberg, A.E., 1988. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. Acoustics Speech Signal Process 36, 871–879.

Stern, R.M., Morgan, N., 2012a. Features based on auditory physiology and perception. In: Virtanen, T., Raj, B., Singh, R. (Eds.), Techniques for Noise Robustness in Automatic Speech Recognition. Wiley, New York, NY, USA:, pp. 207–243.

Stern, R.M., Morgan, N., 2012b. Hearing is believing: biologically inspired methods for robust automatic speech recognition. Signal Process. Magaz. IEEE, 34–43.

Stern, R.M., Wang, D., Brown, G.J., 2006. Binaural sound localization. In: Wang, D., Brown, G.J. (Eds.), Computational Auditory Scene Analysis chapter 5. Wiley-IEEE Press.

Stilp, C.E., Alexander, J.M., Kiefte, M., Kluend, K.R., 2010. Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. Attent. Percept. Psychophys. 72, 470–480.

Stockham, T.G., Cannon, T.N., Ingebretsen, R.B., 1975. Blind deconvolution through digital signal processing. Proc. IEEE 63, 678–693.

Tadeu, A., António, J.M.P., 2002. Acoustic insulation of single panel walls provided by analytical expressions versus the mass law. J Sound Vib 257, 457–475.

Tchorz, J., Kleinschmidt, M., Kollmeier, B., 1996. A psychoacoustical model of auditory periphery as the front end for ASR. J. Acoust. Soc. Am. 105, 1157–1157.

Tchorz, J., Kollmeier, B., 1999. A model of auditory perception as front end for automatic speech recognition. J. Acoust. Soc. Am. 106, 2040–2050.

Togneri, R., Pullella, D., 2011. An overview of speaker identification: Accuracy and robustness issues. IEEE Circ. Syst. Magaz. 11, 23–61.

Tokuda, K., Yoshimura, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2000, Instanbul, pp. 1315–1318.

Trancoso, I., Nunes, R., Neves, L., 2006. Classroom lecture recognition. In: Computational Processing of the Portuguese Language, Proceedings Book Series: Lecture Notes in Artificial Intelligence, pp. 190–199.

Tranter, S., Reynolds, D.A., 2006. An overview of automatic speaker diarization systems. IEEE Trans. Audio Speech Language Process 14, 1557–1565.

Vér, I.L., Beranek, L.L., 2006. Noise and Vibration Control Engineering: Principles and Applications. Second Edition. John Wiley and Sons, Inc, Hoboken, New Jersey.

Vogt, R., Pelecanos, J.W., Scheffer, N., Kajarekar, S., Sridharan, S., 2009. Within-session variability modelling for factor analysis speaker verification. In: Proceedings of Interspeech 2009, Brighton, pp. 1563–1566.

Wang, K., Shamma, S.A., 1994. Self-normalization and noise-robustness in early auditory representations. IEEE Trans. Speech Audio Process. 2, 421–435.

Wang, L., Kitaoka, N., Nakagawa, S., 2007a. Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM. Speech Commun. 49, 501–513.

Wang, L., Kitaoka, N., Nakagawa, S., 2007b. Robust distant speech recognition by combining position-dependent CMN with conventional CMN. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007, Honolulu, pp. 817–820.

Wang, L., Kitaoka, N., Nakagawa, S., 2011. Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm. IEICE Trans. Inform. Syst. E.94.D, 659–667.

Watkins, A.J., Makin, S.J., 1996. Some effects of filtered contexts on the perception of vowels and fricatives. J. Acoust. Soc. Am. 99, 588–594.

Werblin, F.S., Jacobs, A., Teeters, J., 1996. The computational eye. IEEE Spectrum 33, 30–37.

Wölfel, M., 2009a. Enhanced speech features by single-channel joint compensation of noise and reverberation. IEEE Trans. Audio Speech Language Process. 17, 312–323.

Wölfel, M., 2009b. Signal adaptive spectral envelope estimation for robust speech recognition. Speech Commun. 51, 551–561.

Wölfel, M., McDonough, J., 2009. Distant Speech Recognition. Wiley, Chichester. UK.

Yin, S., Rose, R., Kenny, P., 2007. A joint factor analysis approach to progressive model adaptation in text-independent speaker verification. IEEE Transactions on Audio, Speech. and Language Processing 15, 1999–2010.

Yokoyama, R., Nasu, Y., Iwano, K., Shinoda, K., 2013. Detection of overlapped speech using lapel microphones in meeting. Speech Commun. 55, 941–949.

Yoma, N.B., Villar, M., 2002. Speaker verification in noise using a stochastic version of the weighted viterbi algorithm. IEEE Trans. Speech Audio Process. 10, 158–166.

Young, E.D., 2008. Neural representation of speech spectral and temporal information in speech. Philos. Trans. R. Soc. B 363, 923–945.

Young, E.D., Sachs, M.B., 1979. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. J. Acoust. Soc. Am. 66, 1381–1403.

Zilovic, M.S., Ramachandran, R.P., Mammone, R.J., 1998. Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer function. IEEE Trans. Speech Audio Process. 6, 260–267.

Zwicker, E., 1961. Subdivision of the audible frequency range into critical bands (frequenzgrupenn). J. Acoust. Soc. Am. 33, 248–249.