Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task

Emma Jokinen[a,*], Rahim Saeidi[a], Tomi Kinnunen[b], Paavo Alku[a]

[a]*Department of Signal Processing and Acoustics, Aalto University, PO Box 13000, FI-00076 Aalto, Finland*
[b]*School of Computing, University of Eastern Finland, PO Box 111, FI-80101 Joensuu, Finland*

Abstract

In shouting, speakers use increased vocal effort to convey spoken messages over distance or above environmental noise. For automatic speaker recognition systems trained using normal speech, shouting causes a severe vocal effort mismatch between the enrollment and test hence reducing the recognition performance. In this study, two compensation methods are proposed to tackle the mismatch in a shouted versus normal speaker recognition task. These techniques are applied in the feature extraction stage of a speaker recognition system to modify the spectral envelopes of shouts to be closer to those in normal speech. The techniques modify the all-pole power spectrum of the MFCC computation chain with shouted-to-normal compensation filtering that is obtained using a GMM-based statistical mapping. In an evaluation using the state-of-the-art i-vector based recognition system, the proposed techniques provided considerable improvements in identification rates compared to the case when shouted speech spectra were not processed.

Keywords: Speaker recognition; vocal effort mismatch; shouted speech

1. Introduction

Human speech contains a great deal of intrinsic variability, such as changes in fundamental frequency ($F0$) and intonation, different styles of speaking and phonation and different levels of vocal effort. Speaking style modifications, such as the Lombard effect which takes place when speaking in noisy conditions, are naturally used by talkers to make speech more intelligible to human listeners (Summers et al., 1988). The performance of data-driven systems, however, typically suffers when such changes in speaking style occur (Junqua, 1993). The loss in performance is due to the mismatch between the system's training conditions and its testing conditions. In this study, the focus is on severe vocal effort mismatch between the normal speaking mode and shouting. This mismatch condition is studied in automatic speaker recognition where the system has been trained using normal speech but is tested with shouted speech.

Shouting is used in situations where a message needs to be conveyed urgently over a distance or in a noisy situation. Differently from Lombard speech produced also in noisy conditions, shouted speech shows reduced intelligibility for human listeners compared to speech produced in the normal speaking mode (Pickett, 1956). While the vocal effort of Lombard speech rises to some extent from that of normal speech, the corresponding change is much more prominent when changing from normal speech to shouting. In addition to an overall sound level increase (Rostolland, 1982), also a reduction in spectral tilt (Zhang and Hansen, 2007), movement of formant frequencies (Zelinka et al., 2012; Traunmüller and Eriksson, 2000), increase in $F0$ (Rostolland, 1982) and changes in vowel and consonant durations (Rostolland, 1982; Traunmüller and Eriksson, 2000) occur when speakers change their normal speaking style to shouting.

Several studies have indicated that severe vocal effort mismatch between enrollment and test data causes a considerable decrease in recognition rates in automatic speaker recognition (Zhang and Hansen, 2007; Shriberg et al.,

*Corresponding author.
Email addresses: ejjokine@gmail.com (Emma Jokinen), rahim.saeidi@gmail.com (Rahim Saeidi), tkinnu@cs.uef.fi (Tomi Kinnunen), paavo.alku@aalto.fi (Paavo Alku)

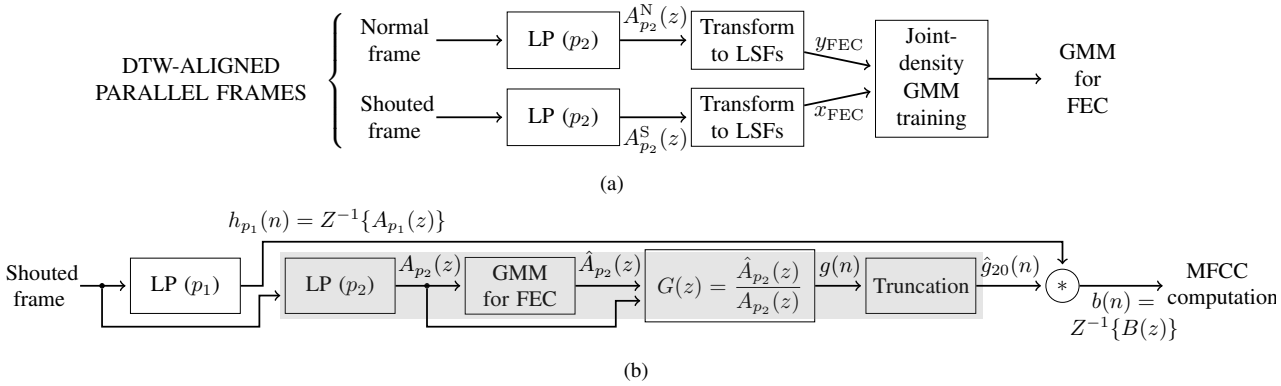


Figure 1: Flowchart of the full envelope compensation (FEC) technique during (a) the training and (b) the test phase. The superscripts S and N in the training phase refer to features computed from shouted and normal speech, respectively. The $\boldsymbol{x}_{\mathrm{FEC}}$ and $\boldsymbol{y}_{\mathrm{FEC}}$ are the line spectral frequency (LSF) features used in the joint-density Gaussian mixture model (GMM) training as specified in Eq. 1

2008; Hanilçi et al., 2013b). To alleviate the problem, two main types of solutions have been proposed: (1) robust feature extraction methods and (2) techniques for compensating the standard features for the vocal effort mismatch. In the first group, Hanilçi et al. (2013a) compared several spectral estimation techniques to obtain Mel-frequency cepstral coefficients (MFCCs) more robust to mismatch in vocal effort. Results showed that *stabilised weighted linear prediction* (SWLP) performed better than the other candidates. Also *mixture* (Pohjalainen et al., 2014) and *power-law adjusted* linear prediction (LP) (Saeidi et al., 2016) have been shown to result in better acoustic features in mismatch conditions.

The second group of techniques are based in the domain of acoustic modeling and consist of different compensation algorithms to mitigate the undesirable effects of vocal effort mismatch. Motivated by the success of Gaussian mixture models (GMMs) in voice conversion (Stylianou et al., 1998), a GMM-based compensation of MFCCs was proposed by Hanilçi et al. (2013b). This technique improved the recognition rates in shouted versus normal mismatch conditions. Similar results were also reported by Ramírez López et al. (2017). A combination of both robust feature extraction and feature compensation is also possible. This kind of setup has been used, for instance, in the context of whispered speech where it improved speaker recognition accuracy in mismatched conditions (Fan and Hansen, 2009).

In this study, two compensation techniques are proposed for handling severe vocal effort mismatch (shouted versus normal) in a speaker identification task. The techniques are applied in the feature extraction stage of the speaker recognition system and their aim is to modify the spectral envelopes in shouts in such a way that the resulting acoustic features are better matched with the features extracted from normal speech in training. The techniques modify the power spectral estimate that is used in the MFCC computation with a shouted-to-normal compensation filter obtained using a GMM-based statistical mapping. One of the techniques aims to compensate for the changes in spectral tilt whereas the other technique uses a more general spectral model in the compensation. The proposed methods are evaluated in a shouted versus normal inset speaker identification task against three different reference techniques.

2. Shouted-to-normal vocal effort compensation

The vocal effort compensation is applied in the first stage of the MFCC (Davis and Mermelstein, 1980) chain, the computation of the signal's power spectrum. In the current study, the power spectrum of the MFCC chain is computed parametrically using LP as was done by Saeidi et al. (2016). Figs. 1 and 2 show the flow diagrams of the proposed two compensation methods: full envelope compensation (FEC) is shown in Fig. 1 and smoothed envelope compensation (SEC) is shown in Fig. 2. Both of the figures are divided into two parts: the training phase and the test phase. In the training phase, frames from parallel normal and shouted samples are aligned using dynamic time warping (DTW) and spectral features, denoted by $A^{\mathrm{N}}_{p_2}(z)$ and $A^{\mathrm{S}}_{p_2}(z)$ for normal and shouted speech, respectively, are computed. The spectral features parameterized are then used to train a joint-density GMM, used as a regression method for mapping shouted speech to normal speech. In the test phase, both methods take as an input a frame of shouted speech and yield

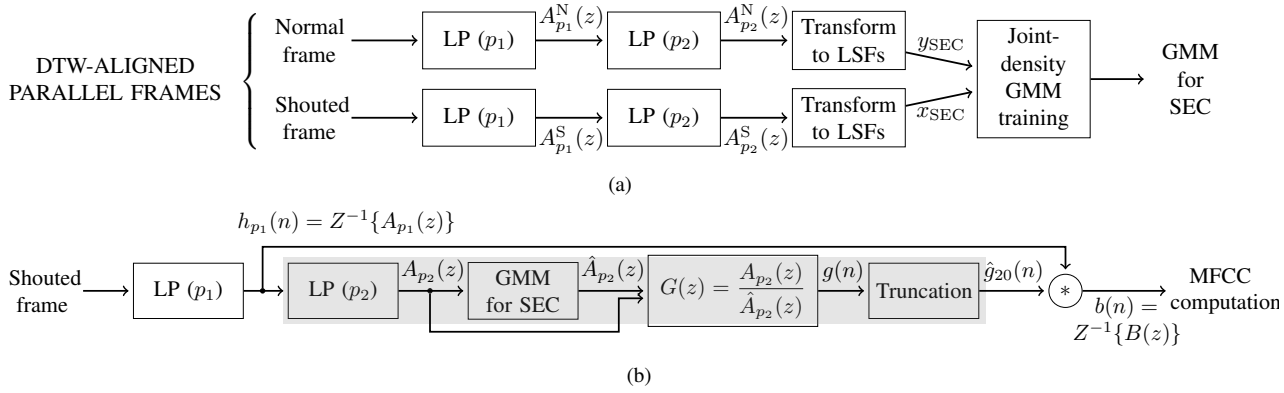


Figure 2: Flowchart of the smoothed envelope compensation (SEC) technique during (a) the training and (b) the test phase. The superscripts S and N in the training phase refer to features computed from shouted and normal speech, respectively. The $\boldsymbol{x}_{\mathrm{SEC}}$ and $\boldsymbol{y}_{\mathrm{SEC}}$ are the line spectral frequency (LSF) features used in the joint-density Gaussian mixture model (GMM) training as specified in Eq. 1

as an output an all-pole spectral model, denoted by $B(z) = Z\{b(n)\}$, which is then used to compute the power spectrum input to the later stages of the MFCC chain. The aim of both methods is to compute such $B(z)$ that fits the spectral envelope of normal speech better than the spectral envelope computed from the original shouted speech frame.

2.1. Training phase

For the training of the statistical mapping, the parallel shouted and normal speech samples were aligned with dynamic time warping (DTW) (Ellis, 2003) using 30-ms frames with 15-ms shift.

In the FEC technique, the features used to train the GMM are the p_2th-order LP analysis computed from the aligned normal and shouted frames. $A_{p_2}(z)$ is therefore an inverse filter and the corresponding all-pole filter models the spectral envelope of the corresponding normal and shouted frames.

The two proposed methods, FEC and SEC, differ in their usage of LP analysis. While the former uses a single-stage LP analysis, in the SEC technique, the impulse response of the first LP analysis, $h_{p_1}(n)$, is fed as a time-domain input to the second LP analysis (this procedure, called *double-LP*, has been used previously by Jokinen et al. (2014) in converting spectral tilt from normal to Lombard speech). Since the second LP analysis with order p_2 ($p_2 < p_1$) models the spectral envelope of the inverse filter $A_{p_1}(z)$, this LP analysis yields a smooth *all-zero* estimate for the spectral envelope of the input frame.

After the spectral features are calculated, they are transformed to line spectral frequencies (LSFs) for GMM training. The statistical dependencies between the LSFs of shouts $\boldsymbol{x}$ and the LSFs of normal speech $\boldsymbol{y}$ are modeled as a GMM

$$p(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i}^{I} w_i N\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \middle| \begin{bmatrix}\boldsymbol{\mu}_{x|i}\\ \boldsymbol{\mu}_{y|i}\end{bmatrix} \begin{bmatrix}\boldsymbol{\Sigma}_{xx|i} & \boldsymbol{\Sigma}_{xy|i}\\ \boldsymbol{\Sigma}_{yx|i} & \boldsymbol{\Sigma}_{yy|i}\end{bmatrix}\right), \tag{1}$$

where the component probabilities are denoted as w_i, the mean vectors as $\boldsymbol{\mu}_i$, and the covariance matrices as $\boldsymbol{\Sigma}_i$. The model parameters are trained with the expectation-maximization algorithm implemented by Paalanen et al. (2005). The minimum mean square error (MMSE) estimate for features $\boldsymbol{y}^*$ that correspond to test input $\boldsymbol{x}^*$ is calculated based on the GMM distribution as (Stylianou et al., 1998)

$$\boldsymbol{y}^* = \sum_{i}^{I} P(i|\boldsymbol{x}^*)\left[\boldsymbol{\mu}_{y|i} + \boldsymbol{A}_i(\boldsymbol{x}^* - \boldsymbol{\mu}_{x|i})\right], \tag{2}$$

where the linear transformations $\boldsymbol{A}_i = \boldsymbol{\Sigma}_{yx|i}\boldsymbol{\Sigma}^{-1}_{xx|i}$ and the posterior probabilities $P(i|\boldsymbol{x}^*)$ are calculated based on the prior probabilities w_i and the feature likelihoods $N(\boldsymbol{x}^*|\boldsymbol{\mu}_{x|i}, \boldsymbol{\Sigma}_{xx|i})$.

Both speaker-dependent GMMs as well as gender-dependent GMMs were used for compensation. For the speaker-dependent models, $I = 5$ full-covariance components were used, whereas for the gender-dependent models, GMMs with $I \in \{10, 15, 20\}$ full-covariance components were evaluated.

2.2. Test phase

In the test phase of both FEC and SEC, an all-pole spectral model, $A_{p_1}(z)$, is first computed from a shouted frame using LP with order $p_1 = 12$. The impulse response of the corresponding LP inverse filter, $h_{p_1}(n)$, is shown by the upper signal paths in both of the test phase flowcharts in Figs. 1 and 2. Paths below the upper ones, shown in gray, depict the GMM-based spectral envelope estimation that is used in compensating the effect of shouting from the impulse response. The two paths are joined at the end using a convolution that yields $b(n)$. Both the FEC and SEC techniques utilize the GMMs trained during the training phase, where the obtained filter coefficient vectors are presented as LSFs. After the mapping, the stability of the output filter is checked and if necessary, the roots outside of the unit circle are replaced with their mirror-image pairs inside the unit circle.

In the FEC technique, the GMM-based compensation uses p_2th-order ($p_2 \leq p_1$) *all-pole* model computed from the shouted frame. The inverse filter $\hat{A}_{p_2}(z)$ of normal speech is estimated with a GMM based on $A_{p_2}(z)$. The compensation needs to be defined in an inverted form because it is applied to the impulse response of the LP *inverse* filter. Therefore, the compensation filter is defined as $G(z) = \hat{A}_{p_2}(z)/A_{p_2}(z)$ and its impulse response is denoted by $g(n)$. Since the feature extraction for the recognizer assumes the spectral envelope to be all-pole, the compensation filter is truncated by cutting its impulse response to 20 samples, yielding an impulse response denoted as $\hat{g}_{20}(n)$. The impulse response of the compensated inverse filter is finally computed as $b(n) = \hat{g}_{20}(n) * h_{p_1}(n)$.

In the SEC technique, the second LP analysis with order p_2 ($p_2 < p_1$) is a smooth *all-zero* estimate for the spectral envelope of the input frame. To obtain a tilt estimate for normal speech, $\hat{A}_{p_2}(z)$, the previously trained GMM is again used. Because the all-zero filter $A_{p_2}(z)$ now models the spectral envelope instead of its inverse, the compensation filter is given as $G(z) = A_{p_2}(z)/\hat{A}_{p_2}(z)$. The impulse response truncation and the convolution are as in the FEC technique. The main steps of both FEC and SEC are demonstrated in Fig. 3 with examples of obtained spectra at each stage.

3. Reference techniques

In order to evaluate the performance of the proposed vocal effort compensation methods in speaker identification, three techniques were selected as reference methods. Two of the reference techniques, 1/3-octave band energy regression fit and LP1 fit, were selected because they aim to model the spectral tilt of speech parametrically and can therefore be used in vocal effort compensation in a similar manner as FEC and SEC. The third reference technique, vocal effort compensation by directly modifying the MFCCs computed from shouted speech, represents a different, more generic approach which does not include a separate parametric spectral tilt estimation phase. Below, these three reference methods are described.

3.1. Techniques based on compensating the spectral tilt

1/3-octave band energy regression fit (REG): the REG method is based on (Lu and Cooke, 2009) where spectral tilts of normal and Lombard speech were parameterized by a regression fit to spectral energies at 1/3-octave bands. This technique has been later used also for spectral tilt -based intelligibility enhancement of telephone speech (Jokinen et al., 2014) and in the present study, the technique is applied in the same form. Differing from the original method, a 4th-order all-pole filter is computed from 1/3-octave band energies and only 15 bands up to 4 kHz are used. Based on the sub-band energies, E_i, a magnitude spectrum is constructed where each component in the ith sub-band is set to $\sqrt{E_i}/N_i$, where N_i is the number of components in the ith sub-band. Autocorrelation is computed from the magnitude spectrum and used to obtain a LP fit with the Levinson-Durbin recursion (Rabiner and Schafer, 2007). These features are transformed to LSFs and then used as an input to a joint-density GMM which is used to map the shouted features to normal features at test time.

LP1 fit (LP1): the LP1 method is based on using first order LP analysis to estimate the spectral tilt. The coefficient of the first order LP analysis is then used as an input to a joint-density GMM which is used to map the shouted features to normal features at test phase.

3.2. MFCC-based compensation

The **MFCC-based compensation (HAN)** was proposed for vocal effort compensation in speaker recognition by Hanilçi et al. (2013b). In their original study, either 16 MFCCs or the entire 48-dimensional feature vector with MFCCs and their first and second time derivatives were mapped using a joint-density GMM. The speaker recognition

experiments were conducted on the same database as in the current study, but the training of the speaker-independent models was done using another database containing emotional speech. For the current study, the MFCC-based compensation technique was adopted to a slightly different feature extraction process: the MFCCs were mapped before the addition of energy to the features and the training and testing was done using the same database. Due to the small number of training samples, the mapping of the full feature vector including first and second time derivatives was not used.

4. Experimental setup

4.1. Speech data

The speech database used in the experiments of this study contains normal and shouted speech from 11 male and 11 female speakers (Pohjalainen et al., 2013). For each speaker, 24 Finnish sentences of approximately 1.5 seconds in duration (Raitio et al., 2013) were recorded. The sentences were first produced in normal voice after which they were repeated in shouted voice. The difference in sound pressure levels between the shouted and normal speech ranged from 15 to 33 dB for male speakers and from 17 to 28 dB for female speakers (Pohjalainen et al., 2013). The original recordings were sampled at 16 kHz but for the current study they were downsampled to 8 kHz.

For each speaker, the speaker recognition system pools together 12 sentences for enrollment and 12 for testing. This selection is circularly shifted by one sample 12 times which results in 12 pairs of enrollment and test data for each speaker (Saeidi et al., 2016). For the spectral envelope compensation, speaker-dependent GMMs were first trained using the same division of samples in enrollment and testing. For the gender-dependent GMMs, the training data was obtained by pooling data from all the speakers of the given gender. In order to keep the training and test data separate, the test utterance was dropped from all the speakers for the training.

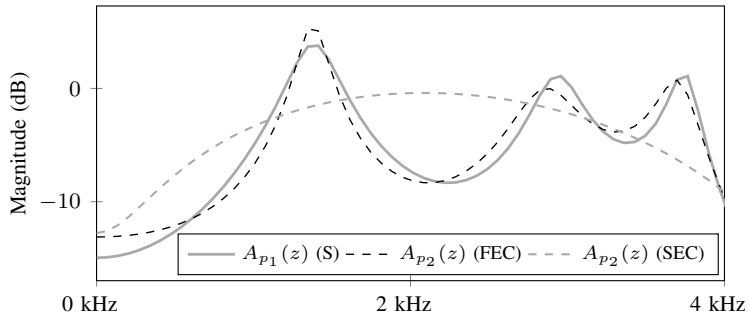
4.2. Speaker recognition system

For the speaker recognition, an i-vector -based system was used (Saeidi and Van Leeuwen, 2012; Saeidi et al., 2013; Dehak et al., 2011) by taking advantage of a gender-dependent universal background model (UBM) with 512 components (Reynolds et al., 2000). The UBM was trained using a subset of the Callfriend, Fisher, Switchboard and NIST SRE 2004 speech corpora. A subset of the NIST SRE 2004-2008, Fisher and Switchboard corpora were used to train the total variability space (Dehak et al., 2011) for factorizing the GMM mean supervectors. The 450-dimensional utterance-level i-vectors were processed further using a linear discriminant analysis projection reducing the dimensionality to 200. The i-vectors were centralized and length-normalized (Garcia-Romero and Espy-Wilson, 2011) before probabilistic linear discriminant analysis (Prince and Elder, 2007) modeling was utilized to calculate the matching scores.

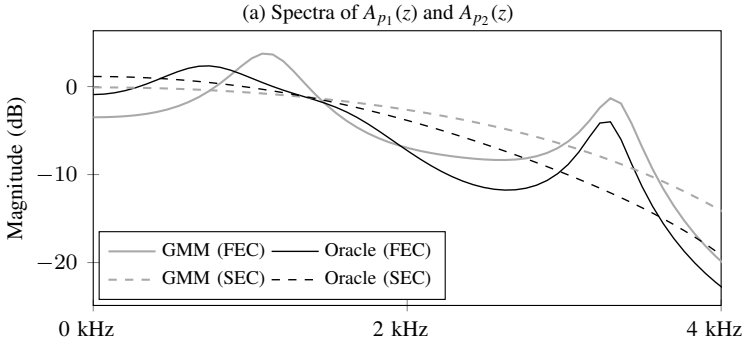
At run-time, the features were extracted by first using conventional LP with $p_1 = 12$ on the pre-emphasized frames for spectrum estimation. At this stage, the spectral envelopes of shouts were compensated with either FEC or SEC except in the baseline normal and shouted conditions. This approach assumes that the system is able to detect whether the incoming speech sample is produced in shouted or normal speaking mode. For the speaker-dependent compensation, the trained GMMs are attached to their respective speaker model and this association is done in the training phase. For the gender-dependent compensation, each speaker is similarly associated with either a male or a female GMM during training time. After this, 19 MFCCs were computed based on the spectral estimate and the energies of the frames were attached to the feature vectors. To obtain 60-dimensional feature vectors, the Δ and $\Delta\Delta$ features were calculated and appended. The features were then filtered with a combination of quantile-based cepstral dynamics normalization and RASTA filtering (Bořil and Hansen, 2010). As described in Section 4.1 (following the protocol used by Saeidi et al. (2016)), the 24 utterances from the speaker were divided equally into enrollment and test data which resulted in approximately 10 seconds of speech in both sets.

5. Results

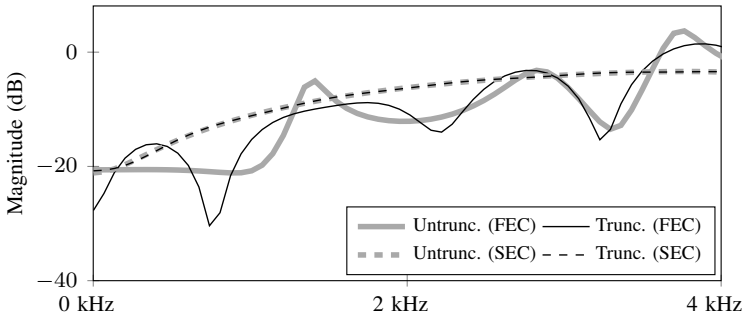
The results of the speaker identification experiments are shown in Tables 1-3. The baseline identification accuracies (without any compensation) for normal and shouted speech are shown in Table 1 and accuracies using the speaker-dependent and gender-dependent GMMs are given in Tables 2 and 3, respectively. Furthermore, Table 2 also



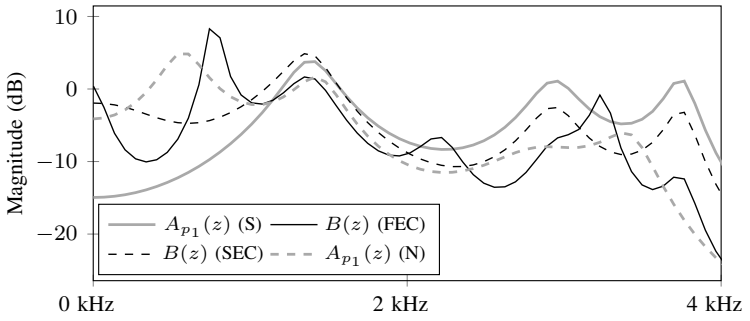
(a) Spectra of $A_{p_1}(z)$ and $A_{p_2}(z)$



(b) Spectra of $\hat{A}_{p_2}(z)$



(c) Spectra of $G(z)$



(d) Spectra of $B(z)$ and $A_{p_1}(z)$ in normal and shouted speech

Figure 3: Steps of the spectral envelope compensation (see Figs. 1 and 2) using the full envelope compensation (FEC, $p_2 = 10$) and the smoothed envelope compensation (SEC, $p_2 = 6$) techniques. The compensation proceeds from (a) the spectral envelope of the shouted frame ($A_{p_2}(z)$) to (b) the GMM predicted normal envelope ($\hat{A}_{p_2}(z)$). These are used to form (c) the compensation filter ($G(z)$), which is finally used to obtain the (d) compensated inverse filter ($B(z)$). In the figure, the power spectral envelopes ($A_{p_1}(z)$) computed from the shouted input frame and the normal target frame are denoted with S and N, respectively.

Table 1: Baseline accuracies (in %) from the speaker identification experiments with normal versus normal and shouted versus normal speech for male, female and all speakers.

Normal vs. normal			Shouted vs. normal		
Male	Female	All	Male	Female	All
90.9	98.5	94.7	59.8	29.5	44.7

Table 2: Identification accuracies (in %) from the speaker identification experiments with shouted versus normal speech for male, female and all speakers using the two vocal effort compensation techniques, FEC and SEC, with different model orders (p_2) as well as the reference techniques, LP1, REG, and HAN, with both oracle and speaker-dependent Gaussian mixture models. In the bottom row, also the baseline scores in the shouted condition (without any compensation) from Table 1 are given to ease comparison.[1]

			Oracle			Speaker-dependent		
		p_2	Male	Female	All	Male	Female	All
PROPOSED	FEC	12	86.4	74.2	80.3	86.4	72.7	79.5
		10	81.1	68.2	74.6	87.1	64.4	75.8
		8	56.8	64.4	60.6	60.6	67.4	64.0
		6	62.1	44.7	53.4	65.9	41.7	53.8
	SEC	10	72.0	59.8	65.9	65.9	56.8	61.4
		8	73.5	54.5	64.0	62.9	58.3	60.6
		6	59.8	59.8	59.8	59.8	43.9	51.9
REF.	LP1		53.0	28.8	40.9	57.6	34.8	46.2
	REG		53.8	43.9	48.9	50.8	40.2	45.5
	HAN		92.4	85.6	89.0	88.6	90.9	89.8
Baseline scores			59.8	29.5	44.7	59.8	29.5	44.7

contains the identification rates of the oracle scenario which correspond to the maximum achievable performance in the identification task and therefore provides further insight into the compensation techniques. The oracle scenario was computed by first using DTW to find the corresponding shouted and normal frames and then replacing the GMM-estimated normal features in the compensation methods with the corresponding features taken directly from the DTW-aligned frame.

5.1. Results of the oracle scenario

The identification rates obtained using normal speech (94.7%) and the best oracle scenario using FEC with $p_2 = 12$ (80.3%) are slightly different. While in this oracle scenario, the compensation filter, when computed without truncation, should be able to fully counter the effects of the difference in vocal effort, several reasons for the gap in recognition rates can be identified. First, the oracle scenarios utilize only the frames that have been aligned by DTW, which considerably reduces the number of frames compared to the recognition condition with normal speech. Second, the energy used as one of the features of the recognizer is computed from the original shouted frames resulting in partially mismatched features. Finally, an additional loss in performance might be caused by the use of a truncated compensation filter instead of the untruncated one. As shown in Fig. 3c, the difference between the untruncated and truncated compensation filters can become considerable with large values of p_2.

The SEC technique (65.9%) improves the recognition rates compared to using shouted speech without compensation (44.7%) which suggests that even rough spectral envelopes, loosely referred to as spectral tilt, play a role in affecting spectral matching for automatic recognition system in the presence of severe vocal effort mismatch. However, the FEC technique is able to produce much higher identification rates compared to the rates obtained with shouted

[1] After the acceptance of this article, a software bug was found that affected the identification accuracies of LP1 and REG. The bug was corrected when examining the article proofs.

Table 3: Identification accuracies (in %) from the speaker identification experiments with shouted versus normal speech for male, female and all speakers using the two vocal effort compensation techniques, FEC and SEC, with different model orders (p_2) as well as the reference techniques, LP1, REG, and HAN, using gender-dependent Gaussian mixture models with 10, 15, and 20 components. In the bottom row, also the baseline scores in the shouted condition (without any compensation) from Table 1 are given to ease comparison.[2]

			$I = 10$			$I = 15$			$I = 20$		
		p_2	Male	Female	All	Male	Female	All	Male	Female	All
PROPOSED	FEC	12	58.3	37.9	48.1	69.7	39.4	54.5	65.9	39.4	52.7
		10	53.0	33.3	43.2	55.3	35.6	45.5	66.7	37.9	52.3
		8	38.6	42.4	40.5	43.2	40.9	42.0	44.7	42.4	43.6
		6	40.9	25.0	33.0	44.7	24.2	34.5	43.2	26.5	34.8
	SEC	10	45.5	28.8	36.7	40.2	31.8	36.0	40.9	31.1	36.0
		8	49.2	28.0	38.6	48.5	32.6	40.5	50.8	40.2	45.5
		6	49.2	40.2	44.7	47.7	42.4	45.1	49.2	43.2	46.2
REF.	LP1		59.8	33.3	46.6	59.1	33.3	46.2	59.1	34.1	46.6
	REG		51.5	31.8	41.7	53.0	32.6	42.6	48.5	32.6	40.5
	HAN		56.8	31.8	44.3	73.5	50.8	62.1	69.7	42.4	56.1
Baseline scores			59.8	29.5	44.7	59.8	29.5	44.7	59.8	29.5	44.7

speech which indicates that a compensation technique including further spectral details is more efficient in bridging the gap between normal and shouted speech. In Fig. 3d, the spectral envelope compensated using FEC shows shifted formants in addition to modified spectral tilt. The highest accuracy in the oracle scenario is achieved by HAN (89.0%) which directly maps the MFCCs of shouted speech to those of normal speech. The two other reference techniques, LP1 and REG, do not improve the baseline accuracy.

5.2. Results of the speaker-dependent compensation

In a realistic system, a GMM mapping is used to estimate $\hat{A}_{p_2}(z)$ which is a less accurate procedure to estimate the spectral tilt of normal speech compared to the one used in the oracle scenario. Using a speaker-dependent GMM results in some loss in identification performance, for instance, with the SEC technique using the smallest model order, $p_2 = 6$ (oracle: 59.8%, speaker-dependent: 51.9%). However, for the FEC and HAN techniques, the differences between the oracle and the speaker-dependent scenario are relatively small. In some cases, the accuracies obtained in the speaker-dependent scenario are even slightly higher than those obtained in the oracle scenario. The identification in the GMM-mapped vocal effort compensation takes advantage of a larger number of frames than the oracle scenarios which are based on the DTW matching of frames. The smaller number of frames results in lower identification rates for the oracle scenarios as was discussed earlier. All of the speaker-dependent results were obtained using $I = 5$ GMM components which gave a good compromise between the amount of required training data and the performance of the model in the preliminary experiments.

The performance of the proposed methods and the MFCC-based reference technique was also compared in a speaker-dependent scenario using different amounts of training data. The evaluation was conducted only with FEC using $p_2 = 12$ and SEC using $p_2 = 10$ to limit the amount of test cases. The training data was reduced from the maximum of 12 parallel shouted and normal utterances to the minimum of 1 parallel utterance pair. The identification accuracies for the different test scenarios are shown in Fig. 4. While the HAN technique outperforms both FEC and SEC with moderate amounts of training data, both of the proposed methods show a clearly better identification accuracy when the number of utterances in the training data is small. Especially, the FEC technique provides improvement over the baseline without compensation even with limited training data.

[2] After the acceptance of this article, a software bug was found that affected the identification accuracies of LP1 and REG. The bug was corrected when examining the article proofs.

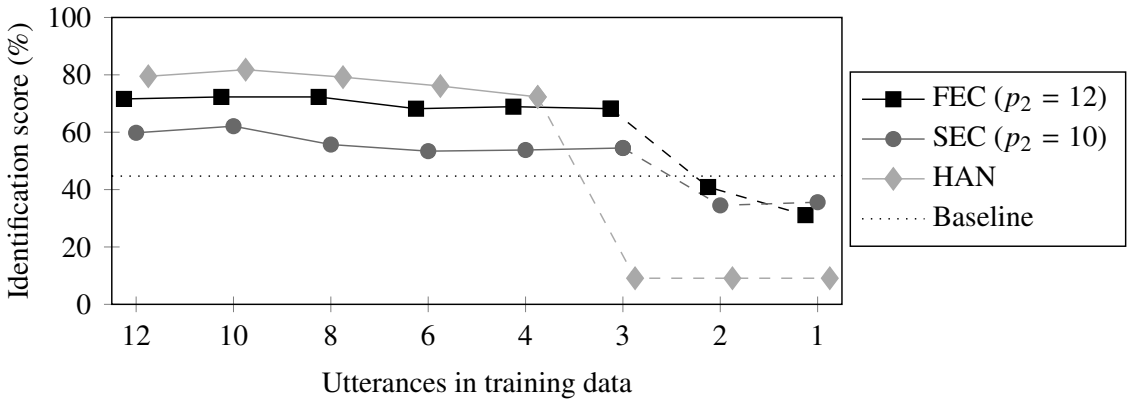


Figure 4: Identification accuracy for all speakers using different compensation techniques with speaker-dependent GMMs. The 2-component GMMs are trained with different number of parallel utterances. The conditions marked with solid line use full-covariance components whereas the dashed line marks the conditions where block-diagonal covariances are used due to the small amount of training data. The baseline score, marked with a horizontal, dotted line, refers to the identification accuracy using shouted speech without any compensation.

5.3. Results of the gender-dependent compensation

While the spectral envelope compensation using speaker-dependent GMMs provides largely improved identification rates compared to the shouted condition, the schemes with gender-dependent GMMs do not increase the performance over the baseline in most cases for male speakers. For FEC with the largest model size ($p_2 = 12$) and with $I = 15$ or $I = 20$ GMM components, the identification accuracies are improved for both male and female speakers. For male speakers otherwise, the identification rates are even slightly lower compared to the shouted condition which suggests that the mapping impairs the speaker-specific features used in the recognition. For female speakers, the gender-dependent compensation achieves higher accuracies compared to the baseline in most cases evaluated. As shown in Table 3, the identification rates generally improve when the number of GMM components is increased from $I = 10$ to $I = 15$ and further to $I = 20$. While the HAN technique still provides slightly higher recognition scores than the proposed techniques, the difference between HAN and the two proposed methods is smaller than in the speaker-dependent setting.

6. Conclusion

Two spectral envelope compensation techniques, full envelope compensation (FEC) and smoothed envelope compensation (SEC), were proposed for vocal effort compensation in a normal versus shouted speaker recognition task. The introduced methods use a statistical mapping to estimate a compensation filter that is applied on the spectral envelope estimates of shouts in the recognizer before the MFCC extraction. The performance of the proposed algorithms was evaluated in a shouted versus normal inset speaker identification task with 22 speakers using the state-of-the-art i-vector based speaker recognition system.

The results show that while both of the proposed techniques were able to provide improvement in an oracle scenario over the baseline identification rate obtained with shouted speech, the compensation taking advantage of a more detailed spectral envelope model produced significantly better results. A similar difference between the two techniques was observed using the speaker-dependent GMM mapping. Compared to the best oracle scenario which increased the identification rate of shouted speech from 44.7% to 80.3%, the compensation based on the speaker-dependent GMM achieved 79.5% identification rate overall. In the gender-dependent experiments, the identification accuracies were improved over the baseline especially in the case of female speakers. While a previously proposed MFCC-based compensation technique, HAN, still outperformed the proposed methods when the number of training utterances was larger than 4, both of the proposed methods gave a better performance in cases when the number of training utterances was less than 4. Therefore, the proposed methods, particularly FEC, could find use in speaker recognition scenarios with highly limited training data.

7. Acknowledgements

This work was supported by the Academy of Finland (project numbers 284671, 312490, and 309629). We acknowledge the computational resources provided by the Aalto Science-IT project.

Bořil, H., Hansen, J., 2010. Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments. IEEE Trans Audio, Speech, Lang. Process. 18 (6), 1379–1393.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech, Signal Process. 28 (4), 357–366.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech and Lang. Proc. 19 (4), 788–798.

Ellis, D., 2003. Dynamic time warp (DTW) in Matlab. Visited 16.03.2014, URL: http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/.

Fan, X., Hansen, J., 2009. Speaker identification with whispered speech based on modified LFCC parameters and feature mapping. In: Proc. ICASSP. pp. 4553–4556.

Garcia-Romero, D., Espy-Wilson, C., 2011. Analysis of i-vector length normalization in speaker recognition systems. In: Proc. Interspeech. pp. 249–252.

Hanilçi, C., Kinnunen, T., Rajan, P., Pohjalainen, J., Alku, P., Ertaş, F., 2013a. Comparison of spectrum estimators in speaker verification: Mismatch conditions induced by vocal effort. In: Proc. Interspeech. pp. 2881–2885.

Hanilçi, C., Kinnunen, T., Saeidi, R., Pohjalainen, J., Alku, P., Ertaş, F., 2013b. Speaker identification from shouted speech: Analysis and compensation. In: Proc. ICASSP. pp. 8027–8031.

Jokinen, E., Remes, U., Takanen, M., Palomäki, K., Kurimo, M., Alku, P., 2014. Spectral tilt modelling with GMMs for intelligibility enhancement of narrowband telephone speech. In: Proc. Interspeech. pp. 2036–2040.

Junqua, J.-C., 1993. The Lombard reflex and its role on human listeners and automatic speech recognizers. J. Acoust. Soc. Amer. 93 (1), 510–524.

Lu, Y., Cooke, M., 2009. The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. Speech Commun. 51 (12), 1253–1262.

Paalanen, P., Kämäräinen, J., Kälviäinen, H., 2005. GMMBayes toolbox for Matlab - Gaussian mixture model learning and Bayesian classification. Visited 22.03.2014, URL: http://www.it.lut.fi/project/gmmbayes/.

Pickett, J. M., 1956. Effects of vocal force on the intelligibility of speech sounds. J. Acoust. Soc. Amer. 28 (5), 902–905.

Pohjalainen, J., Hanilçi, C., Kinnunen, T., Alku, P., 2014. Mixture linear prediction in speaker verification under vocal effort mismatch. IEEE Signal Process. Lett. 21 (12), 1516–1520.

Pohjalainen, J., Raitio, T., Yrttiaho, S., Alku, P., 2013. Detection of shouted speech in noise: Human and machine. J. Acoust. Soc. Amer. 133 (4), 2377–2389.

Prince, S., Elder, J., 2007. Probabilistic linear discriminant analysis for inferences about identity. In: IEEE 11th Int. Conf. on Comput. Vis. pp. 1–8.

Rabiner, L., Schafer, R., 2007. Introduction to digital speech processing. Foundations and Trends® in Signal Processing 1 (1–2), 1–194.

Raitio, T., Suni, A., Pohjalainen, J., Airaksinen, M., Vainio, M., Alku, P., 2013. Analysis and synthesis of shouted speech. In: Proc. Interspeech. pp. 1544–1548.

Ramírez López, A., Saeidi, R., Juvela, L., Alku, P., 2017. Normal-to-shouted speech spectral mapping for speaker recognition under vocal effort mismatch. In: Proc. ICASSP. pp. 4940–4944.

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian mixture models. Digital signal processing 10 (1), 19–41.

Rostolland, D., 1982. Acoustic features of shouted voice. Acta Acustica united with Acustica 50 (2), 118–125.

Saeidi, R., Alku, P., Bäckström, T., 2016. Feature extraction using power-law adjusted linear prediction with application to speaker recognition under severe vocal effort mismatch. IEEE/ACM Trans. Audio, Speech, Lang. Process. 24 (1), 42–53.

Saeidi, R., Lee, K.-A., Kinnunen, T., Hasan, T., Fauve, B., Bousquet, P.-M., Khoury, E., Sordo Martinez, P., Kua, J., You, C., et al., 2013. I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification. In: Proc. Interspeech. pp. 1986–1990.

Saeidi, R., Van Leeuwen, D., 2012. The Radboud University Nijmegen submission to NIST SRE-2012. In: Proc. of the NIST Speaker Recognition Evaluation Workshop.

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

Stylianou, Y., Cappé, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion. IEEE Trans. Speech, Audio Process. 6 (2), 131–142.

Summers, W. V., Pisoni, D., Bernacki, R., Pedlow, R., Stokes, M., 1988. Effects of noise on speech production: Acoustic and perceptual analyses. J. Acoust. Soc. Amer. 84 (3), 917–928.

Traunmüller, H., Eriksson, A., 2000. Acoustic effects of variation in vocal effort by men, women, and children. J. Acoust. Soc. Amer. 107 (6), 3438–3451.

Zelinka, P., Sigmund, M., Schimmel, J., 2012. Impact of vocal effort variability on automatic speech recognition. Speech Commun. 54 (6), 732–742.

Zhang, C., Hansen, J., 2007. Analysis and classification of speech mode: whispered through shouted. In: Proc. Interspeech. pp. 2289–2292.