NORMAL-TO-SHOUTED SPEECH SPECTRAL MAPPING FOR SPEAKER RECOGNITION UNDER VOCAL EFFORT MISMATCH

Ana Ramírez López, Rahim Saeidi, Lauri Juvela, Paavo Alku

Department of Signal Processing and Acoustics
School of Electrical Engineering, Aalto University Finland

ABSTRACT

Speaker recognition performance degrades substantially in case of vocal effort mismatch (e.g. shouted vs. normal speech) between test and enrollment utterances. Such a mismatch is often encountered, for example, in forensic speaker recognition. This paper introduces a novel spectral mapping method which, when employed jointly with a statistical mapping technique, converts the Mel-frequency band energies of normal speech towards their counterparts in shouted speech. The aim is to obtain more robust performance in speaker recognition by tackling vocal effort mismatch between enrollment and test utterances. The processing is performed on the speech signal before feature extraction. The proposed approach was evaluated by testing the performance of a state-of-the-art i-vector-based speaker recognition system with and without applying the spectral mapping processing to the enrollment data. The results show that pre-processing with the proposed approach results in considerable improvement in correct identification rates.

Index Terms— speaker recognition, vocal effort mismatch, shouted speech, spectral mapping

1. INTRODUCTION

Variability in speech recordings of the same speaker may result from extrinsic factors such as the transmission channel or the acoustic environment, or intrinsic factors of the speaker such as age or vocal effort. In speaker recognition, each of these variabilities poses a challenge to the recognition task, and many efforts have been undertaken to cope with them [1]. Research on vocal effort variability was already conducted in the late 1980s by Hansen in his studies on speech under stress [2]. Most of the research to date, however, has focused on extrinsic factors rather than the intrinsic ones. Even though vocal effort mismatch has been shown to considerably affect the recognition performance [3], this challenging condition has not been studied in speaker recognition as actively as, for example, the effect of background noise. Studies conducted in the topic have generally approached the problem by focusing on feature extraction [4, 5, 6, 7, 8], statistical modeling [9, 10] or score calibration [11]. In the feature domain approaches, the solutions proposed can be categorized into three scenarios. The first one corresponds to processing both soft and loud speech such that a middle point is determined in which the difference in acoustic features between the two vocal effort categories is reduced [4, 5, 6]. The second solution corresponds to processing loud speech such that its features match better those of soft speech [7, 8]. Finally, the last category involves processing soft speech to approach loud speech [4, 8]. To our knowledge, there are no previous speaker recognition studies involving normal and shouted speech that belong to the last category.

In this study, we focus on the specific case of shouted-normal speech mismatch, that is, the speaker recognition system uses speech of normal loudness in enrollment and faces shouted speech in test. By normal speech, we refer to speech uttered in normal loudness, such as in ordinary telephone conversations. Shouted speech, on the other hand, refers to producing a very loud acoustical signal in order to enable communication above noise (the Lombard effect [12, 13]) or over distance, or to communicate something urgently. Shouted-normal speech mismatch situations are often encountered in forensic speaker recognition, when the speech recordings under study originate from a speaker that might be in an excited or stressed state, and the recordings are tested with a standard speaker recognition system trained on normal speech. In the present study, we propose to map the short-term spectra of normal speech samples employed during enrollment in an i-vector-based [14], text-independent speaker recognition framework so that they approximate the corresponding spectra of shouted speech. With this processing, we intend to improve the speaker recognition performance by decreasing the spectral mismatch between the enrollment and test samples. The mapping is conducted using a novel perceptually motivated processing method called perceptual spectral matching (PSM), which is employed in conjunction with a GMM-based statistical technique.

2. NORMAL-TO-SHOUTED SPEECH SPECTRAL MAPPING

Increasing the vocal effort changes many acoustical properties of speech. In the spectral domain, for example, raising the vocal effort results in an increase of the fundamental frequency and the first formant [15], as well as in flattening of the spectral tilt [16]. Common short-term spectral features of speaker recognition, like Mel-frequency cepstral coefficients (MFCCs), are thus directly affected by the increased vocal effort which in turn affects the recognition performance.

2.1. Perceptual spectral matching

Perceptual spectral matching (PSM) was proposed in [17] as a novel approach for spectral matching in a perceptual scale between synthetic and natural speech. Motivated by the PSM design in [17], in the current study, given a frame of normal speech $s_{norm}(n)$ and the corresponding shouted speech frame $s_{shout}(n)$, we aim to find a mapping all-pole filter, with the impulse response $h(n)$, such that the convolution $s_{norm}(n) * h(n)$ matches the Mel spectral band energies of $s_{shout}(n)$. More specifically, the mapping filter is estimated by minimizing the distance between the Mel-scale filter bank energies (MFBEs) of the two signals. We indicate the Mel-warped power spectrum of $h(n)$ by $|H(\tilde{\Omega})|^2$ with $\tilde{\Omega}$ as the index of the warped FFT

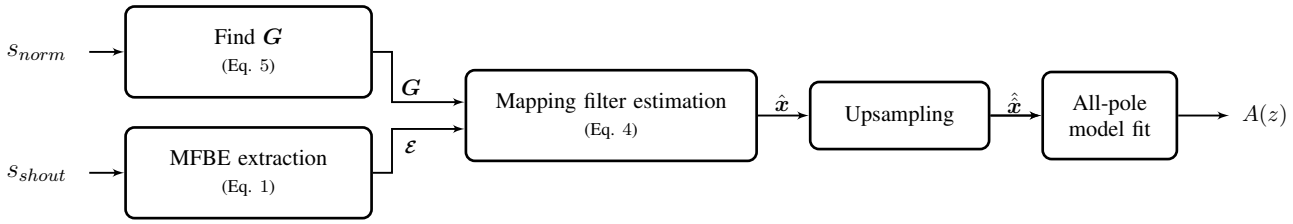

Fig. 1: Flowchart of steps in perceptual spectral matching (PSM) technique. MFBE stands for Mel-scale filter bank energy. $H(z) = 1/A(z)$ is the final all-pole mapping filter, of order p.

bins [18]. A schematic block diagram of the PSM stages is presented in Figure 1.

We first pass the Mel-warped power spectrum of $s_{shout}(n)$, denoted as $|S_{shout}(\tilde{\Omega})|^2$, through a uniform-scale, triangular filter bank $\boldsymbol{T} = [\boldsymbol{t}_0^T, \boldsymbol{t}_1^T, \cdots, \boldsymbol{t}_{M-1}^T]^T$ with M filters $\boldsymbol{t}_i$ ($0 \leq i \leq M-1$) of equal passband width and 50% overlap. The $M \times 1$ vector denoted by $\boldsymbol{\varepsilon}$ consists of the MFBE values of shouted speech, and it is obtained as

$$\boldsymbol{\varepsilon} = \boldsymbol{T}|S_{shout}(\tilde{\Omega})|^2 \tag{1}$$

Mel-warping of the power spectrum is achieved by performing Mel-scale-based spectral interpolation of the FFT power spectrum bins [18]. In the following, all the power spectrum references are defined in the frequency-warped domain.

In general, there is no unique inverse transformation from the MFBE vector associated with the mapping filter's power spectrum to the corresponding full-length spectrum (with N FFT points) due to the dimensionality reduction caused by computing the filter bank energies. However, a unique solution can be achieved by assuming the power spectrum of the mapping filter to be piecewise constant, with $M \times 1$ vector $\boldsymbol{x}$ holding the values of the segments of the power spectrum (see Figure 2). We dub $\boldsymbol{x}$ as the *elementary power spectrum*. With this simplification, the computation of the mapping filter can be derived as follows.

By upsampling $\boldsymbol{x}$ to full-length spectrum, as shown in Figure 2, the ith segment would have a constant value (x_i) in the region where the spectral amplitude of the ith triangular filter $\boldsymbol{t_i}$ is larger than that of its neighbouring filters $\boldsymbol{t_{i-1}}$ and $\boldsymbol{t_{i+1}}$. This construction of $\boldsymbol{x}$ partitions each filter $\boldsymbol{t_i}$ in three different regions: (1) the central region K_i^c, where x_i contributes to the filter output, (2) the lower region K_i^l, where the contributing segment is x_{i-1}, and (3) the higher region K_i^h, where x_{i+1} contributes to the filter output. Given this configuration, we can now express the output of the filter bank, the MFBE vector $\hat{\boldsymbol{\varepsilon}}$ that is mapped from normal speech to shouted speech (from now on, the resulting signal is referred to as shout-like speech). By denoting the warped FFT power spectrum of normal speech as $|S_{norm}(\tilde{\Omega})|^2$, the ith element of the MFBE vector

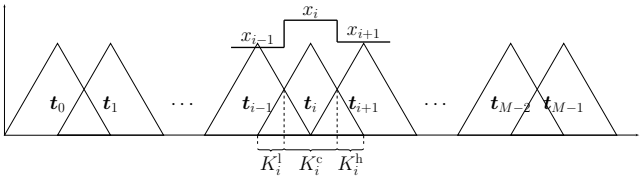

Fig. 2: Piecewise constant upsampling of $\boldsymbol{x}$, the elementary power spectrum of the mapping filter. Note that the triangular filter banks are of equal width due to Mel-warping of the input.

$\hat{\boldsymbol{\varepsilon}}$ of the shout-like speech frame can be written as:

$$\hat{\varepsilon}_i = G_i^l x_{i-1} + G_i^c x_i + G_i^h x_{i+1}, \tag{2}$$

where G_i^l, G_i^c and G_i^h are the dot products of $\boldsymbol{t_i}$ and $|S_{norm}(\tilde{\Omega})|^2$ computed over region K_i^r, with $r = \{l, c, h\}$. Writing the output of all the filters in matrix form, we obtain

$$\underbrace{\begin{bmatrix} G_0^c & G_0^h & 0 & \dots & & 0 \\ G_1^l & G_1^c & G_1^h & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ & & & G_{M-2}^l & G_{M-2}^c & G_{M-2}^h \\ 0 & \dots & & 0 & G_{M-1}^l & G_{M-1}^c \end{bmatrix}}_{\boldsymbol{G}} \underbrace{\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{M-2} \\ x_{M-1} \end{bmatrix}}_{\boldsymbol{x}} = \underbrace{\begin{bmatrix} \hat{\boldsymbol{\varepsilon}}_0 \\ \hat{\boldsymbol{\varepsilon}}_1 \\ \vdots \\ \hat{\boldsymbol{\varepsilon}}_{M-2} \\ \hat{\boldsymbol{\varepsilon}}_{M-1} \end{bmatrix}}_{\hat{\boldsymbol{\varepsilon}}} \tag{3}$$

In order to find $\boldsymbol{x}$, one can use the shouted speech energies $\boldsymbol{\varepsilon}$ as $\boldsymbol{Gx} = \boldsymbol{\varepsilon}$. But the direct solution in the form of $\boldsymbol{x} = \boldsymbol{G}^{-1}\boldsymbol{\varepsilon}$ might return negative values for the elementary power spectrum $\boldsymbol{x}$. Instead of the direct solution, the mapping filter can be formulated as a non-negative least square (NNLS) problem [19], where the objective function to be minimized is $(\boldsymbol{Gx} - \boldsymbol{\varepsilon})^2$, given the element-wise non-negativity constraint $x_i \geq 0$. That is:

$$\hat{\boldsymbol{x}} = \arg\min_{x_i \geq 0} \{(\boldsymbol{Gx} - \boldsymbol{\varepsilon})^2\}. \tag{4}$$

This is solved with the classical NNLS algorithm.

We upsample the resulting $\hat{\boldsymbol{x}}$ in order to obtain an N-point representation in the form of $\hat{\hat{\boldsymbol{x}}}$, before fitting an auto-regressive model, $H(z)$, with $|H(\tilde{\Omega})|^2 = \hat{\hat{\boldsymbol{x}}}$. The elements of $\hat{\boldsymbol{x}}$ can be interpreted as samples of the full-length, N-point power spectrum taken at the triangular filter centers. Then the piecewise constant assumption corresponds to doing nearest value (0th order) interpolation between these samples. In this study, we employed 1st order interpolation in order to obtain piecewise linear spectrum. This is obtained by solving Eq. 4 with

$$\boldsymbol{G} = \boldsymbol{TDT}^T, \tag{5}$$

where $\boldsymbol{D}$ is a diagonal matrix with the vector elements of warped normal speech power spectrum $|S_{norm}(\tilde{\Omega})|^2$ in its diagonal, and upsampling as $\hat{\hat{\boldsymbol{x}}} = \boldsymbol{T}^T\hat{\boldsymbol{x}}$ [17]. The autocorrelation of $\hat{\hat{\boldsymbol{x}}}$ is computed by the inverse Fourier transform, and the obtained autocorrelation is fed into the Levinson-Durbin recursion [20, 21] to obtain the final warped all-pole mapping filter $H(z) = 1/A(z)$, with inverse z-transform $h(n)$.

2.2. Statistical mapping

We employ statistical mapping in the processing chain to achieve automatic spectral mapping of normal speech to its estimated shouted version. That is, we predict the mapping filter corresponding to a given normal speech frame using a trained statistical model. Spectral

mapping from normal to shouted speech using PSM requires aligned normal and shouted speech frames in the training stage.

The statistical dependencies between feature vectors of normal speech ($\boldsymbol{z}$) and feature vectors of the mapping filter ($\boldsymbol{y}$) are modelled using a Gaussian mixture model (GMM):

$$p(\boldsymbol{z}, \boldsymbol{y}) = \sum_{j=1}^{J} w_j \mathcal{N}\left(\begin{bmatrix} \boldsymbol{z} \\ \boldsymbol{y} \end{bmatrix} \middle| \begin{bmatrix} \boldsymbol{\mu}_j^z \\ \boldsymbol{\mu}_j^y \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_j^{zz} & \boldsymbol{\Sigma}_j^{zy} \\ \boldsymbol{\Sigma}_j^{yz} & \boldsymbol{\Sigma}_j^{yy} \end{bmatrix} \right), \quad (6)$$

where w_j is the weight of the jth mixture component; and mean vectors and covariance matrices are denoted as $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$, respectively [22, 23]. Once the GMM is trained, the minimum mean square error estimate of feature $\boldsymbol{y}$ given $\boldsymbol{z}$ is computed as:

$$\hat{\boldsymbol{y}} = \sum_{j=1}^{J} P(j|\boldsymbol{z}) \left[\boldsymbol{\mu}_j^y + \boldsymbol{B}_j (\boldsymbol{z} - \boldsymbol{\mu}_j^z) \right], \quad (7)$$

where the posterior component probabilities $P(j|\boldsymbol{z})$ and linear transformations $\boldsymbol{B}_j = (\boldsymbol{\Sigma}_j^{yz})^{-1} \boldsymbol{\Sigma}_j^{zz}$ are computed from prior component probabilities w_j and likelihoods $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_j^z, \boldsymbol{\Sigma}_j^{zz})$.

3. EXPERIMENTAL SETUP

3.1. Speaker recognition system

In the current study, an i-vector-based speaker recognition system is employed [24]. In this system, a gender-dependent universal background model (UBM) of 512 components was trained using a subset of the Callfriend, Fisher, Switchboard and NIST SRE 2004 speech corpora, and sufficient statistics were computed for each utterance of interest. Total variability matrix was trained by employing a subset of the NIST SRE 2004-2008, Fisher and Switchboard data. Utterance-level, 450-dimensional i-vectors were extracted next. The i-vectors were post-processed using linear discriminant analysis to project the vectors to a 200-dimensional space. In addition, mean removal, length normalization and within-class covariance normalization [25] were applied on the i-vectors. Finally, probabilistic linear discriminant analysis [26] modeling was employed to calculate the recognition score.

The feature extraction stage was started by computing the first 20 Mel-frequency cepstral coefficients (MFCC) from 30-ms, Hamming-windowed speech frames with 50% overlap. The first coefficient, c_0, was discarded. The power spectrum estimate used for the MFCC computation was obtained by computing linear prediction (LP) analysis of order 12. Frame energy was included to the feature vector, after which dynamic Δ and $\Delta\Delta$ features were computed and appended in order to obtain 60-dimensional input feature vectors. Since the duration of the utterances is very short, we retained all the frames in both enrollment and recognition.

In order to evaluate the speaker recognition performance in vocal effort mismatch, the spectral characteristics of the enrollment data (i.e. normal speech) are modified in the current study to correspond more closely to the spectral characteristics of the test data (i.e. shouted speech). This spectral modification, importantly, is conducted at the signal frame level, whereas the speaker recognizer remains unchanged. Only voiced frames are modified; these are selected using an energy threshold criterion. The spectral modification is performed using the proposed PSM method jointly with the GMM mapping technique, hence denoting the overall mapping method as PSM-GMM. The ultimate goal of PSM-GMM is to modify the MFCC feature vectors of normal speech to become more shout-like.

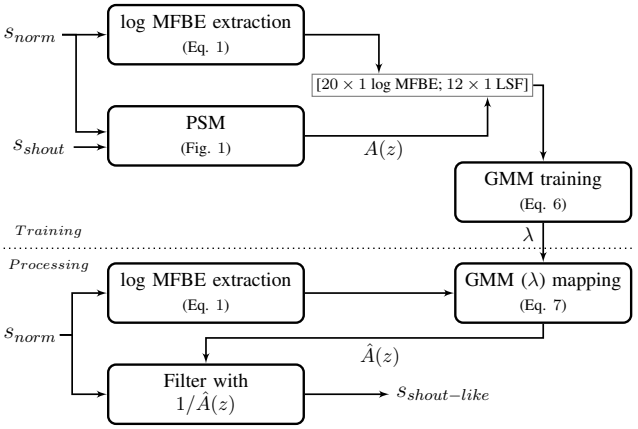

Fig. 3: Flowchart of PSM-GMM for mapping normal speech frames to shout-like speech by employing Eq. 7. The output is shout-like speech frame, $s_{shout-like}$, produced with predicted mapping filter $\hat{H}(z) = 1/\hat{A}(z)$.

The speech dataset employed in the experiments consisted of recordings from 22 native Finnish speakers (11 female, 11 male), each speaker producing 24 short utterances of about 2 seconds produced in two modes: normal and shouting. The speech signals were recorded in an anechoic chamber, first producing normal speech utterances and then repeating the utterances by shouting. No instructions were given to the speakers with respect to achieving a specific sound pressure level (SPL); instead, they were instructed to utter the sentences at a very large vocal effort. The SPL difference between shouted and normal speech varied substantially among the speakers: in a range of 15-33 dB for males and of 17-28 dB for females [27, 28]. Speech signals were originally sampled with 16 kHz, but the data were down-sampled to 8 kHz in the present study. The speech corpus was employed for both the speaker recognition system and PSM-GMM processing.

3.2. PSM-GMM processing

In the present study, PSM-GMM processing was performed with mapping filter $H(z)$ of order $p = 12$ and 6-component full-covariance GMMs. Furthermore, Mel-warping was performed in PSM with warping coefficient $\lambda = 0.31$ [29]. A filter bank of $M = 20$ channels was employed. Figure 3 shows a flowchart of the PSM-GMM algorithm.

Prior to processing the frames, the GMM models were trained with the expectation-maximization algorithm, using a set of normal-shouted speech sample pairs. A dynamic time warping (DTW) algorithm implemented in [30] was used to align the frames of shouted speech to those of normal speech. While such an alignment might not be optimal to pair frames of normal and shouted speech, DTW offers a straightforward and relatively accurate pairing technique that is suitable for training of the GMM mapping. The training feature vectors consisted of: 20-dimensional log MFBE vector of normal speech for $\boldsymbol{z}$, and 12-dimensional line spectral frequency (LSF) coefficient vector corresponding to $A(z)$ for $\boldsymbol{y}$. Once the GMM was trained, the spectral mapping with PSM-GMM could be performed. In case the predicted warped filter $\hat{H}(z) = 1/\hat{A}(z)$ is unstable, this is solved by reflecting, with respect to the unit circle, the pole roots that are outside of it. Finally, shout-like speech was produced with $\hat{H}(z)$ by using the warped filter realization method from [31].

3.3. Evaluation

In the present study, we evaluate a text-independent speaker recognition task under vocal effort mismatch. Similarly to [7], 12 of the 24 utterances available for each speaker are selected and pulled together for enrollment, and the remaining 12 utterances are pulled together for test. This utterance-selection procedure (12 enrollment and 12 test utterances) is repeated using a circular rotation until altogether 12 sets of enrollment-test pairs are obtained for each speaker. No cross-gender comparisons are made, thus resulting in $12 \times 11 = 132$ comparisons for each gender.

The baseline speaker recognition experiments were carried out to demonstrate the system performance in the normal vs. normal ($N - N$) and shouted vs. normal ($S - N$) conditions. As the speaker identification rates presented in Table 1(a) suggest, the presence of shouted speech in the test phase deteriorates the recognition performance. By using the proposed PSM-GMM approach, we form two evaluation scenarios: shout-like vs. shout-like ($\hat{S} - \hat{S}$) and shouted vs. shout-like ($S - \hat{S}$). The $\hat{S} - \hat{S}$ scenario is a resemblance of the $N - N$ condition. The amount of degradation in correct identification rate from $N - N$ to $\hat{S} - \hat{S}$ indicates the degree of speaker-dependent information lost due to applying the PSM-GMM technique. On the other hand, the $S - \hat{S}$ scenario simulates the $S - N$ condition by raising the normal speech to the level of shouted speech in terms of the Mel-frequency band energies. The amount of improvement in correct speaker identification rate from the $S - N$ condition to the $S - \hat{S}$ scenario indicates how effectively the properties of shouted speech can be incorporated to normal speech via the PSM-GMM technique.

In implementing the proposed PSM-GMM approach, we need to first establish the efficiency of the PSM processing without GMM mapping included. We dub this condition of the $S - \hat{S}$ scenario as "oracle" because in producing $\hat{S}$, we assume that the shouted version of the normal speech sample in hand is available. In the oracle condition, the assumption of having a piecewise linear spectrum for the mapping filter, the use of the NNLS algorithm (Eq. 4), and the suboptimal DTW alignment are potential sources of inaccuracy in PSM. In practice, a GMM is trained (Eq. 6) in order to produce shout-like speech for a frame of normal speech. By choosing the type and amount of available information to train such a GMM, we introduce two more conditions for the $S - \hat{S}$ scenario:

SD: A speaker-dependent (SD) PSM-GMM is trained for each trial of speaker recognition where only 12 utterances, considered as enrollment, and their shouted counterparts are available to train PSM-GMM.

GD: All of the normal and shouted utterances for each gender are used to train a gender-dependent (GD) PSM-GMM.

3.4. Results

Table 1 shows the results for both the baseline and the three processing conditions in terms of correct identification rates.

The comparison between the S–N and S–$\hat{S}$ scenarios reveals that in vocal effort mismatch condition, PSM-GMM processing improves the recognition performance over the unprocessed baseline by a large margin for both female and male talkers in the oracle and SD conditions. The relatively high identification rate of 76.9% in the S–$\hat{S}$ scenario for the oracle condition demonstrates a high potential of the PSM technique in converting the spectral properties of normal speech closer to those in shouted speech. Although this identification rate is still well below that of the N–N scenario (95.8%), in the same time, it is well above the identification rate in the S–N condition (42.8%), where no compensation is done. The identification rate of the speaker-dependent (SD) condition in the S–$\hat{S}$ scenario (73.5%) implies that GMM modeling (Eq. 6-7) and its implementation (Figure 3) is capable of providing a fairly accurate mapping filter $\hat{H}(z)$ to produce shout-like speech. Gender-dependent (GD) modeling does not improve the speaker recognition performance when compared to the baseline. Potential reasons for such a behaviour include smoothing caused by the GMM mapping of a fixed model order over all speakers' data.

It should also be noted that in all the mismatch conditions ($S-N$ and $S-\hat{S}$), identification rates are higher for males than for females. This is most likely caused by biasing of the spectral envelopes by harmonics, an effect that is stronger for high-pitched utterances of female talkers. This is in line with the results observed in [7]. Finally, comparing the results of N–N and $\hat{S}$–$\hat{S}$ show that applying PSM-GMM to normal speech does not cause a significant loss of discriminant speaker information.

Table 1: Identification rates (%) for both genders, male (M) and female (F); and average (All). (a): baseline case, (b): PSM-GMM processing conditions. SD and GD are defined at the end of Section 3.3.

(a)

Test-Enroll	N–N			S–N		
	M	**F**	**All**	**M**	**F**	**All**
Baseline	95.5	96.2	95.8	62.1	23.5	42.8

(b)

Test-Enroll	$\hat{S}$–$\hat{S}$			S–$\hat{S}$		
	M	**F**	**All**	**M**	**F**	**All**
Oracle	93.2	93.2	93.2	88.6	65.2	76.9
SD	97.7	92.4	95.1	80.3	66.7	73.5
GD	93.9	96.2	95.1	46.2	37.1	41.7

4. CONCLUSIONS

A perceptual spectral mapping technique was proposed for normal-to-shouted spectral mapping of enrollment speech data in a speaker recognition framework, in which there is vocal effort mismatch between the enrollment (normal speech) and test (shouted speech). The obtained results revealed a substantial improvement in the recognition performance when PSM-GMM processing in the speaker-dependent (SD) condition (73.5%) was compared to a baseline with no spectral mapping (42.8%). The potential of the proposed method prompts to conduct further studies in the topic. One direction is to study the performance of PSM-GMM in a scenario in which its training set consists of speech samples (shouted vs. normal) from an external dataset instead of the scenario that was used in the current study (i.e. the training set contained speech sample pairs from the target speaker). Another direction is to study PSM-GMM in other types of vocal effort mismatch (e.g. whispered vs. normal speech).

5. ACKNOWLEDGEMENTS

This work was supported by the Academy of Finland (project numbers 256961 and 284671). We acknowledge the computational resources provided by the Aalto Science-IT project.

6. REFERENCES

[1] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, 2015.

[2] J. H. L. Hansen, *Analysis and compensation of stressed and noisy speech with application to robust automatic recognition*, Ph.D. thesis, Georgia Institute of Technology, 1988.

[3] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Proc. Interspeech*, 2008, pp. 609–612.

[4] X. Fan and J. H. L. Hansen, "Speaker identification with whispered speech based on modified LFCC parameters and feature mapping," in *Proc. ICASSP*, 2009, pp. 4553–4556.

[5] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertas, "Speaker identification from shouted speech: Analysis and compensation," in *Proc. ICASSP*, 2013, pp. 8027–8031.

[6] J. Pohjalainen, C. Hanilçi, T. Kinnunen, and P. Alku, "Mixture linear prediction in speaker verification under vocal effort mismatch," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1516–1520, 2014.

[7] R. Saeidi, P. Alku, and T. Bäckström, "Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch," *IEEE/ACM Trans. ASLP*, vol. 24, no. 1, pp. 42–53, 2016.

[8] M. Sarria-Paja, M. Senoussaoui, D. O'Shaughnessy, and T. H. Falk, "Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification," in *Proc. ICASSP*, 2016, pp. 5480–5484.

[9] J. H. L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. ASLP*, vol. 17, no. 2, pp. 366–378, 2009.

[10] I. Shahin, "Novel third-order hidden Markov models for speaker identification in shouted talking environments," *Eng. Applic. Artif. Intelligence*, vol. 35, pp. 316–323, 2014.

[11] M. I. Mandasari and R. Saeidi D. A. van Leeuwen, "A study of likelihood ratio calibration in high vocal effort speech for a modern automatic speaker recognition system," in *Proc. IAFPA*, 2012.

[12] H. Lane, B. Tranel, and C. Sisson, "Regulation of voice communication by sensory dynamics," *J. Acoust. Soc. Am.*, vol. 47, pp. 618–624, 1970.

[13] H. Lane and B. Tranel, "The Lombard sign and the role of hearing in speech," *J. Speech Hear. Res.*, vol. 14, pp. 677–709, 1971.

[14] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. ASLP*, vol. 19, no. 4, pp. 788–798, 2011.

[15] J-S. Liénard and M-G. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *J. Acoust. Soc. Am.*, vol. 106, no. 1, pp. 411–422, 1999.

[16] 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.

[17] L. Juvela, "Perceptual spectral matching utilizing Mel-scale filterbanks for statistical parametric speech synthesis with glottal excitation vocoder," M.S. thesis, Aalto University, 2015.

[18] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, no. 2, pp. 142–152, 2008.

[19] C. L. Lawson and R. J. Hanson, *SOLVING LEAST SQUARES PROBLEMS*, Prentice-Hall, 1974.

[20] N. Levinson, "The Wiener RMS error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, no. 4, pp. 261–278, 1947.

[21] J. Durbin, "The fitting of time-series models," *Revue Inst. Int. Stat.*, vol. 28, pp. 233–244, 1960.

[22] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, 1998.

[23] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.

[24] R. Saeidi and D. A. Van Leeuwen, "The Radboud University Nijmegen submission to NIST SRE-2012," in *Proc. NIST SRE Workshop*, 2012.

[25] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Interspeech*, 2006.

[26] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

[27] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio, and P. Alku, "Analysis and synthesis of shouted speech," in *Proc. Interspeech*, 2013, pp. 1544–1548.

[28] J. Pohjalainen, T. Raitio, S. Yrttiaho, and P. Alku, "Detection of shouted speech in noise: human and machine," *J. Acoust. Soc. Am.*, vol. 133, no. 4, pp. 2377–2389, 2013.

[29] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - A unified approach to speech spectral estimation," in *Proc. ICSLP*, 1994, pp. 1043–1046.

[30] D. Ellis, "Dynamic time warp (DTW) in Matlab," Website resource: `http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/`, 2003, Accessed 19.05.2016.

[31] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *J. Audio Eng. Soc.*, vol. 48, no. 11, pp. 1011–1031, 2000.