Comparison of Spectrum Estimators in Speaker Verification: Mismatch Conditions Induced by Vocal Effort

Cemal Hanilçi[1,2], Tomi Kinnunen[2], Padmanabhan Rajan[2], Jouni Pohjalainen[3], Paavo Alku[3], Figen Ertaş[1]

[1]Department of Electronic Engineering, Uludağ University, 16059, Bursa, Turkey
[2]School of Computing, University of Eastern Finland, Joensuu, Finland
[3]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
chanilci@uludag.edu.tr, tkinnu@cs.joensuu.fi, jouni.pohjalainen@aalto.fi, paavo.alku@aalto.fi

Abstract

We study the problem of *vocal effort mismatch* in speaker verification. Changes in speaker's vocal effort induce changes in fundamental frequency (F0) and formant structure which introduce unwanted intra-speaker variations to features. We compare seven alternative spectrum estimators in the context of mel-frequency cepstral coefficient (MFCC) extraction for speaker verification. The compared variants include traditional FFT spectrum and six parametric all-pole models. Experimental results on the NIST 2010 speaker recognition evaluation (SRE) corpus utilizing both GMM-UBM and more recent GMM supervector classifier indicate that spectrum estimation has a considerable impact on speaker verification accuracy under mismatched vocal effort conditions. The highest recognition accuracy was achieved using a particular variant of temporally weighted all-pole model, *stabilized weighted linear prediction* (SWLP).

Index Terms: speaker recognition, vocal effort mismatch, spectrum estimation

1. Introduction

Speaker verification is the task of determining whether a given speech segment is spoken by a claimed speaker [1]. Generally, *mel-frequency cepstral coefficient* (MFCC) features obtained from *discrete Fourier transform* (DFT) magnitude spectrum are used as features to create speaker models. Gaussian mixture model (GMM) [2] and support vector machine (SVM) [3] are two well-known techniques chosen often for this purpose.

Besides the two well-studied problems of channel effects and additive noise, there are also inherent *intra-speaker* variations that cause mismatches to acoustic features of the same speaker. In this study, we focus on combating *vocal effort mismatch* between training and test speech samples. In the presence of background noise, speakers tend to adjust their speech production by increasing vocal effort, a phenomenon known as the *Lombard effect* [4]. For example, in a quiet library environment, speakers may lower their vocal effort to produce whispered speech. High vocal effort causes considerable changes both in time and frequency domain features [4]. In [5], it was reported that fundamental frequency ($F0$) and the first formant ($F1$) are highly correlated with increased vocal effort. In the same study, it was shown that the second and the third formant frequencies ($F2$ and $F3$) do not vary as much but their amplitudes do increase considerably. Generally, high vocal effort causes increase in $F0$, while low vocal effort lowers it. These acoustic changes in F0 and formant parameters reflect as changes in the short-term spectrum, which is the starting point for features used by speaker recognition systems. Thus, vocal effort mismatch between training and test in speaker recognition is expected to result in degraded recognition accuracy. The effect of vocal effort to speaker verification performance was analyzed in the NIST 2010 speaker recognition evaluation (SRE) campaign [6]. Indeed, the general consensus reported by many sites was that the recognition accuracy considerably degrades when speaker models are trained with normal vocal effort but tested with high vocal effort. In [7], it was found that the features extracted from nasal syllables appeared robust to high vocal effort in speaker recognition. In a recent study [8], the effect of vocal effort (whisper, soft, loud, shouted and normal) for speech recognition accuracy was studied. It was found that speech recognition accuracy changes dramatically with changes in speech mode.

Spectrum of high-pitched speech is characterized by a sparse harmonic structure which makes the estimation of the spectral envelope difficult from voices produced in high vocal effort. Several spectrum envelope estimation techniques have been proposed in the literature. *Linear prediction* (LP) method is a well-known spectral envelope estimation technique [9] which models the low-pitched voiced speech well. However, for medium and high-pitched voiced sounds, LP does not provide a reliable estimate of the spectral envelope [10]. Minumum variance distortionless response (MVDR) method [10], also known as *Capon* or *maximum likelihood* method, has been proposed for speech of high $F0$. *Regularized linear prediction* (RLP) [11, 12] has recently been proposed with the same rationale.

In this study, we compare different all-pole model based spectrum estimation methods for robust MFCC feature extraction for speaker verification across varying vocal effort conditions. The standard discrete Fourier transform (DFT) method is compared with baseline LP and its recently proposed temporally weighted extensions, *weighted linear prediction* (WLP) [13] and *stabilized WLP* (SWLP) [14], the MVDR method [10] and the RLP method [11]. The NIST 2010 SRE corpus with GMM-UBM and GMM-supervector classifiers are used in the experiments.

2. Spectrum Estimation

2.1. Methods

Given a Hamming windowed speech frame $\mathbf{s} = [s(0),\ s(1), \ldots,\ s(N-1)]^{\mathrm{T}}$, the most basic form of power spectrum computed by discrete Fourier transform (DFT)

is given by,

$$S_{\mathrm{FFT}}(f) = \left| \sum_{n=0}^{N-1} s(n) e^{-j2\pi nf/N} \right|^2, \quad (1)$$

where $f = \{0,\ 1, \ldots,\ N-1\}$ is the discrete frequency index. Another commonly used spectrum estimation method is based on *linear prediction* (LP) [9]. In LP analysis, it is assumed that a speech sample $s(n)$ can be estimated from its previous p samples, $\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k)$. Here, $s(n)$ is the original speech sample, $\hat{s}(n)$ is the predicted sample and p is the predictor order (time span). Conventional autocorrelation method is generally used to estimate the predictor coefficients, $\{\alpha_k\}_{k=1}^{p}$, by minimizing the energy of the residual, $e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k)$. Optimum coefficients are obtained from,

$$\mathbf{a}_{\mathrm{opt}}^{\mathrm{lp}} = -\mathbf{R}_{\mathrm{lp}}^{-1}\mathbf{r}_{\mathrm{lp}}, \quad (2)$$

where $\mathbf{R}_{\mathrm{lp}}$ is a Toeplitz autocorrelation matrix and $\mathbf{r}_{\mathrm{lp}}$ is an autocorrelation vector. Given the predictor coefficients, a_k, the LP spectrum is obtained by,

$$S_{\mathrm{LP}}(f) = \frac{1}{\left|1 + \sum_{k=1}^{p} a_k e^{-j2\pi fk}\right|^2}. \quad (3)$$

A variant of the standard LP, *temporally weighted linear prediction* (WLP) [15, 13] obtains the optimum prediction coefficients by minimizing the weighted square of the residual, $E = \sum_n e^2(n)\Psi_n = \sum_n (s(n) + \sum_{k=1}^{p} b_k s(n-k))^2 \Psi_n$. Here, Ψ_n is a time-domain weighing function. In this work, we use the short-time energy (STE) as the weighting function, $\Psi_n = \sum_{i=1}^{M} x^2(n-i)$, where M is the length of the STE window. The optimum coefficients, b_k, $k = 1, \ldots, p$, are computed as

$$\mathbf{b}_{\mathrm{opt}}^{\mathrm{wlp}} = -\mathbf{R}_{\mathrm{wlp}}^{-1}\mathbf{r}_{\mathrm{wlp}}, \quad (4)$$

where $\mathbf{R}_{\mathrm{wlp}} = \sum_n \mathbf{s}(n)\mathbf{s}(n)^{\mathrm{T}}\Psi_n$, $\mathbf{r}_{\mathrm{wlp}} = \sum_n s(n)\mathbf{s}(n)\Psi_n$ and $\mathbf{s}(n) = [s(n-1)\ s(n-2)\ \ldots\ s(n-p)]^{\mathrm{T}}$. It can be seen that $\mathbf{R}_{\mathrm{wlp}} = \mathbf{R}_{\mathrm{LP}}$ if and only if $\Psi_n = 1$ for all n. Standard LP method guarantees that resulting all-pole filter is stable (poles are inside the unit circle). However, such guarantee does not hold for WLP. Thus, stabilized WLP (SWLP) was proposed in [14]. In SWLP, the weighted autocorrelation matrix and the weighted autocorrelation vector are expressed as $\mathbf{R}_{\mathrm{swlp}} = \mathbf{Y}^T\mathbf{Y}$ and $\mathbf{r}_{\mathrm{swlp}} = \mathbf{Y}^T\mathbf{y}_0$, respectively (the original article [14] presents the problem in a slightly different form). The columns of the matrix $\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2\ \ldots\ \mathbf{y}_p]$ are calculated by $\mathbf{y}_{k+1} = \mathbf{B}\mathbf{y}_k$ for $0 \le k \le p-1$, where $\mathbf{y}_0 = [\sqrt{\Psi_1}x(1)\ \ldots\ \sqrt{\Psi_N}x(N)\ 0\ \ldots\ 0]^T$ and $\mathbf{B}$ is a matrix where all the elements are zero outside the subdiagonal and the elements of the subdiagonal, for $1 \le i \le N+p-1$, are

$$\mathbf{B}_{\mathrm{i+1,i}} = \begin{cases} \sqrt{\Psi_{i+1}/\Psi_i}, & \Psi_i \le \Psi_{i+1} \\ 1, & \Psi_i > \Psi_{i+1}. \end{cases} \quad (5)$$

In regularized LP (RLP) [11, 12], a penalty measure is introduced in the cost function and optimum predictor coefficients are computed by minimizing the new cost function, $\sum_n \left(s(n) + \sum_{k=1}^{p} c_k s(n-k)\right)^2 + \lambda\phi(\mathbf{c})$, where $\phi(\mathbf{c})$ is the penalty measure which is a function of the predictor coefficients $\mathbf{c}$ and λ is a regularization factor which controls the smoothness of the spectrum. In [11], the penalty function was chosen as,

$$\phi(\mathbf{c}) = \mathbf{c}^{\mathrm{T}}\mathbf{DFDc} \quad (6)$$

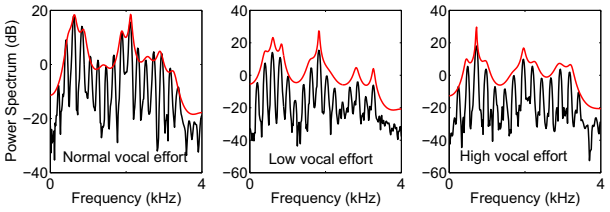

Figure 1: *FFT and LP spectra of the vowel* /I/ *from the same female speaker in different vocal efforts.*

where $\mathbf{D}$ is a diagonal matrix in which each diagonal element is the corresponding row or column number and $\mathbf{F}$ is a matrix of windowed autocorrelation sequence $f(m) = r(m)v(m)$ with Toeplitz form, representing the coarse approximation of the spectral envelope. Here, $r(m)$ is the conventional autocorrelation sequence $r(m) = \sum_n s(n)s(n-m)$ and $v(m)$ is a window function (Boxcar window is used in this study similar to [11]). Optimum prediction coefficients are then computed by,

$$\mathbf{c}_{\mathrm{opt}}^{\mathrm{rlp}} = -(\mathbf{R}_{\mathrm{lp}} + \lambda\mathbf{DFD})^{-1}\mathbf{r}_{\mathrm{lp}}. \quad (7)$$

In [16, 17] the present authors proposed using double autocorrelation (DAC) sequence, $f(k) = \sum_m r(m)r(m-k)$, to compute matrix $\mathbf{F}$ under additive noise. Since noise and speech signals are uncorrelated, the DAC sequence helps to decompose them in the autocorrelation domain [18]. Thus the proposed method improved the recognition performance considerably.

The MVDR spectrum estimation method [10] models the unvoiced or mixed speech spectra by using the LP coefficients. An m^{th} order MVDR spectrum is computed by,

$$S_{\mathrm{MVDR}}(f) = \frac{1}{|\sum_{k=-m}^{m} \mu(k) e^{-j2\pi fk}|^2}, \quad (8)$$

where m is the MVDR filter order and the parameters $\mu(k)$ are computed by a simple non-iterative method from the LP coefficients [10] as follows:

$$\mu(k) = \begin{cases} \sum_{i=0}^{m-k} (m+1-k-2i)a_i a_{i+k}, & k = 0, 1, ..., m \\ \mu(-k), & k = -m,\ \ldots,\ -1, \end{cases} \quad (9)$$

where a_i is the i^{th} LP coefficient.

2.2. Effect of Vocal Effort on Speech Spectrum

It is known that $F0$, shape of the glottal waveform, formant locations and their bandwidths are all affected by changes in vocal effort [5, 4]. To exemplify, Fig.1 shows the DFT and LP spectra of the vowel */I/* in the utterance "*I mean*" spoken by the same female speaker in the NIST 2010 SRE corpus. We can see that the shape of the spectrum radically changes with vocal effort. In particular, sparse harmonic peaks appear in the spectrum produced using high vocal effort.

Table 1 shows the average $F0$ and the first three formants ($F1$-$F3$) and their bandwidths for 2 female speakers in the NIST 2010 SRE corpus. The first speaker produces the utterance "*I mean*" and the second speaker produces "*yeah*" in three different vocal effort condition. $F0$, formants and their bandwidths were computed using *Praat* software[1]. As seen, the average formant frequencies and bandwidths change with vocal effort. In the case of high vocal effort, $F0$ is larger than normal

[1] `http://www.praat.org/`

Table 1: Average $F0$ and first three formant frequencies with their bandwidths (in Hz) for different vocal effort (NVE: normal vocal effort, LVE: low vocal effort and HVE: high vocal effort).

	Speaker 1			Speaker 2		
	NVE	LVE	HVE	NVE	LVE	HVE
$F0$	206	204	241	150	134	159
$F1$	513	582	618	532	567	680
$F2$	1528	1688	1802	1628	1730	1680
$F3$	2509	2332	2419	2215	2314	2342
$BW1$	52	72	71	69	81	75
$BW2$	439	357	568	382	173	111
$BW3$	162	401	308	937	575	301

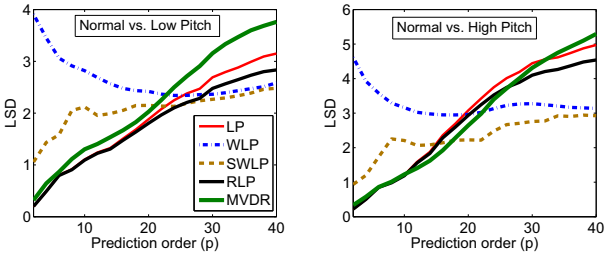

Figure 2: *LSD between original speech and pitch modified speech ($F0$ decreased (left) and increased (right) by 50 %).*

effort for both speakers, as expected. From speaker recognition point of view, these acoustic differences cause intra-speaker variation between training and test whenever there is vocal effort difference. Even though it is not the only acoustic parameter that changes with vocal effort, F0 is one of the most prominent ones. Before proceeding to the speaker recognition experiments with large-scale NIST data, we first analyze spectral distortions implied by changes in F0 changes, under a controlled set-up involving artificial software-based F0 shifting. To this end, Fig. 2 reports average log-spectral distortion (LSD) between normal vocal effort speech and artificial pitch-modified versions (50% increase and decrease) of the same utterance, as a function of prediction order, p, for different all-pole models described in the previous section. $F0$ modification was implemented by *Praat*. The average LSD between two all-pole spectra is defined as,

$$LSD = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{f=0}^{N-1}\left[10\log_{10}S_n^t(f) - 10\log_{10}\hat{S}^t(f)\right]^2}, \quad (10)$$

where $S_n^t(f)$ and $\hat{S}^t(f)$ are the power spectra of the tth frame of the original speech signal and modified signal, respectively. T is the total number of frames. The LSD of the WLP method is a decreasing function of p different from other methods. However, SWLP yields the smallest LSD value for high prediction orders. In general, LSD values between normal and high pitch speech samples are larger than the value between normal and low pitch speech.

3. Experimental Setup

Experiments are carried out on the core task of NIST 2010 SRE corpora, including three different vocal effort sub-conditions:

- **Det 5:** Conversational telephone speech with **normal vocal effort (NVE)** in both training and test, containing 708 target and 29655 impostor trials.

Table 2: EERs (%) for different spectrum estimators and sub-conditions with the GMM-UBM system.

		Equal Error Rate						
	Trials	FFT	LP	WLP	SWLP	RLP	RLP-DAC	MVDR
Det 5	Male	**14.16**	16.51	17.28	15.00	15.59	14.16	15.01
	Female	17.46	17.75	19.15	**15.90**	17.74	16.33	18.58
	All	15.86	17.23	18.36	15.35	16.80	**15.25**	16.66
Det 6	Male	23.85	23.03	24.16	23.03	24.22	22.76	**21.34**
	Female	32.78	27.06	27.58	**23.49**	25.55	26.64	24.04
	All	27.70	25.48	26.03	23.26	24.93	24.65	**22.53**
Det 8	Male	**10.62**	13.44	14.28	11.76	13.44	10.64	15.12
	Female	**12.58**	15.08	13.96	13.40	14.88	12.89	16.23
	All	**11.74**	14.53	13.75	13.08	14.15	12.53	16.10

Table 3: EERs (%) for different spectrum estimators and sub-conditions with the GMM-supervector system.

		Equal Error Rate						
	Trials	FFT	LP	WLP	SWLP	RLP	RLP-DAC	MVDR
Det 5	Male	6.23	5.94	6.13	**5.42**	6.55	8.34	6.78
	Female	8.13	8.18	8.16	**6.47**	7.88	10.12	9.58
	All	7.34	7.06	7.06	**6.00**	7.06	9.32	8.46
Det 6	Male	**7.86**	8.96	8.42	8.42	8.78	11.98	9.57
	Female	12.58	13.56	14.75	**12.56**	14.75	18.05	14.75
	All	10.86	11.91	12.01	**10.80**	12.33	15.85	12.74
Det 8	Male	3.61	4.22	5.04	**2.76**	3.62	6.72	4.20
	Female	6.70	7.62	6.47	**4.74**	6.70	8.37	7.26
	All	5.03	6.28	5.70	**4.02**	5.85	8.05	5.86

- **Det 6:** Conversational telephone speech with **normal vocal effort** condition in training and **high vocal effort (HVE)** telephone speech in test, containing 361 target and 28311 impostor trials.

- **Det 8: Normal vocal effort** telephone speech in training and **low vocal effort (LVE)** telephone speech in test, containing 289 target and 28306 impostor trials.

Two different classifiers are chosen for the experiments. First, we have used simple GMM-UBM system with 128 Gaussian components. GMM-UBM was used because it enables optimizing the control parameters of each spectrum estimation method using the classifier's fast scoring capability without the need for other hyperparameters except the universal background model (UBM). Gender-dependent UBMs are trained using SRE04, SRE05, SRE06 and Switchboard corpuses. Second, we used GMM-supervector classifier [3] with nuisance attribute projection (NAP) channel compensation [19]. In GMM-supervector classifier, gender-dependent UBMs with 512 Gaussians are trained using SRE05, SRE06 and Switchboard databases. Negative examples (background speakers) to train speaker-dependent SVM are selected from SRE03 and SRE04 corpora (395 and 577 speech files for male and female genders, respectively). NAP matrices are trained using 2020 male and 2017 female utterances from the NIST SRE06 corpus. Relevance factor of $r = 8$ is used for adapting the mean vectors.

MFCC features are extracted from 30 ms Hamming windowed frames with 15 ms overlap. To compute the magnitude spectrum of windowed frames, different spectrum estimation methods are considered. Besides standard FFT and LP methods, WLP, SWLP, RLP, RLP-DAC and the MVDR methods are used to compute the spectrum. 18 MFCCs are extracted by multiplying the spectrum with a bank of 27 triangular mel-scale filters. Δ and Δ^2 features are then appended to RASTA filtered [20] MFCCs. Finally, cepstral mean and variance normalization (CMVN) and energy-based voice activity detection (VAD) [21] are applied to the features.

We have used equal error rate (EER) as the performance criterion. EER is the threshold value at which false alarm rate (P_{fa}) and miss detection rate (P_{miss}) are equal. Besides from EER values, detection error trade-off (DET) curves of selected methods are also shown.

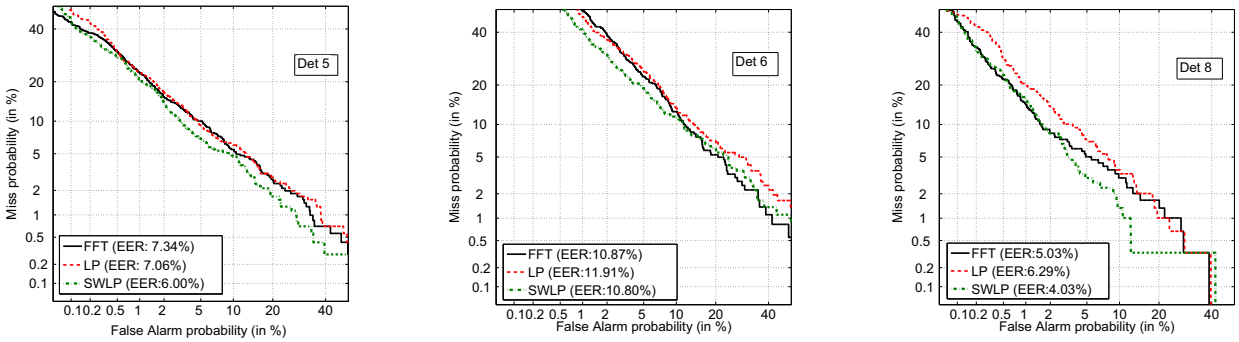

Figure 3: DET curves for Det 5 (NVE-NVE), Det 6 (NVE-HVE) and Det 8 (NVE-LVE) sub-conditions using selected spectrum estimation methods using GMM-supervector classifier.

4. Speaker Verification Results

The spectrum estimators are first compared using the baseline GMM-UBM recognizer. Prediction order is set to $p = 20$ for the all-pole methods. STE windowing with $M = 20$ was used in WLP [15] and SWLP [14]. Regularization parameters $\lambda = 10^{-4}$ and $\lambda = 10^{-9}$ are used in standard RLP [11] and RLP-DAC [16, 17] spectrum estimators. Table 2 summarizes the recognition accuracies of the GMM-UBM classifier for male, female and all trials separately. The minimum EER of each row are highlighted. From these preliminary results we find that:

- Female speakers systematically produce higher EERs independent of the vocal effort condition.
- In Det 5 condition, FFT and RLP-DAC methods yield the smallest EER for male trials.
- For female trials in Det 5, SWLP gives the highest accuracy (15.9 % EER) which corresponds to approximately 9 % relative improvement in EER over FFT (17.46 % EER).
- In Det 6 condition, SWLP shows considerable improvement over FFT for female trials in terms of EER (EER reduced from 32.78 % to 23.49 %, a relative improvement of 28 %). MVDR yields slightly higher EER than SWLP for female speakers. However, it produces the smallest EERs for male trials.
- The performance on the Det 8 condition always gives smaller EER values than Det 5 and Det 6 conditions. Similar observations have been made in [6, 7].

Next we compare the effect of spectrum estimation with the GMM-supervector classifier. The results of GMM-supervector are given in Table 3. Differently from the GMM-UBM results (Table 2), SWLP yields the highest recognition accuracy in comparison to the other methods, irrespective of the vocal effort condition (for Det 6, male speakers with FFT is slightly better). A potential explanation is that SWLP, being a stabilized temporally weighted all-pole model, is capable of computing smooth spectral envelopes in which modelling of formants is less biased by sparse harmonics of high vocal effort speech. In contrast to the GMM-UBM results, the performance improvement obtained with SWLP is larger when low vocal effort is used in test (Det 8). Interestingly, RLP-DAC method gives the highest EER values with GMM-supervector. DET curves for FFT, LP and SWLP methods for Det 5, Det 6 and Det 8 sub-conditions are given in Figure 3. It can be seen that, the performance difference between SWLP and standard methods are larger at low miss rates for Det 5 and Det 8 conditions. However, in Det 6 condition, FFT is better at low miss rates but SWLP at low false alarm rates. The most interesting observation from the experiments is that, Det 8 sub-condition (normal vocal effort training and low vocal effort test) yields the smallest EER than Det 5 and Det 6 conditions independent of the spectrum estimation method used for both the GMM-UBM and GMM-supervector recognizers.

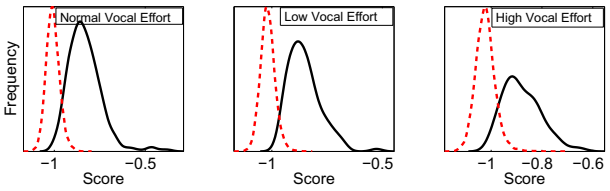

Figure 4: *Target (black solid line) and impostor score (dashed red line) distributions for SWLP method on male trials.*

Figure 4 shows the recognition score distributions for different vocal effort conditions (Det 5, Det 6 and Det 8) for male speakers using SWLP method. In case of high vocal effort (Det 6), the overlap of the target scores within the impostor score distribution is larger than the case of low or normal vocal effort. This is expected because from the results Det 6 sub-condition gives the highest recognition accuracy.

5. Conclusions

We compared different spectrum estimators for MFCC feature extraction in the context of vocal effort mismatch in speaker recognition. From the experimental results conducted on NIST 2010 SRE corpus with GMM-UBM and GMM-supervector classifiers, we found that change in vocal effort affects the recognition performance. With normal vocal effort in training and high vocal effort test (Det 6 sub-condition), the recognition accuracy degraded dramatically. Interestingly, the best recognition accuracy was achieved when low vocal effort was used in test. In general, spectrum estimation has a considerable impact on the speaker recognition performance with different vocal effort conditions. The SWLP method showed the best recognition accuracy in comparison to the remaining six methods independent of the vocal effort condition.

6. Acknowledgements

The work was supported by Academy of Finland (proj. no. 253120)

7. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Comm.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[2] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Dig. Sig. Proc.*, vol. 10, no. 1, pp. 19–41, Jan. 2000.

[3] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Sig. Proc. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[4] H. Boril and J. H. L. Hansen, "Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 18, no. 6, pp. 1379–1393, 2010.

[5] J. S. Liénard and M. G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *Journal of the Acoustical Society of America*, vol. 106, no. 1, pp. 411–422, July 1999.

[6] C. Greenberg, A. Martin, B. Barr, and G. Doddington, "Report on performance results in the NIST 2010 speaker recognition evaluation," in *Proc. Interspeech 2011*, 2011, pp. 261–264.

[7] N. Scheffer, L. Ferrer, M. Gracinarena, S. Kajarekar, E. Shriberg, and A. Stolcke, "The SRI NIST 2010 speaker recognition evaluation system," in *Proc. ICASSP 2011*, 2011, pp. 5292–5295.

[8] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Commun.*, vol. 54, no. 6, pp. 732–742, 2012.

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

[10] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Proc.*, vol. 8, no. 3, pp. 221–239, May 2000.

[11] L. A. Ekman, W. B. Kleijn, and M. N. Murthi, "Regularized linear prediction of speech," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 16, no. 1, pp. 65–73, Jan. 2008.

[12] M. N. Murthi and W. B. Kleijn, "Regularized linear prediction all-pole models," in *IEEE Speech Coding Workshop*, 2000, pp. 96–98.

[13] C. Ma, Y. Kamp, and L. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Comm.*, vol. 12, no. 1, pp. 69–81, March 1993.

[14] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilized weighted linear prediction," *Speech Comm.*, vol. 51, no. 5, pp. 401–411, April 2009.

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

[16] C. Hanilçi, T. Kinnunen, F. Ertaş, R. Saeidi, J. Pohjalainen, and P. Alku, "Regularized all-pole models for speaker verification under noisy environments," *IEEE Sig. Proc. Lett.*, vol. 19, no. 3, pp. 163–166, March 2012.

[17] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertas, "Regularization of all-pole models for speaker verification under additive noise," in *Odyssey, The Speaker and Language Recognition Workshop*, 2012.

[18] T. Shimamura and N. D. Nguyen, "Autocorrelation and double autocorrelation based spectral representations for a noisy word recognition systems," in *Interspeech*, 2010, pp. 1712–1715.

[19] W. C. A. Solomonof and C. Quillen, "Channel compensation for SVM speaker recognition," in *Proc. Speaker Odyssey*, 2004, pp. 57–62.

[20] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Proc.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[21] T. Kinnunen, J. Saastamoinen, V. Hautamäki, M. Vinni, and P. Fränti, "Comparative evaluation of maximum a posteriori vector quantization and gaussian mixture models in speaker verification," *Pattern Recognition Letters*, vol. 30, no. 4, pp. 341–347, 2009.