On Reducing Harmonic and Sampling Distortion in Vocal Tract Length Normalization

Néstor Becerra Yoma, Claudio Garretón, Fernando Huenupán, Ignacio Catalán, and Jorge Wuth Sepúlveda

***Abstract*—This paper proposes a novel feature-space VTLN (vocal tract length normalization) method that models frequency warping as a linear interpolation of contiguous Mel filter-bank energies. The presented technique aims to reduce the distortion in the Mel filter-bank energy estimation due to the harmonic composition of voiced speech intervals and DFT (discrete Fourier transform) sampling when the central frequency of band-pass filters is shifted. This paper also proposes an analytical maximum likelihood (ML) method to estimate the optimal warping factor in the cepstral space. The presented interpolated filter-bank energy-based VTLN leads to relative reductions in WER (word error rate) as high as 11.2% and 7.6% when compared with the baseline system and standard VTLN, respectively, in a medium-vocabulary continuous speech recognition task. Also, the proposed VTLN scheme can provide significant reductions in WER when compared with state-of-the-art VTLN methods based on linear transforms in the cepstral feature-space. The warping factor estimated with the proposed VTLN approach shows more dependence on the speaker and more independence of the acoustic-phonetic content than the warping factor resulting from standard and state-of-the-art VTLN methods. Finally, the analytical ML-based optimization scheme presented here achieves almost the same reductions in WER as the ML grid search version of the technique with a computational load 20 times lower.**

***Index Terms*—Speech analysis, speech recognition, vocal tract length normalization.**

I. Introduction

VOCAL tract length normalization (VTLN) is one of the most popular techniques applied in speech recognition in recent years [1]–[4]. VTLN attempts to reduce the mismatch between training and testing condition in ASR caused by inter-speaker variability as a result of length differences in the human vocal tract. The main idea of VTLN is to align formants between the test speaker and a reference speaker-independent or dependent model. VTLN is usually implemented in the front-end by scaling the frequency axis [5]–[7] or by shifting band-pass filter centre frequencies within filter-banks [1]. Both alternatives can be performed using an optimal warping parameter or factor which is obtained by optimizing a Maximum Likelihood (ML) criterion over the adaptation data via a grid search. As a result, each generated frequency axis or bank-filter per warping factor has to be evaluated [1]–[4] according to the likelihood of the observed feature vector sequence.

Modeling frequency warping as a linear transform (LT) in the cepstral domain is a strategy that has been followed by some authors [4]–[6], [8]–[15]. As mentioned in [8], applying VTLN as a LT in the feature-space in cepstral-based ASR presents substantial benefits. For instance, due to the fact that the transform can be applied in the original cepstral features, there is no need to compute the log-filter-bank energies and the discrete cosine transform (DCT) for each evaluated warping factor in the grid search. As a result, the computational load of the VTLN estimation can be dramatically reduced [9].

The LT that models the spectral warping function can be represented in the cepstra [4], [10] or in the discrete cepstral space [4]–[6], [8]–[15]. Those techniques can be interpreted as a particular case of Maximum Likelihood Linear Regression (MLLR) [16]. In both groups of techniques the optimal warping factor can be estimated by employing the ML grid search or an analytical gradient-based optimization procedure. For instance, in [5], [6], [8], [15] the optimization is performed by making use of the ML criterion with an Expectation-Maximization (EM) auxiliary function [17].

The vocal tract frequency response is a continuous function represented by a spectral envelope. However, this frequency response or spectral envelope is evaluated by using two independent discrete sampling processes: first, band-pass filters are modeled with a DFT, which in turn provides a given number of samples within the filter bandwidth; second, the harmonic components in voiced signals sample the vocal tract frequency response at multiples of F_0. In Mel filter-banks, which are widely employed in ASR, the filter bandwidths follow the Mel scale. As a consequence, shifting the central frequencies of band-pass filters can introduce perturbations in filter energy estimation due to the discontinuities caused by the DFT and the harmonic structure of voiced signals. This problem is especially acute at low frequencies where the filter bandwidth is narrower according to the Mel scale. For instance, Fig. 1 compares the smoothed spectrum obtained with a moving one-bark bandwidth triangular filter with the smoothed spectrum estimated with the linear interpolation of adjacent Mel filters. As can be seen in Fig. 1(a), the smoothed spectrum obtained with the moving triangular filter is clearly distorted, especially at low frequencies, when compared with the reference spectral envelope. In contrast, the linear interpolation of adjacent Mel filter energies results in a smoothed spectrum that is much more similar to the spectral envelope (Fig. 1(b)). The pitch values within a sentence are highly correlated, and we note that the F0 contour does not exhibit large discontinuities [18]. Also, perturbations within a frame result in a likelihood error, which

Manuscript received July 25, 2011; revised February 08, 2012, June 29, 2012; accepted August 02, 2012. Date of publication August 27, 2012; date of current version October 18, 2012. This work was supported by Conicyt-Chile under Grants Fondecyt 1100195 and 11110391. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Brian Mak.

N. B. Yoma, C. Garreton, I. Catalan, and J. Wuth Sepúlveda are with the Speech Processing and Transmission Laboratory, Department of Electrical Engineering, Universidad de Chile, 8370451 Santiago, Chile (e-mail: nbecerra@ing.uchile.cl).

F. Huenupan is with the Department of Electrical Engineering, Universidad de la Frontera, 4811230 Temuco, Chile.

Digital Object Identifier 10.1109/TASL.2012.2215590

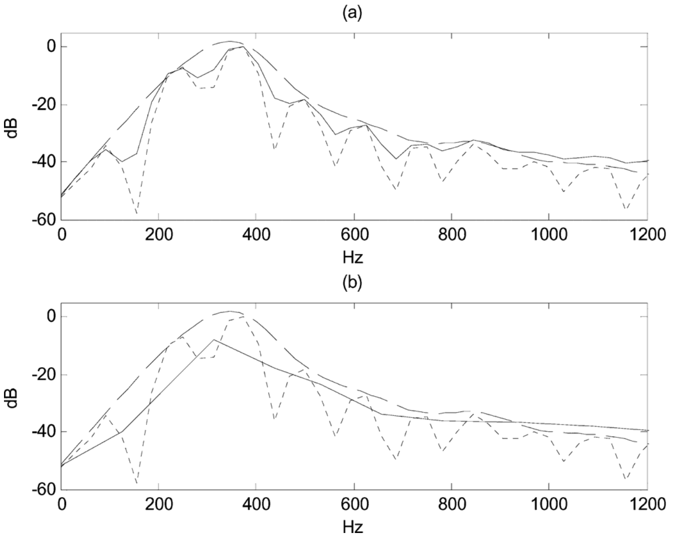

Fig. 1. Spectral density representation of a voiced frame: (a) smoothed spectrum estimated with a moving one-bark bandwidth triangular filters (——); and, (b) smoothed spectrum estimated with the linear interpolation of adjacent Mel filters (——). The original spectral density curve with the harmonic components is represented with (· · · · · · ·) in (a) and (b) and was estimated by using 256 DFT samples. The reference spectral envelope is indicated with (– –) in (a) and (b). The utterance corresponds to a male speaker.

in turn is cumulative on a frame-by-frame basis by definition. Hence, the perturbation due to the harmonic nature of speech will not asymptote to zero as the number of frames increases. Surprisingly, the spectral envelope estimation distortion in VTLN due to the discontinuities caused by the DFT sampling and the harmonic structure of the speech has not been exhaustively addressed in the literature.

In this paper, the warped filter-bank energies are estimated by making use of linear interpolation between contiguous filter energies in the original filter-bank. As a result, the effect of the DFT and harmonic structure of voiced speech intervals is reduced, and hence the perturbation in the spectral envelope estimation is minimized. Moreover, an analytical ML-based optimization scheme to obtain the warping factor is derived to replace the grid search.

The solution presented here could also be seen as a spectral smoothing method. In this sense, the problem of spectral smoothing has also been addressed by other authors in speech processing. In [19] a method to reconstruct a smoothed time-frequency representation of speech was proposed to reduce the interference caused by the periodicity. In [20] the effect of conventional triangular Mel and uniform-bandwidth filters was investigated in the context of recognition performance for children's speech. Accordingly, it is shown that "differences in spectral smoothing lead to loss in recognition performance with conventional VTLN". However, despite the fact that smoothed spectral estimation is a well known problem in the field of speech science and technology, the VTLN method proposed here has not been found in the literature. Observe that the distortion caused by the harmonic nature of voiced speech in the estimation of warped filter energy is much more evident in the speech of children. Nevertheless, there is no reason to assume that this distortion does not exist in the speech of adults as well. In fact, Fig. 1 clearly shows how the harmonic composition of the voice introduces perturbations in the estimation of the spectral envelope.

The contribution of the paper concerns: a) a VTLN model in the filter-bank energy domain based on the interpolation of filter-bank energies (IFE-VTLN); b) an analytical ML estimation of the optimal warping factor according to the IFE-VTLN model; and, c) a comparative analysis of the proposed VTLN method regarding the speaker dependency of the estimated warping factor. It is worth mentioning that the proposed method is also applicable to the interpolation of adjacent filter-bank log energies using a similar mathematical analysis.

As shown later, the interpolated filter-bank energy-based VTLN proposed here leads to a linear transform in the cepstral feature-space by approximating the logarithmic function with a first order Taylor series. Experiments with the LATINO-40 database suggest that the presented method can lead to relative reductions in WER as high as 11.2% and 7.6% when compared with the baseline system and standard VTLN, respectively. When compared with state-of-the-art VTLN methods, the proposed ML grid search scheme leads to significant relative reductions in WER equal to 7.0% on average. Moreover, the proposed analytical ML-based optimization scheme achieves almost the same reductions in WER as the ML grid search version of the technique with a computational load 20 times lower. Finally, the warping factor computed with the VTLN approach described here shows more dependence on the speaker and more independence of the acoustic-phonetic content than the warping factor resulting from standard VTLN and state-of-the-art VTLN methods. This result is observed as a lower gender classification error rate, when the warping factor is employed as a single classifier feature, and as a lower averaged standard deviation per speaker of the warping parameter.

II. Frequency Warping and Filter Energy Interpolation

Consider that ω_m is the central frequency of filter m in a filter-bank composed of M filters. Then, $\hat{\omega}_m$ is the warped central frequency of filter m. By using the linear piece-wise warping function described in [1], [4], $\hat{\omega}_m$ can be written as:

$$\hat{\omega}_m(\alpha) = \begin{cases} \alpha \cdot \omega_m & \omega_m \leq \omega_0 \\ \alpha \cdot \omega_0 + \dfrac{\omega_{\max} - \alpha \cdot \omega_o}{\omega_{\max} - \omega_0}(\omega_m - \omega_0) & \omega_m \geq \omega_0 \end{cases} \tag{1}$$

where $\omega_{\max}$ corresponds to the highest filter-bank frequency, α is the warping factor or parameter, and ω_0 is defined as follows:

$$\omega_0 = \begin{cases} \dfrac{7}{8}\omega_{\max} & \alpha \leq 1 \\ \dfrac{7}{8 \cdot \alpha}\omega_{\max} & \alpha > 1 \end{cases} \tag{2}$$

The energy of filter m at frame i is denoted by $X_{i,m}$. The VTLN method proposed in this paper estimates the energy of warped filter $m, \hat{X}_{i,m}$, as a linear combination of contiguous filter energies in the original filter-bank: if warped filter m is shifted to the left (i.e., $\alpha \leq 1$), the warped filter energy is estimated with a linear interpolation between $X_{i,m-1}$ and $X_{i,m}$; and, if warped filter m is shifted to the right (i.e., $\alpha \geq 1$), the warped filter energy is approximated with a linear interpolation between $X_{i,m}$ and $X_{i,m+1}$. Accordingly, $\hat{X}_{i,m}$ is expressed as:

$$\hat{X}_{i,m}(\alpha) = \frac{X_{i,m} - X_{i,q}}{\omega_m - \omega_q}\left[\hat{\omega}_m(\alpha) - \omega_m^{\mathrm{ref}}\right] + X_{i,m}^{\mathrm{ref}} \tag{3}$$

where,

$$q = \begin{cases} m-1 & \alpha \le 1 \\ m+1 & \alpha > 1 \end{cases} \tag{4}$$

and, $X^{\text{ref}}_{i,m}$ and ω^{ref}_{m} are defined as follows:

$$X^{\text{ref}}_{i,m} = \frac{X_{i,m} + X_{i,q}}{2} \tag{5}$$

$$\omega^{\text{ref}}_{m} = \frac{\omega_m + \omega_q}{2} \tag{6}$$

Conventional VTLN is usually implemented by generating a filter-bank for each and every warping factor α to be evaluated. Then, the optimum α is the one that maximizes the likelihood. According to the model presented here, the filter-bank energies for each α to be evaluated can be computed with (3) without the need to run a filter-bank analysis for each α. Observe that (3) could be replaced by a nonlinear interpolation with the use of $X_{i,m\pm 2}$, $X_{i,m\pm 3}$, etc. However, the motivation here is to use a linear interpolation that can be preserved through the cepstral transform.

III. Maximum Likelihood Estimation of Warping Parameter α

Instead of evaluating several warping factors to choose the one that maximizes the likelihood, it is always desirable to estimate the optimal α analytically. In this section an analytical optimization of α based on ML estimation is proposed. By applying the natural logarithm function to (3), filter m log-energy can be written as:

$$\log[\hat{X}_{i,m}(\alpha)] = \log\left(\frac{X_{i,m} - X_{i,q}}{\omega_m - \omega_q}\left[\hat{\omega}_m(\alpha) - \omega^{\text{ref}}_m\right] + X^{\text{ref}}_{i,m}\right) \tag{7}$$

In order to simplify the formulae, let $L_{i,m} = \log[X_{i,m}]$ and $\hat{L}_{i,m}(\alpha) = \log[\hat{X}_{i,m}(\alpha)]$. By applying the first order Taylor series approximation for the log function according to $\log(a + \Delta a) \cong \log(a) + (\Delta a)/(a)$, if $a \gg \Delta a$, then $\hat{L}_{i,m}(\alpha)$ can be expressed as,

$$\begin{aligned} \hat{L}_{i,m}(\alpha) &\cong \log\left(X^{\text{ref}}_{i,m}\right) + \frac{p_{i,m}}{X^{\text{ref}}_{i,m}}\left(\hat{\omega}_m(\alpha) - \omega^{\text{ref}}_m\right) \\ &= \left[\log\left(X^{\text{ref}}_{i,m}\right) - \frac{p_{i,m}}{X^{\text{ref}}_{i,m}}\omega^{\text{ref}}_m\right] + \frac{p_{i,m}}{X^{\text{ref}}_{i,m}}\hat{\omega}_m(\alpha) \end{aligned} \tag{8}$$

where $p_{i,m} = (X_{i,m} - X_{i,q})/(\omega_m - \omega_q)$. By defining $b^1_{i,m} = (1/X^{\text{ref}}_{i,m})p_{i,m}$ and $b^0_{i,m} = \log(X^{\text{ref}}_{i,m}) - b^1_{i,m} \cdot \omega^{\text{ref}}_m$, (8) can be written as:

$$\hat{L}_{i,m}(\alpha) \cong b^1_{i,m} \cdot \hat{\omega}_m(\alpha) + b^0_{i,m} \tag{9}$$

Observe that the first order Taylor approximation requires $X^{\text{ref}}_{i,m} \gg (X_{i,m} - X_{i,q})/(\omega_m - \omega_q)[\hat{\omega}_m(\alpha) - \omega^{\text{ref}}_m]$. By considering $|\hat{\omega}_m(\alpha) - \omega^{\text{ref}}_m| < |\omega_m - \omega_q|$, the condition that satisfies the first order Taylor series can be evaluated by:

$$\left|\frac{X_{i,q} - X_{i,m}}{X^{\text{ref}}_{i,m}}\right| \le \gamma \tag{10}$$

where γ is a threshold to discard frames if condition in (10) is not satisfied by components $\hat{X}_{i,m}(\alpha)$, where $0 \le m \le M-1$. By incorporating in (9) the definition of $\hat{\omega}_m(\alpha)$ according to (1), $\hat{L}_{i,m}(\alpha)$ can be re-written as:

$$\hat{L}_{i,m}(\alpha) \simeq \begin{cases} b^1_{i,m} \cdot \alpha \cdot \omega_m + b^0_{i,m} & \omega_m \le \omega_0 \\ \alpha \cdot D + E & \omega_m > \omega_0 \end{cases} \tag{11}$$

where $D = b^1_{i,m} \cdot [\omega_0 - (\omega_o/(\omega_{\max} - \omega_0))(\omega_m - \omega_0)]$ and $E = b^1_{i,m} \cdot (\omega_{\max}/(\omega_{\max} - \omega_0))(\omega_m - \omega_0) + b^0_{i,m}$. Consider that the observed unwarped MFCC feature vector sequence is denoted with $C = \{C_i\}_{i=0}^{I-1}$, where: $C_i = \{C_{i,n}\}_{n=0}^{N-1}$ corresponds to the frame at instant i, and I is the number of frames; and $C_{i,n}$ denotes the nth cepstral coefficient at frame i, and N is the number of static cepstral parameters. Then, by applying the DCT, $C_{i,n} = \sum_{m=o}^{M-1} L_{i,m}\cos((\pi \cdot n)/(M)(m - 0.5))$. Consequently, by making use of (11), the nth warped cepstral coefficient at frame i, $\hat{C}_{i,n}$, can be written as:

$$\begin{aligned} \hat{C}_{i,n}(\alpha) &= \sum_{m=o}^{M-1} \hat{L}_{i,m}(\alpha) \cdot \cos\left(\frac{\pi \cdot n}{M}(m - 0.5)\right) \\ &= \sum_{\substack{m=0 \\ m_0/\omega_{m_0} \le \omega_0}} \left(b^1_{i,m} \cdot \alpha \cdot \omega_m + b^0_{i,m}\right) \\ &\quad \times \cos\left(\frac{\pi \cdot n}{M}(m - 0.5)\right) \\ &\quad + \sum_{m=m_0+1}^{M-1} (\alpha \cdot D + E) \cdot \cos\left(\frac{\pi \cdot n}{M}(m - 0.5)\right) \end{aligned} \tag{12}$$

Accordingly, the MFCC warped feature vector sequence is denoted with $\hat{C} = \{\hat{C}_i\}_{i=0}^{I-1}$, where $\hat{C}_i = \{\hat{C}_{i,n}\}_{n=0}^{N-1}$. Observe that the linear piece-wise equation that defines $\hat{L}_{i,m}(\alpha)$ in (11) leads to a DCT representation with two summations: the first one from $m = 0$ to $m = m_0$, where $\omega_{m_0} \le \omega_0$; and, the second one from $m = m_0 + 1$ to $m = M - 1$. If the sums that depend on α are isolated in (12), $\hat{C}_{i,n}(\alpha)$ can be rewritten as:

$$\begin{aligned} \hat{C}_{i,n}(\alpha) = \alpha \cdot \Bigg\{ &\sum_{\substack{m=0 \\ m_0/\omega_{m_0} \le \omega_0}} \left(b^1_{i,m} \cdot \omega_m\right) \\ &\times \cos\left(\frac{\pi \cdot n}{M} \cdot (m - 0.5)\right) \\ &+ \sum_{m=m_0+1}^{M-1} D \cdot \cos\left(\frac{\pi \cdot n}{M}(m - 0.5)\right) \Bigg\} \\ &+ \sum_{\substack{m=0 \\ m_0/\omega_{m_0} \le \omega_0}} b^0_{i,m} \cos\left(\frac{\pi \cdot n}{M}(m - 0.5)\right) \\ &+ \sum_{m=m_0+1}^{M-1} E \cdot \cos\left(\frac{\pi \cdot n}{M}(m - 0.5)\right) \end{aligned} \tag{13}$$

Then, by defining

$$\begin{aligned} W_{i,n} = &\sum_{\substack{m=0 \\ m_0/\omega_{m_0} \le \omega_0}} \left(b^1_{i,m} \cdot \omega_m\right) \cdot \cos\left(\frac{\pi \cdot n}{M}(m - 0.5)\right) \\ &+ \sum_{m=m_0+1}^{M-1} D \cdot \cos\left(\frac{\pi \cdot n}{M}(m - 0.5)\right) \end{aligned}$$

and

$$B_{i,n} = \sum_{\substack{m=0 \\ m_0/\omega_{m_0} \leq \omega_0}} b^0_{i,m} \cdot \cos\left(\frac{\pi \cdot n}{M}(m-0.5)\right) + \sum_{m=m_0+1}^{M-1} E \cdot \cos\left(\frac{\pi \cdot n}{M}(m-0.5)\right)$$

$\hat{C}_{i,n}(\alpha)$ can be expressed as:

$$\hat{C}_{i,n}(\alpha) = \alpha \cdot W_{i,n} + B_{i,n} \tag{14}$$

Observe that (9) allows the interpolated filter-bank energy-based VTLN according to (3) to be applied as a linear transform in the cepstral feature-space, if condition (10) is satisfied. Finally, it is worth highlighting that (14) makes direct use of filter-bank energies instead of modeling them with the inverse discrete cosine transform as in [8], [9].

A. Proposed VTLN Algorithm

The frequency-warping algorithm proposed in this paper makes use of the first decoding pass best hypothesis alignment provided by the Viterbi algorithm. Consider that λ denotes a sequence of context-dependent phoneme HMMs composed of K states, where s_k denotes a state within the composed HMM, with $0 \leq k \leq K-1$. Also $S = \{s_{k(i)}\}_{i=0}^{I-1}$, where S denotes the sequence of states within λ, represents the first decoding pass best hypothesis alignment given by the Viterbi algorithm computed with C. S associates each frame C_i in C with a state within S denoted by $s_{k(i)}$. The presented approach involves three main steps:

Step 1. Given a feature vector sequence C, first decoding pass best hypothesis S is provided by the Viterbi algorithm.

Step 2. By employing the filter-bank interpolation-based frequency warping model proposed here (IFE-VTLN), the optimal warping parameter α is obtained with ML estimation by employing the first decoding pass best hypothesis from Step 1.

Step 3. Finally, the warped MFCC frame sequence $\hat{C}$ is obtained according to (14).

In Step 2, the frequency warping parameter α is estimated by using the ML criterion:

$$\hat{\alpha} = \arg\max_{\alpha}\left\{\log\{p[\hat{C}(\alpha) \mid \lambda, S, \alpha]\} + \log\left|\frac{d\hat{C}(\alpha)}{dC}\right|\right\} \tag{15}$$

where $\hat{\alpha}$ is the optimal frequency warping parameter and $|(d\hat{C}(\alpha))/(dC)|$ is the Jacobian. According to [21], "the warped features are generated by transforming the frequency axis by a suitable warping function before obtaining the cepstra. The models used in the computation of the likelihood of warped features are trained on unwarped features. Therefore the likelihood computation of the warped features with respect to models trained on unwarped features would not be correct unless the Jacobian of the transformation is also taken into account." Due to the fact that the proposed interpolated filter energy model depends on ω_0 in (1) and (2), $\hat{\alpha}$ is estimated by assuming two conditions separately: $\hat{\alpha}_{\text{left}}$ if $\alpha \leq 1$; and, $\hat{\alpha}_{\text{right}}$ if $\alpha > 1$. The ML estimation of $\hat{\alpha}$ is shown in Fig. 2. According to Fig. 2, firstly, $\hat{\alpha}_{\text{left}}$ is computed by considering $\alpha \leq 1$ and $\omega_0 = (7/8) \cdot \omega_{\max}$ as in (2). Then, when $\alpha > 1$, two iterations are proposed to estimate $\hat{\alpha}_{\text{right}}$: first, with $\omega_0 = (7/8) \cdot \omega_{\max}$; second, with $\omega_0 = 7/(8 \cdot \hat{\alpha}^{(1)}_{\text{right}}) \cdot \omega_{\max}$, where $\hat{\alpha}^{(1)}_{\text{right}}$ is the optimum warping factor obtained at the previous iteration. Finally, $\hat{\alpha}$ is chosen between $\hat{\alpha}_{\text{left}}$ and $\hat{\alpha}_{\text{right}}$ according to which one leads to the maximum likelihood of the first decoding pass alignment.

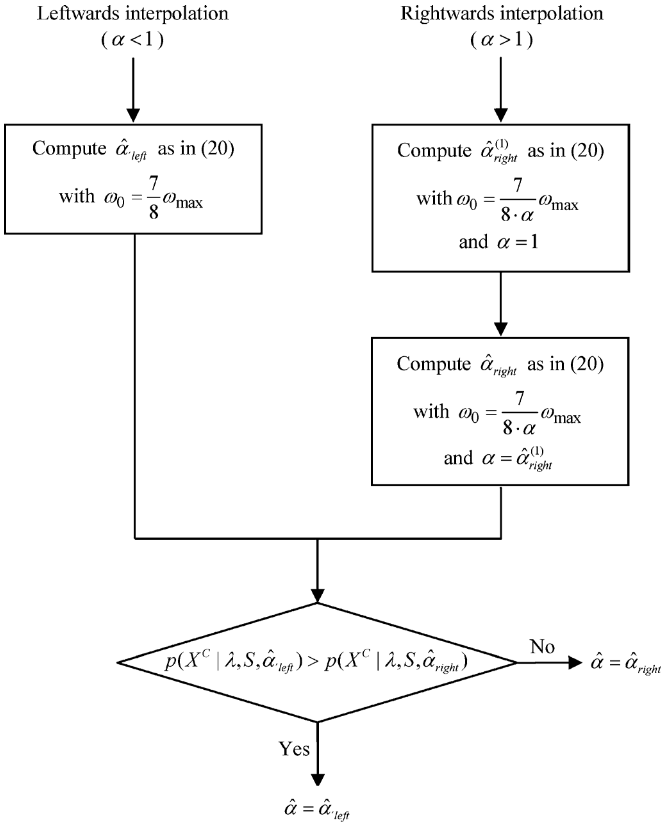
Leftwards interpolation
$(\alpha < 1)$
Rightwards interpolation
$(\alpha > 1)$
Compute $\hat{\alpha}_{left}$ as in (20) with $\omega_0 = \frac{7}{8}\omega_{\max}$
Compute $\hat{\alpha}^{(1)}_{right}$ as in (20) with $\omega_0 = \frac{7}{8 \cdot \alpha}\omega_{\max}$ and $\alpha = 1$
Compute $\hat{\alpha}_{right}$ as in (20) with $\omega_0 = \frac{7}{8 \cdot \alpha}\omega_{\max}$ and $\alpha = \hat{\alpha}^{(1)}_{right}$
$p(X^C \mid \lambda, S, \hat{\alpha}_{left}) > p(X^C \mid \lambda, S, \hat{\alpha}_{right})$
No
$\hat{\alpha} = \hat{\alpha}_{right}$
Yes
$\hat{\alpha} = \hat{\alpha}_{left}$

Fig. 2. Flow chart of the proposed analytical ML-based warping factor estimation method (IFE-VTLN-A).

B. Maximum-Likelihood Estimation of α

As a result of the first decoding pass best hypothesis alignment, the most likely Gaussian per state is chosen. Consequently, state s_k is modeled by a Gaussian function with mean vector $\mu_k = \{\mu_{k,n}\}_{n=0}^{N-1}$ and diagonal covariance matrix Σ_k, and $\phi_k = (\mu_k, \Sigma_k)$. The diagonal components of Σ_k are denoted by $\sigma_k^2 = \{\sigma_{k,n}^2\}_{n=0}^{N-1}$. Then, likelihood $p[\hat{C}_i(\alpha) \mid \phi_{k(i)}, \alpha]$ is defined as:

$$p\left[\hat{C}_i(\alpha) | \phi_{k(i)}, \alpha\right] = \frac{1}{(2\pi)^{\frac{N}{2}} \left|\Sigma_{k(i)}\right|^{\frac{1}{2}}} \times e^{-\frac{1}{2}\sum_{n=0}^{N-1} \frac{[(\alpha \cdot W_{i,n} + B_{i,n}) - \mu_{k(i),n}]^2}{\sigma^2_{k(i),n}}} \tag{16}$$

where $\phi_{k(i)} = (\mu_{k(i)}, \Sigma_{k(i)})$ denotes the set of Gaussian parameters associated to state $s_{k(i)}$ allocated to frame i. Because the Jacobian in (15) is difficult to calculate [4] and is not likely to lead to significant improvements in accuracy [8], [22], the optimal frequency warping parameter can be estimated by maximizing the log-likelihood of the following target function:

$$\begin{aligned}\hat{\alpha} &= \arg\max_{\alpha}\{\log[p(\hat{C}(\alpha)\mid\lambda,S,\alpha)]\}\\ &= \arg\max_{\alpha}\left\{\sum_{i=0}^{I-1}\log[p(\hat{C}_i(\alpha)\mid\lambda,S,\alpha)]\right\}\end{aligned}\tag{17}$$

Notice that the sum on frame index i should consider only those frames whose filter energies comply with condition in (10). Then, by replacing (16) in (17), the optimization can be rewritten as shown in (18) at the bottom of the page, where

$$\sum_{\substack{i=0\\ |(X_{i,q}-X_{i,m})/X^{\mathrm{ref}}_{i,m}|\le\gamma/\forall m}}^{I-1}\log\left[\left((2\pi)^{\frac{N}{2}}\left|\Sigma_{k(i)}\right|^{\frac{1}{2}}\right)^{-1}\right]$$

does not depend on α and is discarded. Then, the optimization in (18) can be solved by computing its partial derivative of (18) with respect to α and setting it to zero:

$$\sum_{\substack{i=0\\ |(X_{i,q}-X_{i,m})/X^{\mathrm{ref}}_{i,m}|\le\gamma/\forall m}}^{I-1}\sum_{n=0}^{N-1}\frac{\left[\alpha\cdot W_{i,n}+B_{i,n}-\mu_{k(i),n}\right][W_{i,n}]}{\sigma^2_{k(i),n}}=0\tag{19}$$

As a result, from (19), $\hat{\alpha}$ is estimated as:

$$\hat{\alpha}=\frac{\displaystyle\sum_{\substack{i=0\\ |(X_{i,q}-X_{i,m})/X^{\mathrm{ref}}_{i,m}|\le\gamma/\forall m}}^{I-1}\sum_{n=0}^{N-1}\frac{W_{i,n}\cdot\left(\mu_{k(i),n}-B_{i,n}\right)}{\sigma^2_{k(i),n}}}{\displaystyle\sum_{\substack{i=0\\ |(X_{i,q}-X_{i,m})/X^{\mathrm{ref}}_{i,m}|\le\gamma/\forall m}}^{I-1}\sum_{n=0}^{N-1}\frac{W^2_{i,n}}{\sigma^2_{k(i),n}}}\tag{20}$$

IV. Experiments

Speaker-independent continuous speech recognition results presented in this paper were obtained by using a medium vocabulary task recorded in a clean environment, the LATINO-40 database [23]. This database is composed of continuous speech from 40 Latin American native speakers, with each speaker reading 125 sentences from newspapers in Spanish. The vocabulary is composed of almost 6000 words. In this paper, experiments were conducted using all 40 speakers as test speakers by employing a non-overlapped "leave-four-out" scheme. As a result, ten sub-experiments were carried out with four test speakers each. One HMM was trained per sub-experiment by employing the utterances from the 36 remaining speakers. Consequently, the training data for each sub-experiment corresponds to 4500 utterances. Also, each sub-experiment contains 500 testing utterances, and hence the whole testing database is composed of 10 sub-experiment × 500 utterances per sub-experiment =5000 utterances. Each utterance is 4.6 seconds long on average and the testing material corresponds to 6.4 hours of recorded speech.

Speech signals were divided into 25-ms frames with fifty percent overlap. The band from 300 to 3400 Hz was covered by 14 Mel DFT filters, and at the output of each channel the logarithm of the energy was computed. The FFT was estimated using 256 samples and thirty-three MFCC parameters (static, delta, and delta-delta coefficients) per frame were computed. Cepstral Mean Normalization (CMN) was also employed. The recognized sentence corresponded to the first hypothesis (the most likely one) within the N-best list obtained from Viterbi decoding. Each triphone was modeled with a three-state left-to-right topology without skip-state transition, with a mixture of eight multivariate Gaussian densities per state with diagonal covariance matrices. The HMMs were trained by using HTK [24] and a trigram language model was employed during recognition. The experiments were conducted by using the recognition engine implemented at the Speech Processing and Transmission Lab., Universidad de Chile. The triphone-based Viterbi algorithm was written by employing ordinary search and pruning techniques in combination with the token passing scheme [25]. The VTLN techniques were applied by estimating the warping factor on an utterance-by-utterance basis with the alignment provided by the best hypothesis in the first Viterbi decoding pass. The baseline system gave a WER equal to 6.42%. The proposed interpolated filter energy model is applied by means of the ML grid search, IFE-VTLN-G, and the proposed ML analytical estimation, IFE-VTLN-A. The proposed method is compared with the linear interpolation of adjacent log filter-bank energies. Also, the VTLN technique presented here is compared with the schemes described in [8], [9], [12], which are denoted by VTLN-LT1, VTLN-LT2 and VTLN-LT3, respectively. Those methods have been recently proposed in the last few years and successfully model VTLN as a LT in the MFCC domain. In the case of [12], an exhaustive grid search was applied to estimate the parameters of the linear combination of adjacent log energies. Observe that there are $3\cdot M-1$ coefficients to optimize for each condition $\alpha\le 1$ and $\alpha>1$, where M is the number of

$$\hat{\alpha}=\arg\max_{\alpha}\left\{\begin{array}{l}\displaystyle\sum_{\substack{i=0\\ |(X_{i,q}-X_{i,m})/X^{\mathrm{ref}}_{i,m}|\le\gamma/\forall m}}^{I-1}\log\left[\left((2\pi)^{\frac{N}{2}}\left|\Sigma_{k(i)}\right|^{\frac{1}{2}}\right)^{-1}\right]\\ \displaystyle -\frac{1}{2}\cdot\sum_{\substack{i=0\\ |(X_{i,q}-X_{i,m})/X^{\mathrm{ref}}_{i,m}|\le\gamma/\forall m}}^{I-1}\sum_{n=0}^{N-1}\frac{\left[\alpha\cdot W_{i,n}+B_{i,n}-\mu_{k(i),n}\right]^2}{\sigma^2_{k(i),n}}\end{array}\right\}\tag{18}$$

TABLE I
WER (%) Obtained With the Baseline System, Standard VTLN, the Proposed IFE-VTLN-G Method, VTLN-LT1, VTLN-LT2 and VTLN-LT3. The Statistical Significances of the Differences With Respect to IFE-VTLN-G Are Presented in Parentheses

	Baseline	Standard VTLN	IFE-VTLN-G	VTLN-LT1	VTLN-LT2	VTLN-LT3
WER (%)	6,42 (p<0.003)	6,17 (p<0.036)	5,70	6,11 (p<0.073)	6,15 (p<0.055)	5.82 (p<0.48)

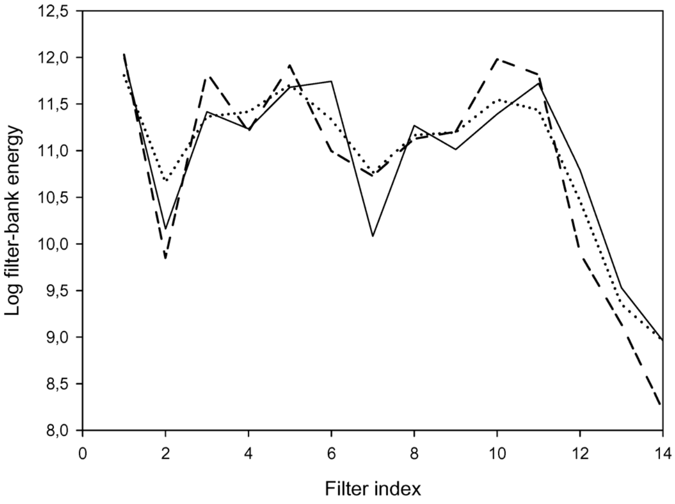

Fig. 3. Log filter-bank energy feature vector that represents the spectral envelope of a voiced speech: original unwarped frame, (——); warped frame with standard VTLN, (– –); and, IFE-VTLN-G, (· · · · · · ·). The utterance corresponds to a male speaker and warping factor was equal to 1.07.

filters. Despite the fact that this strategy requires a high computational load, it allows for comparison with the proposed method without taking into consideration the approximations employed in the MLLR scheme to estimate the linear combination coefficients. Also, the grid search excludes from the analysis the dependence of MLLR on the amount of adaptation data available.

V. Discussion

Fig. 3 shows the log-filter-bank energies of the original and warped filter-banks by employing standard VTLN [1] and IFE-VTLN-G as in (3). According to Fig. 3, the spectral peaks in the filter-bank energy domain provided by IFE-VTLN-G are similar to those with standard VTLN. However, the differences between the spectral peaks and valleys resulting from IFE-VTLN-G are significantly lower than those provided by standard VTLN. This smoothing effect results from the filter-bank energy interpolation. Also, as proposed here, the filter-bank energy interpolation attenuates the discontinuities caused by the DFT sampling and the harmonic structure of the speech spectrum.

Table I shows the WER achieved with the baseline system, standard ML grid search VTLN, IFE-VTLN-G, VTLN-LT1, VTLN-LT2 and VTLN-LT3. The statistical significance of the differences with respect to IFE-VTLN-G is presented in parentheses. When compared with the baseline system, standard VTLN provides a reduction in WER equal to 3.89%. Also, error rates provided by VTLN-LT1 and VTLN-LT2 are very similar to the one obtained with standard VTLN. This result is consistent with those published in [8], [9]. The proposed IFE-VTLN-G scheme leads to relative reductions in WER as high as 11.22%, 7.62%, 6.71% and 7.32% when compared with the baseline system, standard VTLN, VTLN-LT1 and VTLN-LT2, respectively. This result strongly supports the proposed method. The difference in WER between IFE-VTLN-G and VTLN-LT3 is much smaller and is not statistically significant. This result must be due to the fact that the linear interpolation in (9) is a special case of the linear combination of adjacent filter log-energies employed in [12]. Observe that in IFE-VTLN-G only one warping factor α needs to be optimized. In contrast, as explained above, VTLN-LT3 requires the optimization of at least $3 \cdot M - 1$ coefficients per each condition $\alpha \leq 1$ and $\alpha > 1$, which in turn is not feasible for practical purposes. The estimation of these coefficients with MLLR requires the use of approximations and introduces a dependency on the number of adaptation utterances [12]. Consequently, the WER achieved with VTLN-LT3 shown in Table I can be considered a lower bound for the method.

Removing the Jacobian from (15) makes possible the analytical treatment of the optimal warping factor estimation according to (17). To validate this approximation standard ML grid search VTLN and the proposed IFE-VTLN-G were applied with the corresponding Jacobian as in (15). Standard ML grid search VTLN and IFE-VTLN-G with Jacobian give WER equal to 6.10% and 5.7%, respectively. Consequently, including the Jacobian as in (15) leads to no significant reductions in WER with standard VTLN and IFE-VTLN-G, which in turn is consistent with results published elsewhere [8], [22], and validates the optimization according to (17).

The proposed VTLN method attempts to reduce distortion resulting from the harmonic nature of voiced speech in the estimation of warped filter energies (see Fig. 1(a)). As mentioned above, the filter-bank energy interpolation leads to a smoothing effect (see Fig. 1(b)). A similar effect should also be obtained with the interpolation of adjacent log filter-bank energies. If filter energies are replaced with filter log-energies in (3), the ML grid search leads to a WER equal to 5.82%, which in turn is slightly greater than the WER obtained with IFE-VTLN-G. This result suggests that the interpolation of filter energies could be a more appropriate way to approximate the energy of warped filters than the interpolation of log filter-bank energies. Actually, a simple analysis reveals that, for a given warping factor, there is a numerical difference between the warped filter energy estimated with the linear interpolation of filter energies and the one obtained with the interpolation of filter log energies. Nevertheless it is worth emphasizing that the proposed method is equally applicable to both cases with a similar mathematical analysis.

Table II presents the WER provided by the baseline system, standard VTLN, IFE-VTLN-G, VTLN-LT1, VTLN-LT2 and

TABLE II
WER (%) BY GENDER OBTAINED WITH THE BASELINE SYSTEM, STANDARD VTLN, THE PROPOSED IFE-VTLN-G SCHEME, VTLN-LT1, VTLN-LT2 AND VTLN-LT3

	Baseline	Standard VTLN	IFE-VTLN-G	VTLN-LT1	VTLN-LT2	VTLN-LT3
WER-male speakers (%)	6,93	6,65	6,42	6,66	6,30	6,66
WER-female speakers (%)	6,01	5,82	5,11	5.66	6,04	5.19

VTLN-LT3 separated by gender. When compared with the baseline system, IFE-VTLN-G provides a much higher reduction in WER with female speakers than male speakers (14.98% and 7.36%, respectively). A similar result is observed with VTLN-LT3, which in turn models the warped filter log energies as a linear combination of adjacent unwarped log filter-bank energies. Observe that both the linear interpolation of filter energies (or log energies) and the linear combination of log filter energies can be considered spectral smoothing methods. This result strongly supports the hypothesis formulated here and must be due to the fact that female speakers show more separated harmonics in the frequency axis than male speakers. As a consequence, the reduction of the discontinuities due to the harmonic structure of the speech is more relevant for female than male speakers. In contrast, the reductions in WER provided by standard VTLN and VTLN-LT1 with male and female speakers (4.04% vs 3.16% and 3.9% vs 5.8%, respectively) are similar. In the case of VTLN-LT2 significant reductions of WER were observed mainly with male speakers.

As discussed above, the proposed VTLN method shows a higher reduction in WER with female than male speakers. This behavior is not observed with the VTLN schemes implemented here for comparison reasons with the exception of VTLN-LT3, which also results in a spectral smoothing effect. Another procedure to validate the hypothesis related to the distortion caused by the harmonic nature of voiced speech in the estimation of warped filter energies is to apply a spectral smoothing before the Mel filter-bank. Therefore, an experiment was carried out by estimating the spectral envelope with a 20th order LPC filter. Then, the filter-bank energies were estimated and standard ML grid search VTLN was applied to determine the optimal warping factor. Once the optimal warping factor was estimated, the recognition procedure was performed with the original un-smoothed spectrum. This procedure led to a low relative reduction of 1% in WER when compared with standard VTLN. However, not surprisingly, the reduction in WER with female speakers was equal to 3% when male speakers showed an increase of 1% in WER. This result suggests that smoothing techniques have a direct effect in the estimation of warped filter energies. The separation between adjacent harmonics is higher in average in female than in male speakers, which in turn increases the distortion due to the harmonic nature of speech in the estimation of warped filter energies. It is worth mentioning that the warping factor was optimized with the HMMs trained with the original MFCC features without smoothing.

The linear regression according to (14) may suggest a comparison with MLLR [26] that assumes that the adapted HMM parameters can be modeled as a linear transform of the original HMM means and variances. However, significant differences between (14) and MLLR need to be highlighted. First of all, observe that (14) incorporates information of filter-bank log energies. In contrast, MLLR employs a linear regression in the feature vector space (i.e., MFCC). Second, (14) defines warping factor α as the variable to optimize by making use of a well known piecewise linear warping model. On the other hand, MLLR optimizes the whole linear transform. For comparison reasons, fMMLR [27], [28] was implemented on an utterance-by-utterance basis (i.e., the transform is estimated on each utterance) to compensate the unwarped observed feature vector sequence for speaker mismatch. As a result, fMLLR led to a WER equal to 6.09%, which in turn is 5.1% lower than the one provided by the baseline system and is similar to the WER given by standard VTLN, VTLN-LT1 and VTLN-LT2 (see Table I). However, IFE-VTLN-G shows a relative reduction in WER equal to 6.4% $(\mathrm{p} < 0.1)$ when compared with fMLLR. This result suggests that the linear transform optimized by fMLLR can replicate the result achieved with VTLN, but it fails to model the linear regression according to (14). Another comparison is to combine the proposed technique with MLLR. Due to the fact that the linear regression in (14) uses information of both filter log energies and cepstral features, applying MLLR before IFE-VTLN-G makes no sense. When applied after IFE-VTLN-G, MLLR degrades the word recognition accuracy when compared with IFE-VTLN-G alone.

The correlations between the warping factors (α) obtained with standard VTLN and those obtained with IFE-VTLN-G, VTLN-LT1 and VTLN-LT2 are shown in Table III. As can be seen in Table III, the warping factors estimated with IFE-VTLN-G, VTLN-LT1 and VTLN-LT2 are highly correlated with that computed with the standard VTLN. Also, all three correlations are very similar $(0.77 < \text{correlation} < 0.79)$. This result suggests that the proposed technique is as good approximation of standard VTLN as VTLN-LT1 and VTLN-LT2.

Fig. 4 shows the histograms of the estimated warping factors (α) obtained with IFE-VTLN-G and all the grid search-based techniques employed in this paper for comparison purposes: standard VTLN, VTLN-LT1 and VTLN-LT2. The histograms were generated by using all the testing sub-sets and by considering the same range of warping factors (from 0.75 to 1.20). As can be seen in Fig. 4, standard VTLN, VTLN-LT1 and VTLN-LT2 provide rather similar distributions for the warping factor. In contrast, two populations or clusters are clearly identified in the histogram given by IFE-VTLN-G. This result suggests that each cluster of warping factors represent a well defined population of speakers, e.g., speaker gender. This hypothesis is corroborated in Fig. 5, where the gender-depen-

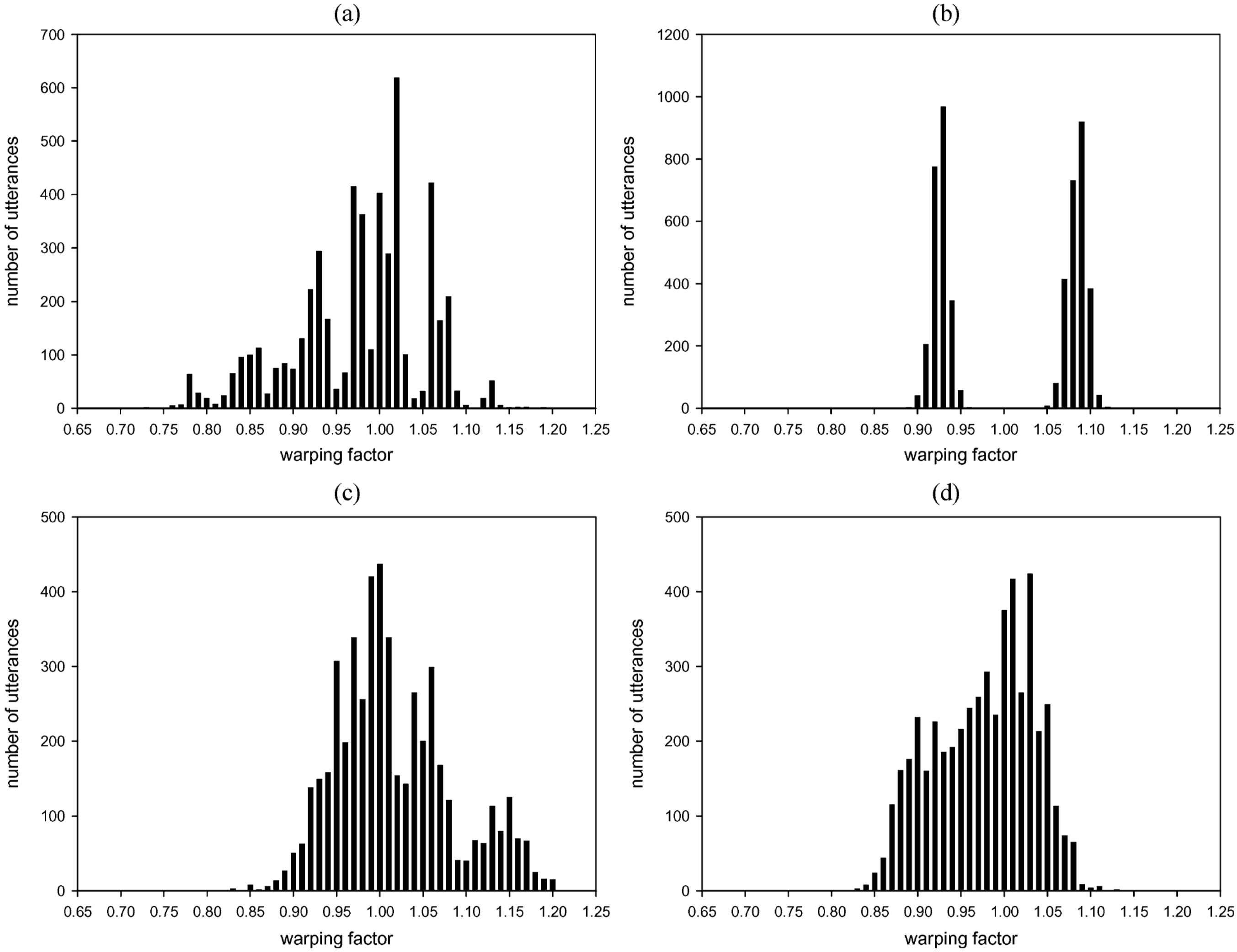

Fig. 4. Histograms of warping factors obtained with: (a) standard VTLN as in [1]; (b) the proposed IFE-VTLN-G scheme; (c) VTLN-LT1 as in [8]; and, (d) VTLN-LT2 as in [9].

TABLE III
CORRELATION BETWEEN THE WARPING FACTOR OBTAINED WITH STANDARD VTLN AND THE WARPING FACTORS OBTAINED WITH THE PROPOSED IFE-VTLN-G TECHNIQUE, VTLN-LT1 AND VTLN-LT2

	IFE-VTLN-G	VTLN-LT1	VTLN-LT2
Correlation	0,79	0,79	0,77

dent histograms of α obtained with IFE-VTLN-G, standard VTLN, VTLN-LT1 and VTLN-LT2 are presented. According to Fig. 5(b), the warping factor estimated with IFE-VTLN-G clearly discriminates between male and female speakers. A similar behavior tends to be observed in Fig. 5(a), (c) and (d). However, the overlap of both populations observed with standard VTLN, VTLN-LT1 and VTLN-LT2 is much higher than the one provided by the proposed IFE-VTLN-G scheme. In fact, the gender classification error rates with IFE-VTLN-G, standard VTLN, VTLN-LT1 and VTLN-LT2 are 4.38%, 9.85%, 10.30% and 20.30%, respectively. This result seems to be very interesting when compared with state-of-the-art gender classification technology that can provide accuracies as high as 95% [29]. Vocal tracts in female speakers are usually shorter than in male speakers, which in turns result in higher formant frequencies. Consequently, the lowest gender classification error rate obtained with IFE-VTLN-G suggests that, given a speaker independent HMM, the warping factor estimated with IFE-VTLN-G should depend more on the speaker and be more independent of the acoustic-phonetic content than the warping factor obtained with standard VTLN, VTLN-LT1 and VTLN-LT2. The gender classification error rate was obtained on a sentence-by-sentence basis.

The distribution of both populations in Fig. 5(b) deserves further discussion. Due to the reduction of the perturbation resulting from the harmonic nature of voiced speech and the linear interpolation of filter-bank energies to estimate warped filter energies, the results presented here suggest that the proposed method generates a warping factor that is more dependent on the speaker than other VTLN schemes. In this sense, a value of α of approximately 1.0 would mean that the testing speaker is perfectly well represented by the speaker-independent HMM. This is not consistent with the uniqueness hypothesis related to the vocal tract. Actually, speaker verification is based on this

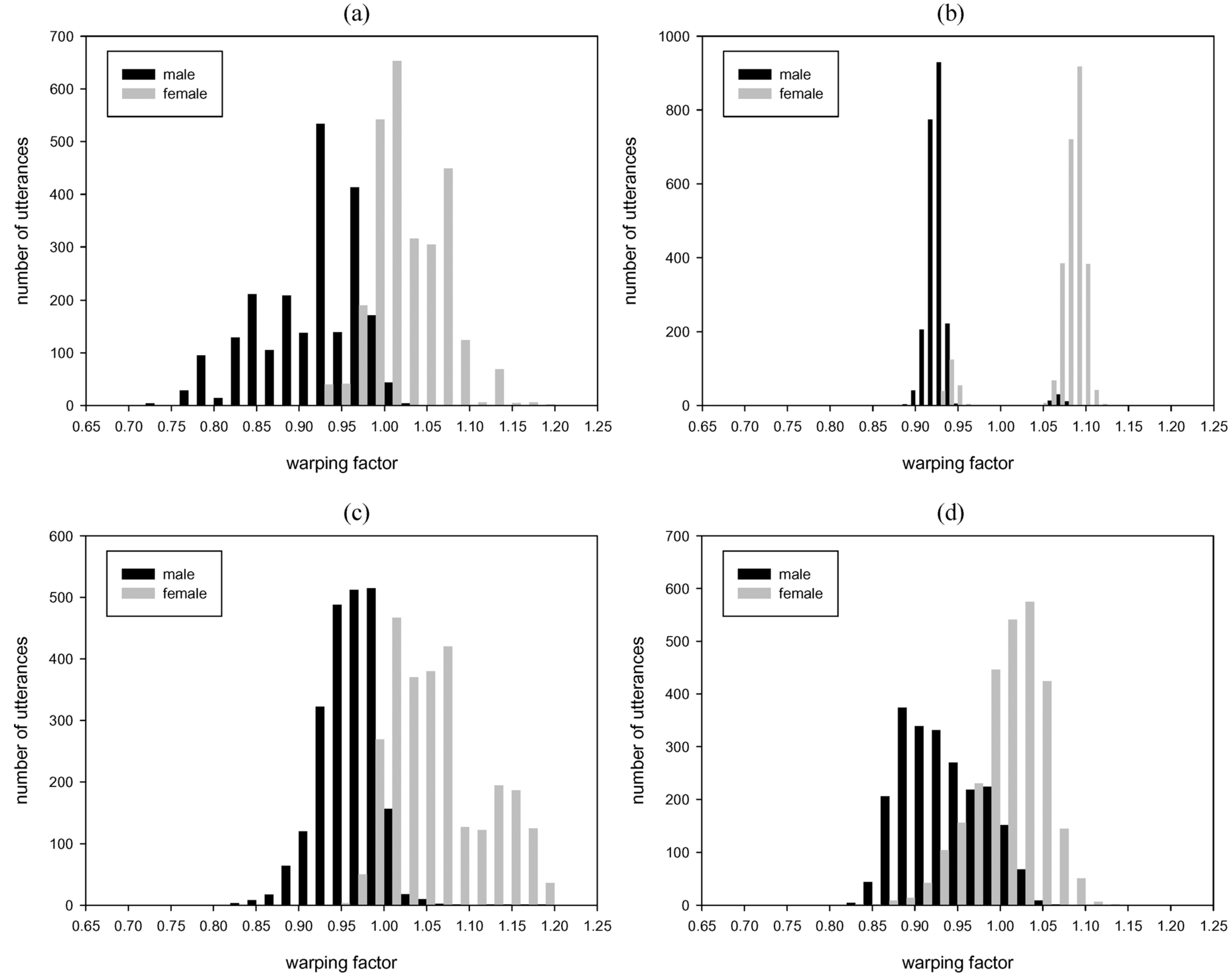

Fig. 5. Histograms of warping factors separated by gender: (a) standard VTLN as in [1]; (b) the proposed IFE-VTLN-G method; (c) VTLN-LT1 as in [8]; and, (d) VTLN-LT2 as in [9].

hypothesis. Although it is not shown in this paper, when the HMM is trained with utterances from the same testing speaker, the warping factor tends to be approximately 1.0.

Table IV shows averaged standard deviation per speaker of the warping factor, $\overline{\sigma_\alpha^{\rm SD}}$, and standard deviation of the warping factor for all the speakers, σ_α, estimated with standard VTLN, the proposed IFE-VTLN-G method, VTLN-LT1 and VTLN-LT2. As presented in Table IV, the average standard deviation per speaker achieved with IFE-VTLN-G is approximately 41% lower than, in average, those provided by regular VTLN, VTLN-LT1 and VTLN-LT2. A similar result is obtained when $\overline{\sigma_\alpha^{\rm SD}}$ is divided by σ_α. The ratio $(\overline{\sigma_\alpha^{\rm SD}})/(\sigma_\alpha)$ is a measure of how concentrated the warping factor is for a given speaker when compared with the overall dispersion of α. This result corroborates the hypothesis discussed above according to which the warping factor estimated with IFE-VTLN-G is more dependent on the speaker and more independent of the acoustic-phonetic content than the warping factor computed with regular VTLN, VTLN-LT1 and VTLN-LT2. It is reasonable to suppose that, given a speaker independent HMM, α is a function of the speaker and should not depend on the acoustic-phonetic content. Consequently, the results shown in Fig. 5 and Table IV also validate the hypothesis that the proposed VTLN scheme can reduce the discontinuities caused by the harmonic structure of the speech, which in turns may cause a less coherent estimation of α.

TABLE IV
AVERAGED STANDARD DEVIATION PER SPEAKER OF THE WARPING FACTOR, $\overline{\sigma_\alpha^{\rm SD}}$, AND STANDARD DEVIATION OF THE WARPING FACTOR FOR ALL THE SPEAKERS, σ_α

	Standard VTLN	IFE-VTLN-G	VTLN-LT1	VTLN-LT2
$\overline{\sigma_\alpha^{SD}}$	0.0330	0.0184	0.0316	0.0291
σ_α	0.0743	0.0797	0.0675	0.0507
$\frac{\overline{\sigma_\alpha^{SD}}}{\sigma_\alpha}$	0.444	0.231	0.468	0.574

TABLE V
WER (%) WITH THE PROPOSED IFE-VTLN-A METHOD VERSUS γ AS DEFINED IN (10). ALSO, THIS TABLE SHOWS THE PERCENTAGE OF FRAMES THAT SATISFIES THE CONDITION IN (10) AND THAT IS EMPLOYED IN THE PROPOSED ML-BASED OPTIMIZATION METHOD ACCORDING TO (20)

			γ			
	0.5	0.6	0.7	0.8	0.9	1.0
WER (%)	5.79	5.87	5.72	5.80	5.75	5.76
% of frames considered	2.38	9.08	23.22	47.07	79.71	100.00

Table V presents the results provided by the proposed analytical ML estimation of the optimum warping factor according to Fig. 2 and (20), IFE-VTLN-A, versus the threshold to discard frames if condition in (10) is not satisfied (γ). As can be seen in Table V, IFE-VTLN-A provides reductions in WER as high as 10.90% and 7.29% when compared with the baseline system and with the standard VTLN, respectively. Also, observe that the improvement due to IFE-VTLN-A slightly depends on γ, which in turn suggests that the linear approximation for the logarithmic function employed in (7) is accurate enough to be used in the analytical ML optimization shown in Fig. 2. Moreover, by comparing the WER achieved by IFE-VTLN-A in Table V with the one obtained with IFE-VTLN-G in Table I, it is possible to conclude that the warping factor α can be reliably estimated by using the analytical ML scheme with just a fraction of frames in the testing utterance.

Regarding the computational load, the proposed IFE-VTLN-G scheme achieves a reduction in WER equal to 7.6% when compared with the standard VTLN with a similar processing time. This processing time does not consider the Viterbi decoding step required to find the optimal alignment and the second Viterbi decoding to obtain the final ASR output after VTLN. When the proposed analytical ML estimation of the warping factor (IFE-VTLN-A) is compared with IFE-VTLN-G, a similar WER is achieved with a computational load 20 times lower.

The correlations between the warping factors (α) obtained with IFE-VTLN-A and those estimated with IFE-VTLN-G and standard VTLN are shown in Table VI. According to Table VI, the warping factors obtained with IFE-VTLN-A are highly correlated with those computed with IFE-VTLN-G $(0.89 < \text{correlation} < 0.94)$. This result indicates than the proposed analytic ML estimation of the warping factor accurately follows the proposed grid search version. Also, the correlation between IFE-VTLN-A and standard VTLN is just slightly lower than the one achieved between IFE-VTLN-G and standard VTLN, as shown in Table III. These results corroborate and validate the analytic optimization scheme according to (20) and Fig. 2. Table VII presents the average standard deviation per speaker of the warping factor versus γ as defined in (10). Comparing with Table IV, IFE-VTLN-A provides an average standard deviation per speaker as low as that achieved with the ML grid search version of the proposed VTLN method.

Fig. 6 depicts the histograms of the estimated warping factors (α) obtained with the proposed IFE-VTLN-A and IFE-VTLN-G schemes. As shown in Fig. 6, IFE-VTLN-A and IFE-VTLN-G provide very similar distributions of the warping factor where two populations or clusters are clearly identified. However it is worth mentioning that the separation between the means of both populations in IFE-VTLN-A is 13% lower than in IFE-VTLN-G. This result is reflected in lower gender discrimination ability when compared with IFE-VTLN-G, which in turns must be due to the low distortion introduced by the linear approximation of the logarithmic function employed in (8). The percentage of the separation was estimated regarding the difference between the average values of both populations.

TABLE VI
CORRELATION BETWEEN THE WARPING FACTOR OBTAINED WITH THE PROPOSED IFE-VTLN-A ALGORITHM AND THE WARPING FACTORS ESTIMATED WITH THE PROPOSED IFE-VTLN-G SCHEME AND STANDARD VTLN

			γ			
	0.5	0.6	0.7	0.8	0.9	1.0
Correlation (IFE-VTLN-G)	0.89	0.93	0.93	0.94	0.94	0.93
Correlation (Standard VTLN)	0.73	0.75	0.75	0.76	0.76	0.76

TABLE VII
AVERAGED STANDARD DEVIATION PER SPEAKER OF THE WARPING FACTOR VERSUS γ AS DEFINED IN (10) PROVIDED BY THE PROPOSED IFE-VTLN-A ALGORITHM

			γ			
	0.5	0.6	0.7	0.8	0.9	1.0
Averaged standard deviation per speaker	0.250	0.176	0.161	0.152	0.147	0.147

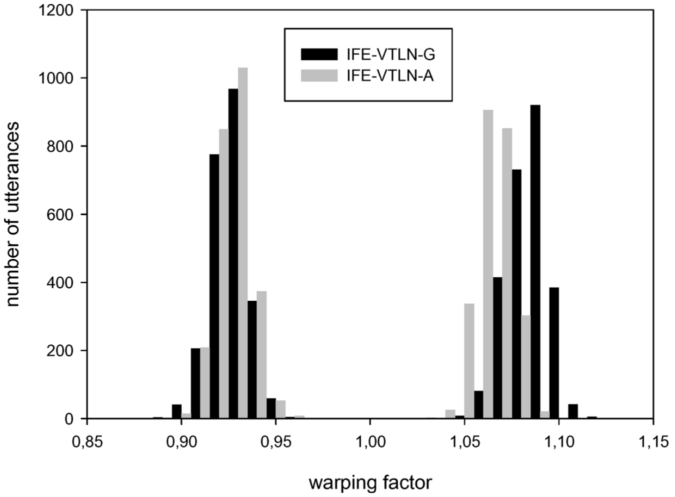

Fig. 6. Histograms of the warping factors obtained with the proposed ML analytical optimization method, IFE-VTLN-A, and the proposed ML grid search, IFE-VTLN-G.

An analysis on stability or robustness in the estimation of γ as defined in (10) provided by the proposed IFE-VTLN-A algorithm is shown in Fig. 7. The entire set of experiments was divided into two subsets, A and B, composed of five sub-experiments each. As can be seen in Fig. 7, the WER is not very dependent on γ in subsets A and B. This result must be due to the fact that γ is defined by a criterion that discards frames to improve the approximation of the first-order Taylor series as described in (8). Consequently, this scheme takes place at a very low level and should be neither task dependent nor speaker dependent.

It is worth noting that, given an original F0 contour, there is already an error in the representation of the "true" spectral envelope by a filter bank as shown in Fig. 1. However, from the point of view of optimization theory, "an optimal policy has the property that whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy

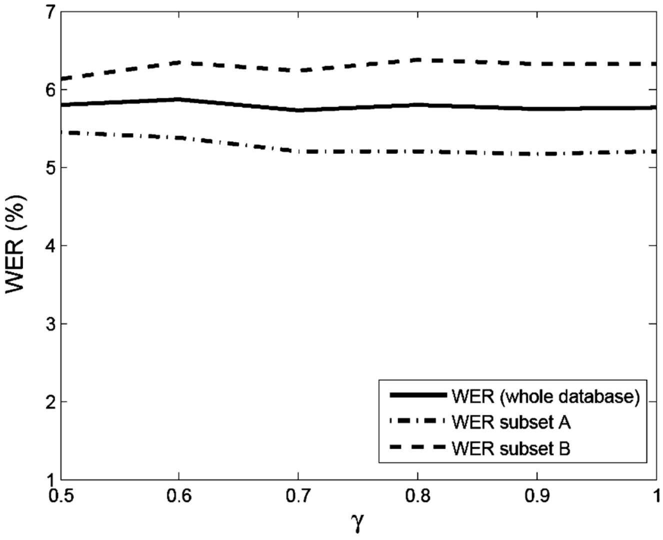

Fig. 7. WER versus. γ as defined in (10) provided by the proposed IFE-VTLN-A algorithm. The whole set of experiments was divided in subsets A and B that correspond to five sub-experiments each.

with regard to the state resulting from the first decision" [30]. Consequently, given an utterance with the F0 contour included, the best one can do is to estimate the VTLN warping factor by reducing the effect of the harmonic representation in the spectral envelope. Observe that the warping factor is estimated by maximizing the likelihood with respect to a speaker-independent HMM. As a result, the warping factor that is determined using the proposed method improves the representation of the observation vector with respect to the current HMM regardless of any possible initial error, which in turn is unavoidable.

It is important to note that the experiments discussed here strongly suggest that the improvements in recognition accuracy resulting from the proposed VTLN method are due to the reduction of the perturbation caused by the harmonic composition of voiced speech by means of the linear interpolation of adjacent filter-bank energies to estimate the energy of warped filters. The estimation of warped filter energies with the linear interpolation of contiguous filter energies results in a spectral smoothing effect, which in turn reduces the distortion due to the harmonic nature of speech and increases the effectiveness of VTLN. Whether the log filter-bank energies are interpolated, or whether the filter-bank energies are interpolated and then a linear approximation is used for the logarithm assuming certain conditions, the resulting transform is a linear regression of the log filter-bank energies and hence the cepstra. In the case of a piecewise-linear frequency warping, the linear regression becomes with respect to just the warping parameter. This reduces the requirements regarding the amount of adaptation data and makes possible the use of ML grid search to estimate α. Also, it should be noted that, according to the results presented here, the interpolation of filter energies leads to slightly better results than the interpolation of filter log-energies.

As a final remark, the proposed approach should be robust with respect to noisy conditions if the spectral distribution of additive noise does not vary much between adjacent filters. According to the linear interpolation in (3), noise should have a low effect in the estimation of $(X_{i,m} - X_{i,q})/(\omega_m - \omega_q)$. Also, $X_{i,m}^{\mathrm{ref}}$ should be reliably compensated with, for example, spectral subtraction. Consequently, $\hat{X}_{i,m}(\alpha)$ could be made robust to noisy conditions by estimating the optimum warping factor with an HMM trained with clean signals. If this is done, the dependence of the optimum warping factor on SNR should be small.

VI. Conclusion

In this paper a feature-space VTLN method that models frequency warping as a linear interpolation of contiguous Mel filter-bank energies (IFE-VTLN) is proposed. The motivation of the presented approach is to reduce the perturbation introduced in the Mel filter-bank energy estimation by the harmonic composition of voiced speech intervals and DFT sampling when the central frequency of band-pass filters is shifted. Also, an analytical Maximum Likelihood optimization method to estimate the warping factor in the cepstral feature-space is proposed. Experiments with a medium-vocabulary continuous speech recognition task show that IFE-VTLN with ML grid search can lead to reductions in WER as high as 11.2% and 7.6% when compared with the baseline system and standard VTLN, respectively. Moreover, IFE-VTLN can lead to significant reductions in WER, equal to 7.0% in average, when compared with state-of-the-art cepstral-based VTLN techniques. It is worth emphasizing that the warping factor estimated with IFE-VTLN is more dependent on the speaker and more independent of the acoustic-phonetic content than the warping factor computed with ordinary and state-of-the-art cepstral-based VTLN methods. In fact, the scheme presented here provides an average standard deviation of the warping factor per speaker approximately 41% lower than regular and state-of-the-art VTLN techniques. Also, IFE-VTLN leads to a gender classification error rate that is at least 50% lower than those obtained with standard and current cepstral-based VTLN methods. Also, the reduction in WER resulting from the proposed analytical ML based optimization scheme is practically the same as the one achieved with the proposed ML grid search, with a computational load 20 times lower. Suggested topics for future research include the combination of the interpolated filter energy VTLN approach with additive noise and channel compensation techniques, and exploring the applicability of the proposed method to gender classification and to speaker adaptive training.

Acknowledgment

The authors would like to thank Prof. Simon King, from CSTR/University of Edinburgh (UK), and Prof. Richard Stern, from CMU (USA), for having proofread this paper.

References

[1] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 49–60, Jan. 1998.

[2] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. ICASSP '96*, 1996, pp. 339–341.

[3] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP '96*, 1996, pp. 346–349.

[4] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 930–944, Sep. 2005.

[5] S. Wang, X. Cui, and A. Alwan, "Speaker adaptation with limited data using regression-tree-based spectral peak alignment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2454–2464, Nov. 2007.

[6] X. Cui and A. Alwan, “Adaptation of children’s speech with limited data based on formant-like peak alignment,” *Comput. Speech Lang.*, vol. 20, no. 4, pp. 400–419, 2006.

[7] A. Acero and R. Stern, “Robust speech recognition by normalization of the acoustic space,” in *Proc. ICASSP ’91*, 1991, vol. 2, pp. 893–896.

[8] S. Panchapagesan and A. Alwan, “Frequency warping for VTLN and speaker adaptation by linear transformation of standard MFCC,” *Comput. Speech Lang.*, vol. 23, no. 1, pp. 42–46, 2009.

[9] S. Umesh, A. Zolnay, and H. Ney, “Implementing frequency-warping and VTLN through linear transformation of conventional MFCC.,” in *Proc. Interspeech ’05*, 2005, pp. 269–272.

[10] J. McDonough, T. Schaaf, and A. Waibel, “Speaker adaptation with all-pass transforms,” *Speech Commun. Spec. Iss. Adapt. Meth. Speech Recognit.*, vol. 41, no. 1, pp. 75–91, 2004.

[11] D. R. Sanand, R. Schlüter, and H. Ney, “Revisiting VTLN using linear transformation on conventional MFCC,” in *Proc. Interspeech ’10*, 2010, pp. 538–541.

[12] G. Ding, Y. Zhu, C. Li, and B. Xu, “Implementing vocal tract length normalization in the MLLR framework,” in *Proc. ICSLP ’02*, 2002, pp. 1389–1392.

[13] T. Claes, I. Dologlou, L. Bosch, and D. Compernolle, “A novel feature transformation for vocal tract length normalization in automatic speech recognition,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 603–616, 1998.

[14] D. Giuliani, M. Gerosa, and F. Brugnara, “Improved automatic speech recognition through speaker normalization,” *Comput. Speech Lang.*, vol. 20, no. 1, pp. 107–123, 2006.

[15] T. Emori and K. Shinoda, “Rapid vocal tract length normalization using maximum likelihood estimation,” in *Proc. Eurospeech ’01*, Aalborg, Denmark, 2001.

[16] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Comput. Speech Lang.*, vol. 9, pp. 171–185, 1995.

[17] T. K. Moon, “The expectation-maximization algorithm,” *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1997.

[18] D. Joho, M. Bennewitz, and S. Behnke, “Pitch estimation using models of voiced speech on three levels,” in *Proc. ICASSP ’07*, 2007, vol. IV, pp. 1077–1080.

[19] H. Kawahara, “Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited,” in *Proc. ICASSP ’97*, 1997, vol. 2, pp. 1303–1306.

[20] S. Umesh and R. Sinha, “A study of filter bank smoothing in MFCC features for recognition of children’s speech,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2418–2430, Nov. 2007.

[21] R. Sinha and S. Umesh, “A method for compensation of Jacobian in speaker normalization,” in *Proc. ICASSP*, 2003, pp. 560–563.

[22] S. P. Rath, A. K. Sarkar, and S. Umesh, “Effect of Jacobian compensation in linear transformation based VTLN under matched and mismatched speaker conditions,” in *Proc. Nat. Conf. Commun. (NCC)*, Jan. 29–31, 2010, pp. 1–5.

[23] J. Bernstein, LATINO-40 Spanish Read News. Philadelphia, PA, 1995, Linguistic Data Consortium.

[24] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book*. Cambridge, U.K.: Cambridge Univ. Eng. Dept., 2006.

[25] S. J. Young, N. H. Russell, and J. H. S. Thornton, “Token Passing: A simple conceptual model for connected speech recognition systems,” Cambridge Univ. Eng. Dept., Tech. Rep. CUED/F-INFENG/TR.38, 1989.

[26] M. J. F. Gales and P. C. Woodland, “Mean and variance adaptation within the MLLR framework,” *Comput. Speech Lang.*, 1996.

[27] A. Ghoshal *et al.*, “A novel estimation of feature-space MLLR for full-covariance models,” in *Proc. ICASSP ’10*, 2010, pp. 4310–4313.

[28] D. Povey and G. Saon, “Feature and model space speaker adaptation with full covariance Gaussians,” in *Proc. Interspeech ’06*, 2006, paper 2050-Tue2BuP.14.

[29] S. Yen-Liang and M. Iseli, “The role of voice source measures on automatic gender classification,” in *Proc. ICASSP ’08*, 2008, pp. 4493–4496.

[30] D. A. Pierre*, Optimization Theory With Applications*. New York: Dover, 1982.

Néstor Becerra Yoma received the Ph.D. degree from the University of Edinburgh, UK, and the M.Sc. and B.Sc. degrees from UNICAMP (Campinas State University), Sao Paulo, Brazil, all of them in electrical engineering, in 1998, 1993 and 1986, respectively. Since 2000, he has been a Professor at the Department of Electrical Engineering, Universidad de Chile, in Santiago, where he is currently lecturing on telecommunications and speech processing, and working on robust speech recognition/speaker verification, language learning, dialogue systems and voice over IP. In 2011 he was promoted to the Full Professor position. At the Universidad de Chile he started the Speech Processing and Transmission Laboratory (LPTV) to carry out research on speech technology applications on the Internet and telephone line. He is the author of 24 journal papers, over 30 conference papers and two awarded patents. Professor Becerra Yoma is a member of the Institution of the Electrical and Electronic Engineers, and the International Speech Communication Association.

Claudio Garretón was born in Santiago, Chile, in 1982. He received the B.Sc., M.Sc. and Ph.D. degrees in electrical engineering from Universidad de Chile, Santiago, Chile in 2005, 2007 and 2011, respectively. From 2005 to 2011 he was a research student at the Speech Processing and Transmission Laboratory (LPTV) at Universidad de Chile, where he worked on techniques for channel distortion and noise canceling in speech recognition, speaker verification and second language learning. He has been a co-author in seven journal papers and seven conference articles in the last years.

Fernando Huenupán received his B.Sc. in electronic engineering from Universidad de La Frontera, Temuco, Chile, in 2004, and the Ph.D. degree in electrical engineering from Universidad de Chile, Santiago, Chile, in 2010. He has been an assistant professor in the Dept. of Electrical Engineering at Universidad de La Frontera since 2010. From 2006 to 2010 he was a research student at the Speech Processing and Transmission Laboratory (LPTV) at Universidad de Chile, where he worked on robust speaker verification based on multiple classifier fusion. His research interests include robustness in speech technology, multiple classifier systems and pattern recognition.

Ignacio Catalán was born in Santiago, Chile, in 1986. He received the B.Sc. degree in electrical engineering from Universidad de Chile, Santiago, Chile in 2011. In 2010 he was a research assistant at the Speech Processing and Transmission Laboratory (LPTV) at Universidad de Chile.

Jorge Wuth Sepúlveda received his electrical engineering degree from Universidad de Chile, Santiago, Chile, in 2007. Since 2007, he has been a research associate at the Speech Processing and Transmission Laboratory (LPTV) where he is currently carrying out his research on speech recognition and computer aided pronunciation training and language learning. Mr. Wuth is the co-author of 3 journal articles and 2 conference papers. His research interests include speech recognition and web-based human machine interfaces.