Frequency Warping and Robust Speaker Verification: A Comparison of Alternative Mel-Scale Representations

Tomi Kinnunen[1], *Md. Jahangir Alam*[2,4], *Pavel Matějka*[3],
Patrick Kenny[2], *Jan “Honza” Černocký*[3], *Douglas O’Shaughnessy*[4]

[1]School of Computing, University of Eastern Finland (UEF), Joensuu, Finland
[2]Centre de Recherche Informatique de Montréal (CRIM), Montréal, Canada
[3]Brno University of Technology (BUT), Brno, Czech Republic
[4]INRS-EMT, Montréal, Canada

tkinnu@cs.uef.fi

Abstract

Accuracy of speaker verification is high under controlled conditions but falls off rapidly in the presence of interfering sounds. This is because spectral features, such as Mel-frequency cepstral coefficients (MFCCs), are sensitive to additive noise. MFCCs are a particular realization of warped-frequency representation with low-frequency focus. But there are several alternative, potentially more robust, warped-frequency representations. We provide an experimental comparison of five warped-frequency features. They use exactly the same frequency warping function, the same number of coefficients and postprocessing, but differ in their internal computations. The compared variants are (1) conventional MFCCs from discrete Fourier transform (DFT), followed by Mel-scaled filterbank, (2) MFCCs via direct warping of DFT, followed by linear-scale filterbank, (3) warped linear prediction features, (4) perceptual minimum variance distortionless features and (5) recently proposed sparse Mel-scale histogram features. Experiments carried out on a subset of the SRE 10 corpus using a scaled-down i-vector system indicate that direct DFT warping outperforms conventional MFCCs in most of the cases.

Index Terms: speaker recognition, noise, frequency warping

1. Introduction

Thanks to mobile technology, we can nowadays communicate with each other in a wide range of environments. Unfortunately, this also sets up challenges in building robust recognizers. Despite tremendous amount of research on the topic, mostly in the context of automatic speech recognition (ASR) [1], recognition in noisy, unforeseen environments remains challenging. While back-end techniques such as integrated speaker and noise models [2] and multicondition training [3, 4] are useful in reducing mismatch across training and run-time utterances, they are either specific to a particular model or require careful selection of additional training utterances representative of the expected operating conditions. We focus on an alternative, potentially more universal solution to reduce mismatch induced by noise - robust acoustic front-ends.

There are a number of ways to increase environmental robustness at the front-end, see [5] for a recent review in speaker recognition context. First, plug-in speech enhancement methods such as spectral subtraction or Wiener filter can be used for increasing signal quality as a pre-processing step [6]. Second, one may enhance the spectral magnitudes either by weighting reliable regions [7, 8] or by even completely discarding uncertain time-frequency bins via *missing feature* techniques [9, 3, 10, 11, 12]. All these techniques require an estimate of either the noise spectrum or spectral reliability values and, consequently, their utility is largely determined by the quality of these estimates. Third, feature normalization techniques [13] can also be useful, when combined with other techniques, but by no means sufficient by themselves.

The fourth class of methods – our focus – is to revise the feature extractor itself. As an example, one may use any robust power spectrum estimator in standard front-ends, such as Mel-frequency cepstral coefficient (MFCC) [14] and perceptual linear prediction (PLP) extractors [15]; see [16] for a recent comparison. Here we focus on a particular facet of feature extraction, *frequency warping*, inspired by psychophysical principles. New types of auditorily-motivated features including cochlear [17] and long-term temporal features [18, 19], sparse coding [17], power normalization [20] and frequency masking [21] have shown promise recently.

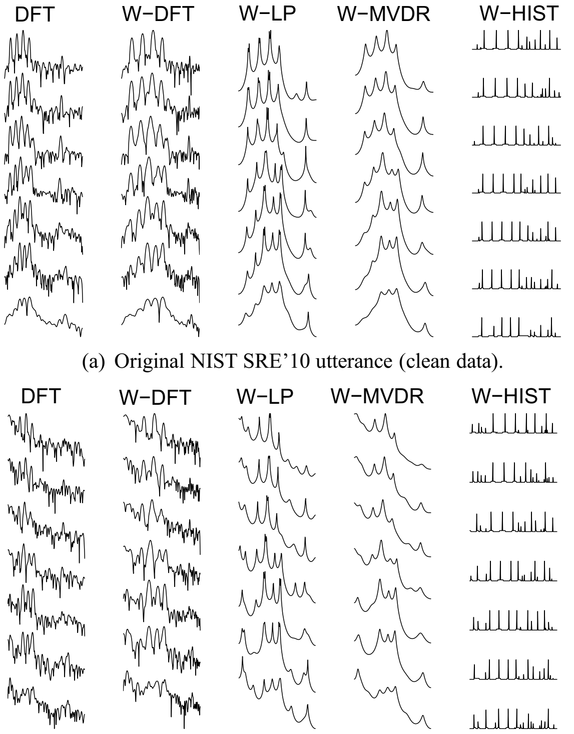
In the context of modern i-vector speaker verification [22], we present a comparison of five feature extractors involving Mel-frequency warping. We apply exactly the same frequency warping, implemented via warped discrete Fourier transform, and compare different parameterizations. The compared methods (Fig. 1) include conventional MFCCs and an alternative implementation via direct warping, two warped all-pole models [23, 24] and a recent histogram-based technique producing sparse features [21]. Even though some of the methods have been studied in both ASR [25] and speaker verification [26, 27], we feel that it is time to present a self-contained summary and comparison within a single study, using a modern i-vector system [22]. We hope the reader finds our study a useful summary of methods otherwise scattered across the literature.

2. Feature extractors

2.1. Conventional MFCCs via DFT

Let $x[n], n = 0, 1, \ldots, N' - 1$ denote the PCM samples of a speech frame of N' samples, including possible trailing zeros so that N' is a power of two. We first compute the windowed discrete Fourier transform (DFT),

$$X[k] = \sum_{n=0}^{N'-1} x[n]w[n]e^{-j\omega_k n}, \quad k = 0, 1, \ldots, N' - 1, \quad (1)$$



(a) Original NIST SRE'10 utterance (clean data).

(b) HVAC-style noise at 6 dB signal-to-noise-ratio.

Figure 1: Running spectra of (a) clean and (b) noisy data. Time runs from bottom to up, frequency axis from left to right. For DFT, the frequency axis is linear in [0,4000] Hz; for the rest, it is linear in [0, 35] *Mels*. **DFT**: discrete Fourier transform, **W-DFT**: warped DFT, **W-LP**: warped linear prediction ($p = 24$); **W-MVDR**: warped minimum variance distortionless response (order $p = 24$), **W-HIST**: warped histogram ($B = 20$).

where $w[n] = \frac{1}{2}\{1 - \cos(2\pi n/N')\}$ is the Hann window and $\omega_k = 2\pi k/N'$ denotes the discrete frequency sampled uniformly on $[0, 2\pi]$. The DFT operation can also be written as $\mathbf{X} = \mathbf{F}\mathbf{x}$, where $\mathbf{X} = (X[0], \ldots, X[N'-1])^{\mathrm{T}}$, $\mathbf{x} = (x[0]w[0], \ldots, w[N'-1]x[N'-1])^{\mathrm{T}}$ and $F_{k,n} = e^{-j\omega_k n}$ is the (k, n)th element of the complex-valued DFT matrix, $\mathbf{F} \in \mathbb{C}^{N' \times N'}$. In practice, DFT is obtained using the fast Fourier transform (FFT). For alternatives to DFT, refer to [16].

The next step is to locally integrate the spectral energy using a frequency-domain filterbank. Due to spectral symmetry, we only need to process the first half of the power spectrum. Let $\mathbf{s} = (|X[0]|^2, \ldots, |X[N]|^2)^{\mathrm{T}}$, where $N = N'/2$, to denote the $N + 1$-point DFT power spectrum and let $\mathbf{H} \in \mathbb{R}^{M \times (N+1)}$ be a filterbank matrix with real and nonnegative elements. It contains M triangular-shaped filter responses in its rows. The MFCC extractor is written as $\mathbf{c} = \mathbf{D} \log(\mathbf{H}\mathbf{s})$, where $\mathbf{c} \in \mathbb{R}^M$ is the MFCC vector, $\log(\cdot)$ is applied element-by-element and $\mathbf{D} \in \mathbb{R}^{M \times M}$ is the orthonormal DCT matrix containing cosine bases on its rows. The MFCC vector is further cut to include the first 12 to 20 elements appended with delta and double delta features.

A key element of any MFCC extractor is the filterbank matrix, $\mathbf{H}$. We follow a classical design (e.g. [28]), whereby the center frequencies of the triangular filters are uniformly placed on the Mel-frequency axis. Let $\mathrm{Mel}(f_{\mathrm{Hz}}) = 2595 \log_{10}(1 + f_{\mathrm{Hz}}/700)$ and $\mathrm{Mel}^{-1}(f_{\mathrm{Mel}}) = 700(10^{f_{\mathrm{Mel}}/2595} - 1)$ denote, respectively, the mapping and its inverse from Hertz to Mel scale. Given the desired number of filters, M, we find the filter centerpoints in Mel-scale as $\mu_m = \frac{m}{M} \times \mathrm{Mel}(f_s/2), m = 1, 2, \ldots, M$ where f_s is the sampling rate, in our case $f_s = 8$ kHz. The uniformly spaced values μ_m are then mapped back to the Hertz domain using $\mathrm{Mel}^{-1}(\cdot)$ and rounded off to the index of the nearest DFT bin. The support region of each filter spans from the center frequencies of the neighboring filters, with shape defined by piecewise linear function. The filters are normalized to have approximately constant energy per each band, implying that the high-frequency filters have smaller peak response[1].

2.2. MFCCs via Direct Warping of DFT (W-DFT)

In all the remaining four methods, we consider an alternative frequency warping strategy by *direct* warping of the spectrum, rather than the filterbank. We start again from the Hann-windowed frame $\mathbf{x}$ but apply a *warped* DFT (WDFT) matrix to obtain a warped spectrum: $\tilde{\mathbf{X}} = \tilde{\mathbf{F}}\mathbf{x}$. Here, $\tilde{\mathbf{X}} \in \mathbb{C}^{N'}$ and $\tilde{\mathbf{F}} \in \mathbb{C}^{N' \times N'}$ are, respectively, the complex-valued warped spectrum and the WDFT matrix. Its elements are given by $F_{k,n} = e^{-j\tilde{\omega}_k n}$, where $\tilde{\omega}_k = \frac{2\pi}{N'}\tilde{k}$, $\tilde{k}$ being uniformly placed on the Mel scale. The direct computation via matrix-vector multiplication of complexity $\mathcal{O}(N'^2)$ is more CPU-intensive than conventional DFT utilizing $\mathcal{O}(N' \log_2 N')$ FFT implementation. One alternative way to achieve direct frequency warping is re-mapping FFT bins following spectral interpolation (e.g. [24]). But as our primary focus is not computational efficiency, we prefer the direct computation due to its direct analogy to conventional DFT. Indeed, as seen in Fig. 1, the warped spectra (W-DFT) have similar features like the conventional DFT spectra, but more frequency bins assigned to the important low-frequency region.

The warped power spectrum, $(|\tilde{X}[0]|^2, \ldots, |\tilde{X}[N]|^2)^{\mathrm{T}}$ with $N = N'/2$, undergoes exactly the same steps as conventional MFCC extraction – filterbank integration, log-compression and DCT. But since the spectrum is already pre-warped, the triangular filters are now uniformly spaced in the Hertz, rather than Mel, scale. They require no energy normalization as the bandwidths (and area) are nearly equal by construction.

2.3. Warped Linear Prediction (W-LP)

Linear prediction (LP) [29] is a classic parametric spectral modeling technique. In the time domain, LP predictor equation is $\hat{x}[n] = -\sum_{m=1}^{p} a_m x[n-m]$, where the nth speech sample is predicted from the past p values, using predictor coefficients a_m. The coefficients are solved by minimizing the residual energy, $\sum_n (x[n] - \hat{x}[n])^2 = \sum_n (x[n] + \sum_{m=1}^{p} a_k x[n-m])^2$. In the *autocorrelation* method of LP, the solution turns out to be $\mathbf{a} = -\mathbf{R}^{-1}\mathbf{r}$, where $\mathbf{R}$ and $\mathbf{r}$ are, respectively, Toeplitz autocorrelation matrix and autocorrelation vector. Given the so-obtained predictor coefficients, $\mathbf{a} = (a_1, a_2, \ldots, a_p)$, the LP spectral envelope sampled at frequency bins $k = 0, 1, \ldots, N'-1$ is obtained as $S_{\mathrm{lp}}[k] = 1/|1 + \sum_{m=1}^{p} a_m e^{-j\omega_k m}|^2$ where $\omega_k = 2\pi k/N'$ as above.

Warped LP (W-LP) modeling dates back to at least [23]; see [30] for a general overview of warped audio signal processing. In our case, the warped LP variant is easily obtained by noting that autocorrelation sequence can be computed as the

[1] We used RASTAmat to implement the filterbank: `http://labrosa.ee.columbia.edu/matlab/rastamat/`

inverse DFT of the squared DFT magnitude spectrum. Thus, we first compute the warped DFT spectrum $\tilde{\mathbf{X}}$ as above, take its squared modulus and use inverse DFT to obtain a warped autocorrelation sequence. The warped LP coefficients, say $b_m, m = 1, \ldots, p$, are then obtained using Levinson-Durbin and converted into a warped LP envelope the same way as above, using b_m's in place of a_m's in $S_{\mathrm{lp}}[k]$. As Fig. 1 indicates, W-LP spectra contain similar features as W-DFT, picking up F0 harmonics and formants. However, due to reduced degrees of freedom in all-pole modeling ($p = 24$ coefficients vs. $N' = 256$ bins), the spectra are generally much smoother. This potentially results in improved noise robustness over W-DFT.

2.4. Warped MVDR (W-MVDR)

Minimum variance distortionless response (MVDR) [31] is another popular all-pole technique with attractive properties. MVDR defines a filter that leaves the signal at frequency of interest undistorted (i.e. with unity gain) while suppressing the other frequencies in an optimal way. Given LP coefficients a_m, the MVDR coefficients, μ_m, are computed via straightforward non-iterative mapping, $\mu_m = \frac{1}{P_e}\sum_{i=0}^{p-m}(p + 1 - m - 2i)a_i a_{i+m}$ for $m = 0, 1, \ldots, p$ and $\mu_m = \mu_{-m}$ for $m = -p, \ldots, -1$. Here, P_e is the predictor error variance. The MVDR spectrum is then obtained from the μ_m coefficients as, $S_{\mathrm{mvdr}}[f] = 1/|(\sum_{m=-p}^{p} \mu_k e^{-j\omega_k m}|^2$ where $\omega_k = 2\pi k/N'$ as before.

To obtain a *warped* MVDR model (see [24]), we proceed the most obvious way: the warped LP coefficients b_k defined in the previous subsection are converted to the MVDR coefficients using the above formula for μ_k, by using b_ks in place of a_ks. These are then substituted to $S_{\mathrm{mvdr}}[f]$ as above. As seen from Fig. 1, the W-MVDR spectra obtained this way are even smoother than W-LP spectra.

2.5. Warped Histogram (W-HIST)

This interesting recent method [21], which initially inspired us to explore frequency warping in detail, uses a radically different approach to compute features. It is roughly based on the idea that the human auditory system can form a critical band around *any* frequency, not just around the arbitrarily-fixed centerpoints of a typical MFCC filterbank. The starting point is the warped DFT spectrum $\tilde{\mathbf{X}} = (|\tilde{X}[0]|^2, \ldots, |\tilde{X}[N]|^2)$ as above. Let

$$k_{\max}(\ell, B) = \ell + \arg\max_{0 \leq p < B} |\tilde{X}[\ell + p]|^2 \qquad (2)$$

be the *index* of a local maximum of warped spectrum in the index range $[\ell, \ell+B-1]$ where B defines the search bandwidth (in units of W-DFT bins). While $\arg\max(\cdot)$ chooses the index of maximum, adding ℓ ensures indexing relative to the original bins of $\tilde{\mathbf{X}}$. We increment ℓ in steps of one bin and count the total number of times each of the spectral bins was activated.

Whenever the warped spectrum contains a prominent spectral peak at a certain bin, it tends to be the $\arg\max(\cdot)$ in multiple overlapping spectral windows. The nonnegative and integer-valued power-spectrum like representation obtained this way favors locally dominant spectral values, such as speech harmonics, while effectively suppressing or zeroing out low-magnitude components. The bandwidth parameter B can be used for controlling spectral sparsity; high values of B produce highly sparse representations that may also remove useful components. As the right-most columns in Fig. 1 indicate, this method indeed picks up the frequency locations of the prominent W-DFT components. They correspond roughly to the location of harmonics in W-DFT spectra.

3. Experimental setup

For the experiments, we adopt female trials (more difficult than males) in the common condition 5 of the NIST 2010 SRE corpus involving normal vocal effort telephone conversations. It consists of 3704 target and 233077 nontarget trials. In addition, we have two simulated noise conditions involving *heating, ventilation and airconditioning* (HVAC) and *crowd* noises at signal-to-noise ratio (SNR) of 6 dB added using FaNT[2]. HVAC noises are taken from FreeSound [3] and crowd noises by summing up hundreds of random conversations. We report both the equal error rate (EER) and the normalized minimum detection cost (MinDCF), which is $\min_\theta\{0.10P_{\mathrm{miss}}(\theta) + 0.99P_{\mathrm{fa}}(\theta)\}$, with the "classical" cost values used[4]. Here, $P_{\mathrm{miss}}(\theta)$ and $P_{\mathrm{fa}}(\theta)$ are the miss and false alarm rates at threshold θ.

We adopt an i-vector recognizer [22] to compare the five feature sets. One gender-independent universal background model with 512 diagonal covariances Gaussians is trained from telephone data in NIST 2004–2005 SRE. One gender-independent 200-dimensional i-vector extractor is trained on the same data. The i-vectors are scaled down to 150 dimensions using linear discriminant analysis (LDA) and normalized to unity length [32]. Full-rank probabilistic LDA (PLDA) is trained on the resulting vectors. Both matrices are gender-independent and estimated on the same data as the i-vector extractor. For details, see [33]. We use the following short-hands for the methods:

DFT: Conventional MFCCs from discrete Fourier transform (DFT), followed by Mel-scaled filterbank, log and DCT.

W-DFT: MFCCs from warped DFT, followed by linear-scale filterbank, log and DCT.

W-LP: Warped linear prediction spectrum from WDFT, followed by linear-scale filterbank, log and DCT.

W-MVDR: Warped MVDR spectrum followed by linear-scale filterbank, log and DCT.

W-HIST: Warped histogram features [21], followed by DCT.

All the five methods lead to a set of base coefficients computed using a 25 ms Hann window every 10 ms. We retain the 19 first coefficients appended further with log-energy. The 20-dimensional feature vector then undergoes short-time mean and variance normalization over a 3-sec sliding window. Deltas and double deltas are estimated from the context of 5 frames to give 60-dimensional features. Speech activity detection uses BUT's Hungarian phoneme recognizer [34] and relative average energy thresholding. Short segments are pruned out and speech segments merged together.

As we do not use multicondition training, any front-end is expected to experience severe degradation under noise. Since standard speech enhancement can be effective in reducing mismatches introduced by noise [14, 6], we include a simple spectral oversubtraction method. Given observed noisy speech power, $|X|^2$ (we omit frame and frequency indices) and estimated noise power, $|\hat{N}|^2$, we form Wiener gain, $g = \max\big(1 - \alpha\frac{|\hat{N}|^2}{|X|^2},\ \min(g_h, \beta\frac{|\hat{N}|^2}{|X|^2})\big)$, where α is an oversubtraction factor,

[2] http://dnt.kr.hsnr.de/download.html

[3] www.freesound.org

[4] http://www.itl.nist.gov/iad/mig//tests/sre/2010/NIST_SRE10_evalplan.r6.pdf

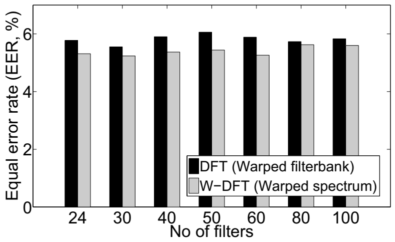


Figure 2: Comparison of conventional MFCCs via warped filterbank (black) and MFCCs via direct spectrum warping and linear-frequency filterbank (gray) on **clean** data (SRE 10 without added noise) for different number of triangular filters.

$g_h = 1.00$ is maximum gain for noise floor and $\beta = 0.01$ determines maximum noise attenuation. Noisy speech magnitude is multiplied with g, combined with noisy phase and reverted to time domain with overlap-and-add. In our implementation[5] the maximum oversubtraction factor is set to 1 and the noise estimate $|\hat{N}|^2$ obtained using MMSE method of [35].

4. Results

We first compare the two variants to compute MFCCs, DFT (conventional MFCCs) and W-DFT (warped spectrum) in Fig. 2 on the original SRE 10 data without added noise, in terms of EER (same trends hold for MinDCF). There is a slight improvement due to direct warping. This could be due to the increased computational precision at the low frequency range. The number of triangular filters does not have much effect, which might be because the number of cepstral coefficients is 19, which limits detail in the corresponding Mel filter outputs. We fix the number of filters for all the methods to $M = 24$.

The two all-pole model variants, W-LP and W-MVDR, require setting the model order p. Speaker verification studies involving non-warped all-pole models [36, 37] suggest useful values from $p = 18$ to $p = 22$ and speech recognition experiments with warped MVDR [24] from $p = 20$ to $p = 24$. We arbitrarily fix $p = 24$ based on these earlier studies.

The histogram method requires setting the spectral bandwidth B for searching the local maxima. Using $N' = 256$ point warped spectrum, we tried values $B = \{10, 15, 20, 25, 30, 40, 50\}$ on clean data. We fix $B = 20$ as this gave the lowest EER on clean data. We also tried including the filterbank and/or use log-compression on the histogram counts, analogous to steps in the four other methods. These attempts lead to increased error rates. Taking simply the DCT of the histogram counts, as suggested in [21], was chosen.

With all the parameters fixed on clean data as described above, the complete results on all data are summarized in Tables 1 (no speech enhancement) and 2 (with speech enhancement). For noisy data without speech enhancement, EERs increase roughly 4- and 6-fold over clean conditions for HVAC and crowd noises, respectively. HVAC appears more challenging noise type. Speech enhancement improves accuracy considerably, though the error rates under noise remain high. The results on clean data also slightly improve, because the initial tunings of spectral subtraction were done on clean data. Comparing DFT and W-DFT, the same ordering as in Fig. 2 holds for noisy data as well. Comparing W-DFT and W-LP, all-pole modeling provides a systematic boost as one might hypothesize from Fig. 1. Interestingly, further variance reduction from W-LP to W-MVDR conversion improves recognition accuracy further. As Fig. 1 suggests, W-MVDR spectra has reduced peakiness over W-LP – a known property of MVDR processing.

Table 1: Comparison of feature extractors without speech enhancement. HVAC: heating, ventilation and air-conditioning.

	EER (%)			MinDCF		
	Clean	HVAC	Crowd	Clean	HVAC	Crowd
DFT	5.79	32.9	21.1	0.25	0.98	0.82
W-DFT	5.31	32.5	20.2	0.24	0.98	0.76
W-LP	5.13	30.7	18.1	**0.22**	0.96	0.75
W-MVDR	**4.82**	**29.3**	**17.3**	**0.22**	**0.95**	**0.70**
W-HIST	6.75	38.0	18.8	0.33	0.99	0.77

Table 2: Same as Table 1 but with speech enhancement.

	EER (%)			MinDCF		
	Clean	HVAC	Crowd	Clean	HVAC	Crowd
DFT	5.38	28.0	19.7	0.24	0.94	0.78
W-DFT	5.03	26.1	18.4	0.23	0.91	0.74
W-LP	**4.90**	25.3	17.2	**0.22**	**0.88**	0.69
W-MVDR	4.94	**24.3**	**15.3**	**0.22**	**0.88**	**0.64**
W-HIST	6.75	30.0	17.6	0.32	0.97	0.74

Unlike in [21], the histogram method does not perform well in our case; it outperforms conventional MFCCs only in EER for crowd noise. While the reasons for the discrepancy should be studied carefully, we note that [21] involved different corpus, spoken language, channel conditions and a classifier without session compensation. It may also be that such features would require different types of feature normalization or classifier.

5. Conclusions

We compared five alternative Mel-frequency warped feature representations using a modern i-vector recognizer. MFCCs computed through directly warped spectrum showed improvement slightly over conventional MFCCs. The two parametric warped all-pole models, on the other hand, outperformed conventional MFCCs. Especially the warped MVDR method appears promising for further studies. The sparse histogram features were not found robust and reasons for this should be studied. Overall, alternative warping variants show some promise, though with increased computation. Immediate future work, therefore, should address speeding up the computations.

Our results were presented for female trials only. A recent study [38] on SRE 2010 found that females may actually benefit from linear-scale, rather than Mel-scale filterbank. This suggest to address both male experiments and alternatives to Mel-warping to find a possible joint effect of the gender and the choice of the warping function.

6. Acknowledgements

The work was initiated during the BOSARIS 2012 workshop held in Brno. It was partly supported by the Academy of Finland (proj. 253120), the Defense Advanced Research Projects Agency (DARPA) Contract No. D10PC20015, Technology Agency of the Czech Republic grant No. TA01011328 and IT4 Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

[5] Taken from Voicebox, `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`

7. References

[1] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE Trans. Audio, Speech and Language Proc. (to appear)*, 2013.

[2] R. Rose, E. Hofstetter, and D. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, April 1994.

[3] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in unknown noisy conditions," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.

[4] D. Garcia-Romero, X. Zhou, and C. Espy-Wilson, "Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. ICASSP 2012*, Kyoto, Japan, March 2012, pp. 4257 – 4260.

[5] R. Togneri and D. Pullella, "An overview of speaker identification: accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011.

[6] S. Sadjadi and J. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. Interspeech 2010*, Makuhari, Japan, September 2010, pp. 2138–2141.

[7] B. Nasersharif and A. Akbari, "SNR-dependent compression of enhanced Mel sub-band energies for compensation of noise effects on MFCC features," *Pattern Recognition Letters*, vol. 28, pp. 1320–1326, 2007.

[8] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Robust feature extraction for speech recognition by enhancing auditory spectrum," in *Proc. Interspeech 2012*, Portland, Oregon, USA, September 2012.

[9] A. Drygajlo and M. El-Muliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Proc. ICASSP 1998*, 1998, pp. 121–124.

[10] D. Pullella, M. K uhne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," in *Proc. ICASSP 2008*, 2008, pp. 4833–4836.

[11] X. Zhao, Y. Shao, and D. Wang, "Robust speaker identification using a CASA front-end," in *Proc. ICASSP 2011*, Prague, Czech Republic, 2011, pp. 5468–5471.

[12] D. Ribas, J. Villalba, E. Lleida, and J. Calvo, "Missing feature techniques combination for speaker recognition in noisy environment," in *FALA 2010: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop*, Vigo, Spain, June-July 2010, pp. 115–118.

[13] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)*, Crete, Greece, June 2001, pp. 213–218.

[14] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, 2010.

[15] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Communication (accepted)*, 2012.

[16] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertas, J. Sandberg, and M. Hansson-Sandsten, "Comparing spectrum estimators in speaker verification under additive noise," in *Proc. ICASSP 2012*, Kyoto, Japan, March 2012, pp. 4769–4772.

[17] D. You, J. Han, T. Zheng, and G. Zheng, "Sparse-based auditory model for robust speaker recognition," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 7, pp. 1–12, 2012.

[18] S. Sadjadi, T. Hasan, and J. Hansen, "Mean hilbert envelope coefficients (MHEC) for robust speaker recognition," in *Proc. Interspeech 2012*, Portland, USA, September 2012.

[19] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-d autoregressive models for speaker recognition," in *Proc. Odyssey 2012: The Speaker and Language Recognition Workshop*, Singapore, June 2012.

[20] E. Ambikairaijah, J. Kua, V. Sethu, and H. Li, "PNCC-ivector-SRC based speaker verification," in *Proc. Asia-Pacific Signal & Inf. Proc. Assoc. Annual Summit and Conf. (APSIPA ASC 2012)*, Hollywood, USA, December 2012, pp. 1–7.

[21] J. Montalvão and M. R. R. Araujo, "Is masking a relevant aspect lacking in MFCC? a speaker verification perspective," *Pattern Recognition Letters*, vol. 33, 2012.

[22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[23] H. W. Strube, "Linear prediction on a warped frequency scale," *Journal of the Acoustic Society of America*, vol. 68, 1980.

[24] U. Yapanel and J. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Communication*, vol. 50, pp. 142–152, 2008.

[25] R. Muralishankar and D. O'Shaughnessy, "A comparative analysis of noise robust speech features extracted from all-pass based warping with MFCC in a noisy phoneme recognition," in *The 3rd Int. Conf. on Digital Telecommunications*, Bangalore, India, June-July 2008, pp. 180–185.

[26] M. Wölfel, Q. Yang, Q. Jin, and T. Schultz, "Speaker identification using warped MVDR cepstral features," in *Proc. Interspeech 2009*, Brighton, UK, September 2009, pp. 912–915.

[27] G. Liu, Y. Lei, and J. Hansen, "Robust feature front-end for speaker identification," in *Proc. ICASSP 2012*, Kyoto, Japan, March 2012, pp. 4233–4236.

[28] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: a Guide to Theory, Algorithm, and System Development.* New Jersey: Prentice-Hall, 2001.

[29] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 561–580, April 1975.

[30] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," in *Proc. 108th AES CONVENTION*, Paris, France, May 2000, pp. 1–42.

[31] M. Murthi and B. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. on Speech and Audio Processing*, vol. 8, 2000.

[32] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech 2011*, Florence, Italy, August 2011, pp. 249–252.

[33] P. Matějka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Černocký, "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Proc. ICASSP 2011*, pp. 4828–4831.

[34] P. Schwarz, P. Matějka, and J. Černocký, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. ICASSP 2006*, Toulouse, France, 2006, pp. 325–328.

[35] T. Gerkmann and R. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech and Language Processing*, vol. 20, pp. 1383–1393, 2012.

[36] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for speaker verification in additive noise," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 40–46.

[37] C. Hanilci, T. Kinnunen, F. Ertas, R. Saeidi, J. Pohjalainen, and P. Alku, "Regularized all-pole models for speaker verification under noisy environments," *IEEE Signal Processing Letters*, vol. 19, no. 3, pp. 163–166, March 2012.

[38] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus mel frequency cepstral coefficients for speaker recognition," in *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU 2011)*, Hawaii, 2011, pp. 559–564.