Feature Extraction Using Power-Law Adjusted Linear Prediction With Application to Speaker Recognition Under Severe Vocal Effort Mismatch

Rahim Saeidi, Senior Member, IEEE, Paavo Alku Senior Member, IEEE, Tom Bäckström, Senior Member, IEEE

Abstract—Linear prediction is one of the most established techniques in signal estimation, and it is widely utilized in speech signal processing. It has been long understood that the nerve firing rate of human auditory system can be approximated by power law non-linearity, and this has been the motivation behind using perceptual linear prediction in extracting acoustic features in a variety of speech processing applications. In this paper, we revisit the application of power law non-linearity in speech spectrum estimation by compressing/expanding power spectrum in autocorrelation-based linear prediction. The development of so-called LP-α is motivated by a desire to obtain spectral features that present less mismatch than conventionally used spectrum estimation methods when speech of normal loudness is compared to speech under vocal effort. The effectiveness of the proposed approach is demonstrated in a speaker recognition task conducted under severe vocal effort mismatch comparing shouted versus normal speech mode.

I. INTRODUCTION

Linear prediction (LP) is a fundamental tool that can be used in many areas of science and technology to model, compress and recognize digital time-domain signals. In speech processing, LP has been used practically in all main technological areas such as speech analysis [1], [2], coding [3], synthesis [4], speech recognition [5], [6], speaker recognition [7]–[10], and enhancement [11]. The basic form of LP defines an optimal predictive filter, an FIR with a given prediction order p, by minimizing the mean-square error (MSE) of the prediction error, the residual [12]. By assuming that the signal waveform is zero outside a given frame, it is well-known [13] that this 2-norm optimization results in a set of Wiener-Hopf equations that are determined from the autocorrelation function of the speech signal. In the frequency domain, minimization of the squared residual is equivalent to the minimization of the integrated ratio of two power spectra, the signal and its LP model [14]. As a parametric spectral estimation technique, LP belongs to the family of all-pole modeling methods because it optimizes a digital filter that has only poles (outside infinity) in the z-domain.

A. Modifications of LP

The basic form of LP has been a subject to many modifications in speech processing, such as perceptual LP (PLP) [5], weighted LP (WLP) [1], regularized LP (RLP) [10], [15], [16] and frequency domain LP (FDLP) [17], [18]. Given the fact that LP is defined from the autocorrelation function, an apparent way to modify it is to affect the computation of the autocorrelation. Since autocorrelation and power spectrum form a Fourier transform pair, an obvious means to affect the autocorrelation is to first modify the power spectrum and then to compute the inverse Fourier-transform of the processed spectrum. This kind of approach was used in [19], where the square root operation was used in the power spectrum domain to obtain better all-pole models for the speech spectra of a large dynamic range and spectral dips. This technique, called square root autocorrelation (SQRTA), is shown to yield a prediction error signal that models a sequence of impulses more accurately than the residual computed by LP. A similar approach, called spectral transform LP [20] (STLP), was studied in the analysis of high-pitch speech by using a general r-th root transform instead of the square root operation.

The idea of transforming the power spectrum in all-pole modeling has also been used in PLP [5] in the form of a power law non-linearity in order to better simulate the human auditory system. In PLP, the power spectrum of a speech segment is multiplied by the equal loudness curve and the resulting spectrum is then raised to the power of 0.33 (cubic root compression) before computing an all-pole model. The cubed root compression is motivated by Stevens’ power-law [21] of hearing, which constitutes the relation between sound intensity and resulting loudness for human perception. Other power-law non-linearity was recently proposed in extracting power-normalized cepstral coefficients (PNCC) acoustic features [22]. In the PNCC computation chain, the power spectrum passed through the Gammatone filter bank is “bias-corrected” and then raised to the power of 0.1 (compression) before calculating the cepstral coefficients. The application of the power function non-linearity in PNCC results in output that is close to zero if the input is very small, which corresponds to the linear-threshold auditory nerve firing rate function in human auditory processing [23].

B. Speech under increased vocal effort

Many previous studies have used the basic form of LP and its modifications as a feature extraction method in speech recognition or speaker recognition. Several of these studies have succeeded in proposing an LP-based feature extraction method that is more robust than, for example, the widely used
Mel Frequency Cepstral Coefficient (MFCC) method when speech is corrupted by external distortion such as background noise [22], reverberation [6], [24], or a communication channel [25], [26]. It is worth noting that these investigations have almost exclusively used speech data produced under normal loudness. Real-life speech, however, is never entirely of normal loudness because humans have an amazing ability to vary the vocal effort in their voice production. Recognition systems built upon speech of normal loudness face severe difficulties when tested with speech that is produced in vocal effort levels different from the normal mode. Such situation occurs in forensic speaker recognition, which has been long considered as a challenging task in presenting support of audio evidence to the court. The main difficulty arises from the fact that the speech material recorded in a forensic case is often of very low quality, degraded and/or intentionally modified. Most importantly, the recording may have taken place in circumstances where the talker is excited and produces his or her speech by shouting, hence causing a vocal effort mismatch between the training and testing of the recognition system.

Vocal effort is a source of intrinsic variability and is defined as the quantity that talkers change to adapt their voice in response to, for example, increased or decreased communication distance. This adaptation can occur also while communicating in noisy conditions with the intention to increase intelligibility. This is referred to as speech under the Lombard effect. Shouting can occur due to extreme emotional state or in emergency situations. The literature regarding the effect of vocal effort on speech spectrum and prosody provides insight into several parameters conveying salient information like sound pressure level, duration lengthening/shortening, spectral center of gravity, frame energy distribution, and spectral tilt [27], [28].

Increased vocal effort involves increase in the muscle tension that affects the characteristics of the vocal tract resonances. The position of the first formant frequency tends to increase for speech under elevated vocal effort. This phenomenon has been consistently reported in the literature (more prominent for high-pitched speakers), and the amount of shift depends, to some extent, on the exact vowel [28].

When the vocal effort is increased from normal loudness to shouting, many acoustical properties of the voice change. In addition to the obvious effect of an increased sound pressure level (SPL) in shouting, also segmental durations and spectral features of speech differ between shouted and normal speech. A large difference in the fundamental frequency (F0) between shouted speech and speech of normal loudness for both male and female speakers is reported in [29], [30]. In a follow-up study, in comparison to French vowels which were produced with a normal effort, increased values for the frequency of the first formant (F1) were reported for shouted French vowels [31]. The amplified articulatory movement patterns in loud speech relative to normal speech can be explained perceptually by relating them to the importance of maintaining the Bark distance between F1 and F0: since F0 increases in loud speech and shouting, the frequency of F1 must also shift up in order to maintain the correct phonetic identities [32].

Increased vocal effort also corresponds to an increase in F0 due to the raise of sub-glottal pressure. However, this increase is mostly gender- and speaker-specific and low-pitched speakers tend to show higher relative increase in F0 in producing speech under higher vocal effort. The spectral tilt declines towards increased vocal effort except in whispering [27], [28]. Due to the absence of periodic glottal excitation in whispering, spectral tilt tends to be lower compared to other speech modes. Finally, by raising the vocal effort, temporal duration of vowels increases while consonants get shorter [33].

C. Contribution of this paper

In this paper, we revisit LP analysis in a transform domain, where autocorrelation function is derived from the transformed power spectrum. The aim is to develop acoustic features that present less mismatch than conventional features when capturing speaker-specific characteristics in speech of normal loudness and in shouting. More specifically, we look into a situation of forensic speaker recognition, where matching shouted and normal speech is of interest. For automatic speaker recognition systems, the conventional acoustic feature extraction using LP produces an unbearable mismatch between shouted and normal speech. We show that the performance of the state-of-the-art i-vector-based [34] speaker recognition using conventional linear prediction can be improved by reducing the sensitivity of LP analysis to the peaks of the power spectrum (compression) in shouted speech. In addition, by making the peaks of power spectrum more prominent in speech of normal loudness (expanding), the resulting features become more similar to their shouted counterparts.

The novelty of this paper is twofold: (1) in search of adjustable acoustic features, we revisit the power spectrum transformation in LP analysis and present a better match between speech of normal loudness and shouting; (2) we experiment with a state-of-the-art speaker recognition system, comparing system performance with different LP analysis methods in conditions when there is either (a) no mismatch between training and testing or (b) a severe vocal effort mismatch caused by having speech of normal loudness in the speaker modeling and shouting in test phase.

The implication of power spectrum compression/expansion in calculating autocorrelation function of speech is detailed in Section II. The speaker recognition task for comparison of shouted speech versus normal speech is explained in Section III. In reporting the experimental results, we assess the performance of the recognition system in comparison to periodogram (FFT), conventional LP, weighted LP (WLP) [1] and stabilized WLP (SWLP) [2] along with Gaussian mixture LP (GMLP) [35] which has recently been shown to be efficient in speaker verification in handling high vocal effort versus normal mode. Section IV provides a comprehensive discussion on the findings and concludes the paper.

II. POWER SPECTRUM COMPRESSION/EXPANSION

In this section, we present the proposed new all-pole modeling technique, LP-α. The name of the method stems from the conventional LP, therefore “LP” is involved in the name, but the method also raises the speech power spectrum to a power
Fourier transform

\[ X(e^{j\omega}) = \mathcal{F}(x[n]) \]

Power spectrum

\[ S(\omega) = X(e^{j\omega})X^*(e^{j\omega}) \]

Power-law non-linearity

\[ S_\alpha(\omega) = (S(\omega))^\alpha \]

Inverse Fourier transform

\[ r_\alpha[n] = F^{-1}(S_\alpha(\omega)) \]

LP-\(\alpha\) analysis filter

Fig. 1: Incorporating power law non-linearity into autocorrelation calculation in LP-\(\alpha\) processing. In the LP-\(\alpha\) processing, the AR modeling is computed using modified autocorrelation \(r_\alpha[n]\).

A. Signal representation

Let us assume that a time-domain speech signal is treated in short-time frames of \(N\) samples by windowing each frame with, e.g., the Hamming window. Let us denote the output of this signal-framing and -windowing by \(x[n]\), with \(n\) as the time sample index \(n \in [0, N - 1]\). By Fourier-transforming signal \(x[n]\) we obtain

\[ X(e^{j\omega}) = \mathcal{F}(x[n]) = \sum_{n=0}^{N-1} x[n]e^{-j\omega n}, \omega \in \mathbb{R}, \]

where \(X(e^{j\omega})\) is the spectrum of \(x[n]\). The power spectrum of \(x[n]\) is computed as

\[ S(\omega) = |X(e^{j\omega})|^2. \]

Autocorrelation of \(x[n]\) is obtained by inverse Fourier-transforming the power spectrum

\[ r[n] = F^{-1}(S(\omega)) = \frac{1}{2\pi} \int_{-\pi}^\pi S(\omega)e^{j\omega n} d\omega. \]

The frequency transforms above are typically treated in discrete forms both in the time and frequency domain by using the discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT). With these transforms we can write the discrete spectrum and power spectrum of \(x[n]\) as

\[ X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi kn/N} \quad \text{and} \quad S[k] = |X[k]|^2, \]

respectively. Similarly, using IDFT, autocorrelation of \(x[n]\) can be written as

\[ r[n] = 1/K \sum_{k=0}^{K-1} S[k]e^{j2\pi kn/K}. \]

The index \(k, k = 0, \ldots, K - 1\) is a discrete frequency index with \(K\) standing for the number of frequency bins. In order to take advantage of fast Fourier transform, the number of frequency bins is always chosen to be a power of 2. When \(K > N\), zero padding is performed on \(x[n]\), and we end up having \(K\) time-domain samples. In the following, we consider having \(K\) samples for \(x[n]\) after zero-padding.

B. LP analysis

Linear prediction is an autoregressive (AR) model for signal estimation. In LP, a signal \(x[n]\) can be predicted as

\[ \hat{x}[n] = \sum_{k=1}^{p} a_k x[n-k], \]

with \(\hat{x}[n]\) standing for the estimated value of the signal at discrete time index \(n\) and \(\{a_k\}\) as AR model parameters with a prediction order \(p\). In the following, discrete signals and LP coefficients are treated in a vector form by using the following notations:

\[ \mathbf{x} = [x[0], x[1], \ldots, x[K-1]]^T, \]

\[ \mathbf{\hat{x}} = [\hat{x}[0], \hat{x}[1], \ldots, \hat{x}[K-1]]^T, \]

\[ \mathbf{a} = [a_1, a_2, \ldots, a_p]^T, \]

where \(\{\}\)^T denotes transposition. Considering the \(l_2\) norm of the residual \(e[n] = x[n] - \hat{x}[n]\), the autocorrelation method is one of the basic approaches to arrive at LP spectrum of a signal. Let us define \(\tilde{r} = [r[0], r[1], \ldots, r[K-1]]^T\) and \(\mathbf{r}\) by using the first \(p\) elements of \(\tilde{r}\), excluding \(r[0]\) and denoting \(\mathbf{r} = [r[1], r[2], \ldots, r[p]]^T\). In the autocorrelation method, the LP coefficients are calculated as

\[ \mathbf{a} = \mathbf{R}^{-1} \mathbf{r}, \]

where \(\mathbf{R}\) is a \(p \times p\) principal submatrix of the Toeplitz matrix \(\tilde{\mathbf{R}}\) with elements

\[ \tilde{\mathbf{R}} = \begin{bmatrix}
  \vdots & \vdots & \ddots & \vdots & \vdots \\
\end{bmatrix}. \]

In this method, the Toeplitz structure of the autocorrelation matrix \(\mathbf{R}\) allows fast and straightforward calculation of autoregressive linear prediction model parameters \(\mathbf{a}\). By using a \(K \times K\) unitary discrete Fourier transform matrix \(\mathbf{F}\) with elements \(F(k,n) = \frac{1}{\sqrt{K}}e^{-j2\pi kn/K}, 0 \leq n, k \leq K - 1\), the estimation of LP coefficients can be easily represented by matrix operations as:

\[ \mathbf{\hat{x}} = \mathbf{Fx} \]

\[ \mathbf{\tilde{r}} = \frac{1}{\sqrt{K}} \mathbf{F}^{-1}(\mathbf{\hat{x}} \odot \mathbf{\hat{x}}^*) \]

\[ \mathbf{a} = \mathbf{R}^{-1} \mathbf{r} \]

The symbol \(\odot\) stands for the element-wise product. Let \(H(z)\) stand for the LP model in the form of an all-pole filter denoted as

\[ H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \]

with \(G\) standing for filter gain. Considering \(P\) complex conjugate pole pairs and \(Q\) real poles, the autocorrelation function of this AR process is a sum of decaying exponentials where
in case of a pair of complex conjugate roots, the damped exponential oscillates [37]. The stability of the frequency response is guaranteed in conventional linear prediction by relying on the fact that the autocorrelation matrix is positive definite [14], [38].

C. LP-α analysis

In the LP-α method, the AR modeling is performed on autocorrelation derived from transformed power spectrum. More specifically, LP-α coefficients, \( a_\alpha \), are calculated from autocorrelation function derived as \( r_\alpha[n] = \mathcal{F}^{-1}(S_\alpha(\omega)) \), where \( S_\alpha(\omega) = (S(\omega))^{\alpha} \). The power law in the form of \((S(\omega))^{\alpha}\) emphasizes or de-emphasizes large spectral components depending on \( \alpha > 1 \) or \( \alpha < 1 \), respectively. The LP-α analysis method based on power-law non-linearity is depicted in Fig. 1 and it can be expressed in a matrix form as:

\[
\tilde{r}_\alpha = \frac{1}{\sqrt{K}} \mathcal{F}^{-1}((\tilde{x} \otimes \tilde{x}^*)^{\alpha})
\]

\[
a_\alpha = R_\alpha^{-1}r_\alpha
\]

The operation \((.)^{\alpha}\) is element-wise. The \( \alpha \)-autocorrelation vector \( r_\alpha \) and \( \alpha \)-autocorrelation matrix \( R_\alpha \) are formed using the elements of \( \tilde{r}_\alpha \) in a fashion similar to that of \( r \), and \( R \) are formed using the elements of \( \tilde{r} \), respectively.

D. Example of LP-α analysis

In order to exemplify the LP-α analysis, we look into a process with power spectral density \( S(\omega) \) specified as

\[
S(\omega) = \frac{1}{(1 - re^{j\phi}e^{-j\omega})(1 - re^{-j\phi}e^{-j\omega})^2},
\]

with \( r \) and \( \phi \) as constants. The conventional LP analysis of this process with \( p = 2 \) yields the AR model

\[
H(z) = \frac{1}{1 - 2r \cos \phi z^{-1} + r^2 z^{-2}},
\]

where the frequency response has a pair of conjugate poles located at \( z = re^{\pm j\phi} = r \cos \phi \pm jr \sin \phi \). Fig. 2 provides insight into the behavior of LP-α analysis corresponding to the process presented in Equation 6 with a pair of conjugate poles located at \( re^{\pm j\phi} = \sqrt{0.5e^{\pm j\pi/4}} \). The plots in Fig. 2 present LP-α analysis (order \( p = 2 \)) performed on the autocorrelation derived from power spectral density in the form of

\[
S_\alpha(\omega) = \frac{1}{((1 - re^{j\phi}e^{-j\omega})(1 - re^{-j\phi}e^{-j\omega})^2)^{\alpha}}.
\]

The application of \( \alpha < 1 \) results in a smoother spectrum when compared to the conventional LP. On the other hand, raising the value of \( \alpha > 1 \) increases the peakiness of the spectrum.

E. Stability of LP-α

The autocorrelation matrix \( R \) is symmetric, real, and has a Toeplitz structure. If \( R \) is positive definite, then \( H(z) \) is minimum-phase, i.e., it has all poles inside the unit circle, and consequently \( H(z) \) is stable [14], [38], [39]. In the

stability analysis of LP-α, we note that in transforming \( S(\omega) \) to \((S(\omega))^{\alpha}\), as long as one keeps the resulting power spectral density real, bounded and strictly positive, the resulting \( \alpha \)-autocorrelation matrix \( R_\alpha \) is positive definite and consequently produces a stable all-pole filter. In the following corollary, we seek to attain a strictly positive definite autocorrelation matrix \( \tilde{R} \), which assures that its inverse (and inverse of its principal submatrix) is also positive definite.

**Theorem** Autocorrelation matrices are strictly positive definite if and only if the discrete Fourier transform of the autocorrelation sequence (i.e., power spectrum) is real and strictly positive.

Since the theorem is an equivalence (if and only if), it follows that if the power spectrum is strictly positive, then autocorrelation matrices constructed thereof are always positive definite.

**Proof** First observe that when applying the discrete Fourier transform of length \( K \) on a time-domain autocorrelation \( r[n] \), we assume that the signal is cyclostationary such that \( r[n] = r[K + n] \). On the other hand, autocorrelation sequences are symmetric with respect to zero, \( r[n] = r[-n] \). It follows that when constructing the \( K \times K \) autocorrelation matrix \( \tilde{R} \) we
have
\[
\mathbf{R} = \begin{bmatrix}
    r[-1] & r[0] & \ldots & r[K-3] & r[K-2] \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
    r[-K+2] & r[-K+3] & \ldots & r[0] & r[1] \\
    r[-K+1] & r[-K+2] & \ldots & r[-1] & r[0] \\
    r[0] & r[1] & \ldots & r[-2] & r[K-1] \\
    r[K-1] & r[0] & \ldots & r[-3] & r[K-2] \\
    \vdots & \vdots & \ddots & \vdots & \vdots \\
\end{bmatrix}
\] (9)

In other words, \( \mathbf{R} \) is a circulant Toeplitz matrix. The eigendecomposition of circulant matrices is of the form \( \mathbf{R} = \mathbf{F} \mathbf{S} \mathbf{F}^H \), where \( \mathbf{S} \) is a diagonal matrix whose coefficients \( S[k] \) correspond to the discrete Fourier transform of \( r[n] \) [40]. We then observe that \( \mathbf{R} \) is strictly positive definite if and only if the coefficients \( S[k] \) are real and strictly positive \( S[k] > 0 \). Conversely, any sequence \( S[k] > 0 \) will correspond to a positive definite autocorrelation matrix \( \mathbf{R} \). Moreover, every \( p \times p \) principal submatrix of \( \mathbf{R} \) will also be positive definite.

In implementing LP and LP-\( \alpha \), a small amount of white noise is added to the speech signal in order to ensure that the power spectral density has positive values (i.e., \( S[k] > 0 \)) for all of the frequency components. The positive definiteness of autocorrelation might be lost when using a finite word length, and it is known that this might result in unstable all-pole filters even in the case of conventional LP [41]. Finite word length results in rounding-off error that might cause the autocorrelation matrix to become ill-conditioned in LP-\( \alpha \), particularly when expanding the dynamic range of the power spectrum using \( \alpha \gg 1 \). Hence, when using LP-\( \alpha \) with \( \alpha \gg 1 \) in applications such as synthesis where filter stability is necessary, a check for the stability of the all-pole filter and use of well-known stabilization procedures are recommended (e.g., [39], [41], [42]).

### F. Time-domain implications

Let us consider a filter having \( p \) poles of \( \beta \)-order in the form of
\[
B(z) = \frac{1}{(1 - \sum_{k=1}^{p} b_k z^{-k})^{\beta}},
\] (10)
where \( \beta \in \mathbb{Q}^+ \) is a rational positive number. \( B(z) \) can be regarded as a cascade of \( p \) filters such that each filter has a fractional pole of order \( \beta \). The power spectral density of \( B(z) \) can be calculated as
\[
S(\omega) = B(z)B(z^{-1})|_{z=e^{j\omega}},
\]
\[
= \frac{1}{(1 - \sum_{k=1}^{p} b_k e^{-j\omega})^\beta} \frac{1}{(1 - \sum_{k=1}^{p} b_k e^{j\omega})^\beta},
\]
\[
= \left[ \frac{1}{(1 - \sum_{k=1}^{p} b_k e^{-j\omega})^\beta} \right] \left[ \frac{1}{(1 - \sum_{k=1}^{p} b_k e^{j\omega})^\beta} \right]^\beta,
\]
\[
= \left[ \frac{1}{(1 - \sum_{k=1}^{p} b_k e^{-j\omega})^2} \right]^\beta.
\]
It is immediately evident that, if we have a filter in the form of \( C(z) = 1/(1 - \sum_{k=3}^{p} b_k z^{-k}) \), the power spectral density of \( B(z) \) and \( C(z) \) are related with a \((.)^\beta \) operation. This suggests that the application of power law on power spectrum of an AR model corresponds to replacing its real and complex conjugate poles of simple order with fractional order [43] poles at the same locations in \( z \)-plane. The multinomial expansion of \( (1 - \sum_{k=1}^{p} b_k z^{-k})^{\beta} \) would result in an infinite order AR model involving fractional delay [44] components for filter.

As it is addressed in [45], a filter with one fractional order pole in the form of
\[
D(z) = \frac{1}{(1 - dz^{-1})^\beta},
\]
can be presented using coefficients of binomial series for \(|z| < |d|, |d| < 1 \) with an impulse response in the form of
\[
d[n] = (-1)^{\beta} \binom{\beta}{n} d^n u[n].
\] (11)
The impulse response is a product of an exponentially decaying part and a polynomial part. This observation implies that a fractional order model is able to represent an impulse response with a slowly varying and a fast varying part. The impulse response of a filter with one fractional pole as given in Equation 11 is proportional to \( n^{-\beta-1} d^n \), which lets us conclude that for \( 0 < \beta < 1 \) the impulse response is slowly decaying. For \( \beta > 1 \), the resulting impulse response is not guaranteed to be bounded in general, and the behavior of time domain signal depends on the specific values of \( 0 < d < 1 \) and \( \beta \). When the term \( d^n \) is decaying faster compared to the ascending rate of \( n^{-\beta-1} \) for \( \beta > 1 \), the impulse response stays bounded and oscillations converge to zero. In raising the power spectrum to the power of \( \alpha \), LP-\( \alpha \) analysis introduces non-linear transformation on power spectrum of signal and then provides a stable AR model for the \( \alpha \)-transformed time domain signal. In fractional LP [45], the fractional order \( \beta \) is an integrated part of the model that is estimated jointly with filter coefficients. Although LP-\( \alpha \) analysis and fractional LP share the view of raising the power spectrum to a power of rational number, the underlying assumptions and parameter estimation methods for the model are different. The resulting filter for fractional LP can be unstable, while LP-\( \alpha \) is guaranteed to give a stable filter as shown in Section II-E. However, under a certain assumption that an AR model \( C(z) \) exists for a signal, LP-\( \alpha \) analysis on the power spectrum of \( C(z) \) can be considered as an approximation for an \( \alpha \) order fractional LP whose power spectrum coincides with the power spectrum of \( C(z) \) raised to the power of \( \alpha \).
Fig. 3: (a) The compression of power spectra (application of \( \alpha < 1 \)) decreases the dynamic range of the spectrum as shown in Fig. 2c but the formant cues are preserved. (b) The spectrum of a short-time window of shouted speech is estimated using conventional LP (\( \alpha = 1 \)) and LP-\( \alpha \) with \( \alpha = 0.1 \).

G. Application of LP-\( \alpha \)

The original studies in [19], [20] approached the transform domain LP analysis by using the power law with the perspective of robust analysis of high-pitched speech. In implementing SQRTA [19], the resulting LP synthesis filter is cascaded with itself to match the spectral dynamics of the original speech signal. STLP [20] also entails application of inverse general transform on power spectrum after LP analysis. In contrast, we utilize the LP-\( \alpha \) filter as such in the spectral feature extraction.

The frequency response plot in Fig. 2c suggests that using \( \alpha \ll 1 \) results in almost flat spectrum. A normalized magnitude spectrum plot of Fig. 2c is depicted in Fig. 3a. The use of normalized amplitude is justified because in most feature extraction algorithms the absolute energy of a short-term speech window is not of interest. Even in a case where the short-term energy is included in the features, a normalization step is utilized to make the features suitable for modeling with for example Gaussian mixture models. Some examples of such common practices in automatic speech or speaker recognition are: excluding the zero-th cepstral coefficient in Mel-frequency cepstral coefficient (MFCC) feature extraction, application of cepstral mean and covariance normalization [46] or employing feature warping [47].

In a typical chain of acoustic feature extraction which includes the above-mentioned normalizations, the resulting features from different values of \( \alpha \) will not depend on the absolute energy of frames. This implies that in mismatched recognition scenarios where shouted speech is compared to the speech of normal loudness, the optimal match can be a result of using LP-\( \alpha \) with \( \alpha > 1 \) for speech in the normal mode and \( \alpha < 1 \) for shouted speech. Such operations are meant to push the spectrum of normal vocal effort speech towards a more peaky form (by application of \( \alpha > 1 \)) while smoothing out the peaky spectrum of shouted speech (by application of \( \alpha < 1 \)).

Estimating the spectrum of speech uttered under high vocal effort using conventional LP often gives rise to sharp peaks. Such situation is plotted in Fig. 3b, where the application of LP-\( \alpha \) on a frame of shouted speech is visualized. It should be pointed out that there is no explicit relationship between the LP model of an acoustic event produced in normal loudness and its corresponding shouted version. As a result, optimal analysis of speech under vocal effort is subject to the application at hand. In the example provided in Fig. 3b, we note that by utilizing LP-\( \alpha \) with \( \alpha = 0.1 \), despite the reduced dynamic range, the resulting smooth spectrum retains important acoustical cues. A smooth spectrum for an acoustic event realized in a high vocal effort is arguably more similar to the spectrum of the same acoustic event phonated in normal loudness.

The spectrum of a normal loudness sound along with the corresponding shouted version is estimated using LP-\( \alpha \) in Fig. 4. The shape of the spectrum in the shouted mode becomes more similar to the spectrum of normal speech when lower values of \( \alpha \) are employed in the analysis of shouted speech. Using LP-\( \alpha \) with a small \( \alpha \) value on both normal and high vocal effort speech decreases dissimilarity at the cost of smoothing out important acoustical cues (i.e., formants). However, LP-\( \alpha \) also allows using an opposite strategy (i.e., \( \alpha > 1 \)), which converts the spectrum of normal speech to have larger dynamics (and therefore closer to the spectra of high vocal effort speech) with formant cues preserved. For recognition systems operating in mismatched vocal effort conditions, it is essential to have similar acoustic features for different realizations of the same acoustic event.

III. SPEAKER RECOGNITION UNDER VOCAL EFFORT MISMATCH

Regardless of several successful attempts in making speaker recognition systems robust against extrinsic variabilities like
In the current study, we experiment with LP-α in text-independent speaker recognition under severe mismatch caused by having speech of normal loudness in speaker modeling and shouting in the test phase. The speech corpus used in our experiments was collected from 11 males and 11 females [53]. Each speaker, all native Finnish, produced 24 utterances using two modes of vocal communication: normal vocal effort and shouting. First, speech samples of normal vocal effort were produced, after which the speakers repeated the same sentences by shouting them. In order not to violate natural production of high-intensity sentences, speakers were not instructed to achieve any pre-defined SPL value but they were advised, instead, to use a very large vocal effort of their own choice when shouting. The output was then controlled by the organizers of the experiment both by listening and monitoring the signal waveform on the computer. If the intensity of shouting was not adequate, the talker was asked to repeat the sentence. The sentences with the average duration of ~1.5 seconds [54] were recorded using a high-quality microphone in an anechoic chamber. The original data were sampled at 16 kHz, but for our experiments the signals were re-sampled at 8 kHz. The difference in SPL between a speaker’s shouted speech and normal speech ranged from 15 to 33 dB for the male speakers and from 17 to 28 dB for the female speakers [53].

A. Experimental Setup

In our speaker recognition system configuration, we fixed all the system parts but the spectrum estimation method in the feature extraction stage. The block diagram of the feature extraction stage is depicted in Fig. 5a. A prediction order of \( p = 20 \) was used for all of the experiments in this paper involving linear prediction for spectrum estimation. First, 19 Mel-frequency cepstral coefficients were extracted and appended by frame energies. After RASTA filtering [55], \( \Delta \) and \( \Delta \Delta \) features were calculated to form 60-dimensional feature vectors. Finally, active speech was retained based on frame-level energy and feature warping [47] was applied. Two configurations for speaker identification were considered: Gaussian mixture model (GMM) [56] and a state-of-the-art i-vector framework [34]. With the GMM model, we tested the speaker recognition system with short utterances of almost 2 seconds. The i-vector-based recognition system was evaluated by having almost 10 seconds of active speech in each of the enrollment and test segments.

The structure of our i-vector-based recognition system is shown in Fig. 5b. We aimed to present the results for the GMM-based system as a proof of concept and considered the i-vector-based system for analyzing the results of this study. For the i-vector recognition system, as it happens in real forensic applications, we took a state-of-the-art recognition system [57], [58] off-the-shelf where recognition system parameters cannot be adapted to the test condition because of data scarcity. The i-vector-based system requires a handful of feature vectors for reliable extraction of sufficient statistics – which is rare in very short utterances –, and we employed an utterance selection protocol different from the one in the GMM case. The utterance selection is exemplified in the pie chart below Tables II and I with 24 segments standing for 24 utterances.

The GMMs with a diagonal covariance structure were trained with 32 Gaussians. The i-vector-based recognition system was developed in Radboud University Nijmegen as a
TABLE I: One utterance is left out and a GMM is made with the remaining 23 utterances for each of the speakers. The left-out utterance for each speaker is tested against all models (no cross-gender comparison). Altogether, 264 samples are compared with 11 models for each gender. The identification rate as percentage, for both genders, is calculated as the number of correct identifications divided by 264. An average identification rate is reported under “All”.

<table>
<thead>
<tr>
<th>Spectrum estimation</th>
<th>Normal vs. normal</th>
<th>Shouting vs. normal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males Females All</td>
<td>Males Females All</td>
</tr>
<tr>
<td>FFT</td>
<td>96.6 88.3 92.4</td>
<td>27.7 8.0 17.8</td>
</tr>
<tr>
<td>LP</td>
<td>97.0 91.3 94.1</td>
<td>33.0 15.2 24.1</td>
</tr>
<tr>
<td>WLP [1]</td>
<td>96.6 92.4 94.5</td>
<td>33.0 15.5 24.2</td>
</tr>
<tr>
<td>SWLP [2]</td>
<td>94.7 92.0 93.4</td>
<td>33.0 17.8 25.4</td>
</tr>
<tr>
<td>GMLP [35]</td>
<td>94.3 91.7 93.0</td>
<td>28.0 12.5 20.3</td>
</tr>
<tr>
<td>LP-α (α = 0.33)</td>
<td>96.2 93.2 94.7</td>
<td>31.8 12.5 22.2</td>
</tr>
<tr>
<td>LP-α (α = 0.1)</td>
<td>92.8 91.7 92.2</td>
<td>33.0 15.5 24.2</td>
</tr>
</tbody>
</table>

TABLE II: In experiments with i-vectors, we pull together 12 segments for enrollment and the rest 12 segments for testing. This results in having almost 10 seconds of active speech in each side. The utterance selection for enrollment and testing is circularly shifted to result in 12 possible pairs for each speaker (no cross-gender comparison). Altogether, there are 12 × 11 = 132 comparisons for each gender. The identification rate is calculated as the number of correct identifications divided by 132.

<table>
<thead>
<tr>
<th>Spectrum estimation</th>
<th>Normal vs. normal</th>
<th>Shouting vs. normal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males Females All</td>
<td>Males Females All</td>
</tr>
<tr>
<td>FFT</td>
<td>100.0 98.5 99.2</td>
<td>47.0 11.4 29.2</td>
</tr>
<tr>
<td>LP</td>
<td>100.0 100.0 100.0</td>
<td>34.1 15.9 25.0</td>
</tr>
<tr>
<td>WLP [1]</td>
<td>100.0 100.0 100.0</td>
<td>32.6 18.9 25.8</td>
</tr>
<tr>
<td>SWLP [2]</td>
<td>100.0 100.0 100.0</td>
<td>40.9 24.2 32.6</td>
</tr>
<tr>
<td>GMLP [35]</td>
<td>99.2 100.0 99.6</td>
<td>31.1 14.4 22.7</td>
</tr>
<tr>
<td>LP-α (α = 0.33)</td>
<td>100.0 100.0 100.0</td>
<td>56.8 23.5 40.2</td>
</tr>
<tr>
<td>LP-α (α = 0.1)</td>
<td>100.0 100.0 100.0</td>
<td>60.6 14.4 37.5</td>
</tr>
</tbody>
</table>

part of the university’s submission to NIST speaker recognition evaluation in 2012 [57], [58]. A gender-dependent universal background model (UBM) [48] with 2048 components was trained using a subset of NIST SRE 2004–2006, Switchboard cellular phase 1 and 2, and the Fisher English corpora. To factorize the GMM mean supervectors, the total variability space [34] was trained with the same data as for UBM with 400-dimensions. In post-processing of utterance-level i-vectors, we used linear discriminant analysis (LDA) projection to enhance the separability of classes (speakers) and reduce the i-vectors’ dimension to 200. Prior to probabilistic linear discriminant analysis (PLDA) [51] modeling, we removed the mean, performed whitening using within-class covariance normalization (WCCN) [49] and normalized the length of i-vectors [50].

It should be noted that in experimenting with the i-vector-based system, the change in spectrum estimation affects acoustic features, which means that system parameters in [57] such as UBM, total variability transform and PLDA do not exactly match with the new features in the off-the-shelf scenario. We accept that this mismatch limits us in our ability to achieve the full potential of the recognition system for each spectrum estimation method. However, as it is shown in Fig. 5, both enrollment and test segments pass the same transformation chain in the recognizer and arguably both of the utterances map into the same space even when the transformations do not match completely with acoustic features. The validity of this hypothesis can be confirmed by looking at “normal vs. normal” results in Table II where the system is able to perform with almost no errors when it utilizes different spectrum estimation methods.

B. Application of the same α value for both normal loudness and shouted speech

The results of speaker identification are presented in Tables I and II. In both of the tables, the degradation in identification performance from “normal vs. normal” to “shouting vs. normal” is large and raises the question of tractability of comparing shouted voices to speech of normal loudness. We present the correct identification rate for the periodogram (FFT), conventional LP and LP-α for two widely used values of α = 0.33 (as in PLP) and α = 0.1 (as in PNCC). We show also the recognition results by using weighted LP (WLP) [1] and stabilized WLP (SWLP) [2] along with Gaussian mixture LP (GMLP) [35]. The SWLP method for spectrum estimation has shown superior performance in speaker recognition under vocal effort mismatch [59] compared to other techniques like minimum variance distortionless response (MVDR) [60] and regularized LP (RLP) [10]. GMLP has recently shown effective speaker verification performance in handling high vocal effort versus normal speech.

The conventional LP analysis serves as the base for comparisons, and the recognition results with FFT are provided for reference. Compared to LP, the application of WLP introduces a slight improvement in the identification performance when dealing with shouted speech. However, with SWLP a boost in the recognition performance is observed in Table II.
when comparing shouted versus normal speech. This indicates that having a temporally weighted all-pole spectral modeling method that guarantees stability is not only important from the point of view of all-pole synthesis but it also helps in producing more reliable spectral features. With the use of GMLP [35], the overall recognition performance does not improve. It is worth observing that the results reported in [35] are based on the NIST SRE 2010 high vocal effort data, which hardly resemble shouting of the kind found in the data of the current study.

It is often observed in the NIST speaker recognition evaluations that recognition systems produce inferior performance for females when compared with males [61]. This is sometimes attributed to higher F0 present in females’ speech, yielding lower resolution in sampling the vocal tract shape, which in turn results in more uncertainty in spectrum estimation. In comparing shouted speech to a model trained with speech of normal loudness, as it is reported in Tables I and II, identifying female speakers also turns out to be more challenging than identifying male speakers. Since increased F0 is one of the characteristics of shouted speech, mediocre identification rates for females when compared to males continue to exist in matching shouting with speech of normal loudness.

The speaker identification performance for shouting versus speech of normal loudness shows improvement in Table II compared to Table I, which could be partly due to having a balanced amount of speech for training and testing in the experiments with i-vectors. The transforms utilized in the i-vector-based system were trained on a set of NIST SRE corpora which supposedly does not include high vocal effort data and practically, the speech was of telephone quality or recorded in an interview session. Nevertheless, by utilizing a well-trained recognition system that shares most of its components with the configuration in the present study, the baseline identification rate when comparing shouted speech versus speech of normal loudness stands on 25.0% when conventional LP is used for the spectrum estimation. In practice, a speaker recognition system with the identification rate of 25.0% is not usable, and a dramatic increase in identification rate is required. By utilizing LP-α, a significant improvement in speaker identification for shouted versus normal speech is observed for both α = 0.1 and α = 0.33. It brings the overall identification rate from 25.0% for conventional LP to 40.2% for LP-α (α = 0.33) for the i-vector-based recognition system.

### C. Application of different α values for both speech of normal loudness and shouted speech

In doing experiments with LP-α for speaker recognition under vocal effort mismatch, two different questions are of interest:

- Is there an optimal value of α that can be independently applied on training and testing utterances so that the recognition performance improves in shouted vs. normal mode comparison without compromising the performance in the normal vs. normal comparison?
- If the speaker models are already trained using conventional LP in the spectrum estimation stage, what is the optimal value of α yielding the best recognition performance in a mismatch condition caused by having, in training, speech intensity characteristics that greatly differ from those used in testing?

In order to answer these questions, we employ a range of values for α and let enrollment and test utterances be characterized using different values of α in LP-α analysis. This inconsistent setting of parameters between enrollment and test utterances is not usually considered in normal recognition scenarios. However, forensic speaker recognition scenario where shouting is compared to speech of normal loudness is hardly a normal recognition task. The fact that the test utterance is phonated under extreme vocal tract stretching and abnormal excitation signal compared to a neutral phonation allows us to consider a different parametrization for shouted speech. In this experiment, we stick to LP-α modeling by using different values of alpha for speech of normal loudness and shouted speech. In doing so, we implicitly assume that the system is aware of the vocal effort mode when selecting the alpha parameter in spectrum estimation. Such a prior knowledge is arguably available in handling forensic cases. Otherwise, considering reported equal error rates as low as 3% for automatic shouting detection in noisy environments [62], prior knowledge of the speech mode can be relaxed by utilizing the posterior probability of vocal effort mode.

As it is shown in Fig. 6a, the performance for normal speech both in training and testing is not particularly affected by the selection of α in the range of 0.5 ≤ α ≤ 1.5. However, other choices of α produce different levels of deteriorated performance. By looking at the corresponding range of α in the experiment comparing shouted versus speech of normal loudness in Fig. 6b, it can be clearly observed that conventional LP (α = 1) does not provide the best recognition performance. When conventional LP is used in training, the best performance for the recognition system is attained when the test utterance is parameterized using LP-α with α = 0.05. There is another local maximum occurring around α = 2, which can be attributed to the amplification of higher formants, as can be seen in Figures 3b and 4. Since acoustical cues of higher formants are important for speaker recognition, the observed improved performance for α = 2 could be explained by the occurrence of more prominent high-frequency resonances in the all-pole spectra computed by LP-α. Using α ≪ 1 for shouted speech smoothes out the picky spectrum of LP analysis for high vocal effort speech and brings better matching with speech of normal loudness. The best identification rate for shouted versus normal speech using LP-α is attained by using α = 1.5 for normal vocal effort speech and α = 0.01 for shouted speech.

### IV. Discussion and Conclusion

We presented an acoustic feature extraction scheme in the analysis of shouted speech and speech of normal loudness for efficient speaker recognition. In the spectrum estimation stage, instead of using conventional LP, we took advantage of adjusted LP where a power non-linearity is applied on the power spectrum of speech. An α parameter is utilized in
Fig. 6: The application of LP-α on speaker identification using the i-vector-based system depicted in Fig. 5. Using a different α value for training and test utterances is examined for (a) normal vs. normal and (b) shouting vs. normal. The identification rates are reported for the “All” scenario where both males and females are considered. Two performance points marked as “Conventional LP” and “Proposed setting for LP-α” indicate that using LP-α with α = 1.5 for enrollment utterances of normal loudness and α = 0.01 for shouted test utterances results in a considerable boost in speaker recognition performance compared to using conventional LP (α = 1) in both enrollment and testing (Fig. 6b). The same comparison with speech of normal loudness both in enrollment and testing indicates a negligible performance decline for LP-α compared to conventional LP (Fig. 6a).

As demonstrated by Fig. 3 and Fig. 4, using α ≪ 1 results in a remarkable reduction in the dynamic range and subsequently also in the energy of the all-pole spectrum. Hence, with α ≪ 1 the all-pole spectra computed from speech of normal loudness become more similar to those extracted from shouted speech, and one would be tempted to assume that this is reflected in improved recognition performance. Our experimental results in Fig. 6b, however, indicate that using an excessively reduced value of α = 0.005 for both speech of normal loudness and shouted speech does not give the best recognition performance. The best performance is obtained by making the all-pole spectrum of speech of normal loudness more peaky (application of α = 1.5) in enrollment and smoothing out the peaky all-pole spectrum of shouted speech (application of α = 0.01) for test utterances.

As shown in Table III, when the training and test signal are in the same domain, the GMM-based recognizer is able to perform reasonably well. As suggested in [64], we have also experimented with mixed training of models. In doing so, the test protocol presented in Table I does not change and only the corresponding 23 shouted utterances will be added to each enrollment. This results in having both normal and shouted versions of each utterance present for the GMM training.

Table III: Effect of training with shouted speech and mixed training. The test protocol is the same as in Table I and conventional LP spectrum estimation is being used.

<table>
<thead>
<tr>
<th>Enrollment</th>
<th>Normal Males</th>
<th>Normal Females</th>
<th>All</th>
<th>Shouted Males</th>
<th>Shouted Females</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>97.0</td>
<td>91.3</td>
<td>94.1</td>
<td>33.0</td>
<td>15.2</td>
<td>24.1</td>
</tr>
<tr>
<td>Shouting</td>
<td>27.7</td>
<td>12.1</td>
<td>19.9</td>
<td>96.6</td>
<td>87.5</td>
<td>92.0</td>
</tr>
<tr>
<td>Mixed training</td>
<td>80.7</td>
<td>86.4</td>
<td>83.5</td>
<td>93.9</td>
<td>82.6</td>
<td>88.3</td>
</tr>
</tbody>
</table>
The recognition results for the mixed training in Table III reveal that by introducing both shouting and normal speech in GMM training, the model captures an average acoustic space and produces favorable recognition results for testing both with speech of normal loudness and with shouted speech. The feature compensation approach for speaker identification under vocal effort mismatch presented in [65] was aimed at constructing an intermediate “vocal effort free” feature space. The acoustic features would be then mapped to this space before enrollment and testing. Marginal improvement over conventional LP is reported in [65] for comparing shouted speech versus speech of normal loudness, highlighting the difficulty of finding a general mapping to strip away the vocal effort effect for speaker recognition.

The results in Table I are indicative and serve the purpose of showing the recognition rates for a basic recognition system. In terms of the results for the GMM-based recognition system, LP-$\alpha$ performs in line with other spectrum estimation techniques. The results in Table II are for the state-of-the-art i-vector-based recognition system. For LP-$\alpha$, we are reporting the results where, in both the training and testing phases, the same value of $\alpha$ is used. One of the most important results of the study is presented in Fig. 6, where a recognition rate of 49% is attained by selecting $\alpha = 1.5$ for utterances of normal loudness in enrollment and $\alpha = 0.01$ for shouted utterances in test. Such selection of $\alpha$ results in slight performance degradation (from 100% to 97%) when using speech of normal loudness both in enrollment and test. Considering the best recognition performance of the baseline spectrum estimation methods in comparing shouted speech to speech of normal loudness (32.6% obtained by SWLP), the recognition rate of 49% for LP-$\alpha$ with $\alpha = 1.5$ in training and $\alpha = 0.01$ in test is remarkably better.

The human sensory system shows adaptation behavior which helps neural systems to efficiently encode non-stationary signals by changing their coding strategy or computation [66]. Unlike in a situation where the focus is on recognizing the speaker, the encoding strategy of the auditory system can change when the focus is on recognizing phonemes. It is possible that speaker recognition mechanisms in humans utilize expansive or compressive power-law adaptation in order to match the information in critical bands of speech spectrum when shouted speech is being compared to normal speech. Such adaptation could happen in the course of time and be dependent on the dynamic range of short-term power spectral density. Based on findings in [67], it is viable to search for an adaptive power law expression depending on the F0 of the short-time speech segment. The parameter $\alpha$ can be adaptively found from the speech signal. Utterance-level speech characteristics, such as average spectral tilt, can point out an $\alpha$ to be used where vocal effort is unchanged in an utterance. The compression/expansion factor can be more specifically tuned for each time-frame of a speech signal based on for example spectral flatness [53] or fundamental frequency.

The prominent recognition performance by SWLP shown in Table II suggests that LP-$\alpha$ can benefit from temporal weighting. Another important point is that identification errors are not distributed uniformly across speakers. In our experiments, we have noted that if we can tune the $\alpha$ parameter in the analysis of shouted speech for a particular speaker, independent of other speakers, a marginally higher identification rate is attained. There could be a correlation between measured sound pressure levels for individual speakers, the $\alpha$ parameter in LP analysis and respective identification rate for shouted versus normal speech. Further analysis of speaker-dependent factors and automatic tuning of the $\alpha$ parameter is left for future research.

V. Acknowledgement

This work was supported by the Academy of Finland (project numbers 256961 and 284671). The authors thank Dr. Jouli Pohjalaainen for fruitful discussions. We acknowledge the computational resources provided by the Aalto Science-IT project.

REFERENCES


Rahim Saeidi received his Ph.D. degree in computer science from the University of Eastern Finland (UEF, formerly University of Joensuu) in 2011. He was a Marie Curie post-doctoral fellow in an EU project (Bayesian biometrics for forensics) when working at Radboud University Nijmegen in Netherland from 2011 to 2013. Currently, he is a postdoctoral researcher at the Department of Speech Processing and Acoustics, Aalto University, Finland. His research interests are robust speaker and speech recognition and speech enhancement.

Pauvo Alku received his M.Sc. Lic.Tech. and Dr.Sc.(Tech) degrees from Helsinki University of Technology, Espoo, Finland in 1986, 1988, and 1992, respectively. He was Assistant Professor at the Asian Institute of Technology, Bangkok, Thailand, in 1993, and the University of Turku, Finland, from 1994 to 1999. Currently, he is Professor of speech communication technology at Aalto University, Espoo, Finland. His research interests are analysis and parameterization of speech production and their utilization in HMM-based speech synthesis, spectral modelling of speech, speech enhancement and cerebral processing of speech.

Tom Bäckström (S’02–M’04) is Professor of Speech Coding at the International Audio Laboratories Erlangen (AudioLabs), which is a joint research unit between the Friedrich-Alexander University Erlangen-Nürnberg (FAU) and Fraunhofer Institute of Integrated Circuits IIS, Erlangen, Germany. He received his M.Sc. and D.Sc. (tech.) degrees from Helsinki University of Technology, Finland, in 2001 and 2004, respectively, and continued there as a post-doc until 2008. Since then he has worked at AudioLabs, first as a researcher and from 2013 as professor. His research interests include speech and audio coding and perception, speech production, time-frequency analysis, matrix and polynomial algebra.