SPEAKER IDENTIFICATION FROM SHOUTED SPEECH: ANALYSIS AND COMPENSATION

Cemal Hanilçi[1,2], *Tomi Kinnunen*[2], *Rahim Saeidi*[3], *Jouni Pohjalainen*[4], *Paavo Alku*[4], *Figen Ertaş*[1]

[1] Department of Electronic Engineering Uludağ University, Bursa, Turkey
[2] School of Computing, University of Eastern Finland, Joensuu, Finland
[3] Centre for Language and Speech Technology in Radboud University Nijmegen, Netherlands
[4] Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
chanilci@uludag.edu.tr, tkinnu@cs.joensuu.fi, rahim.saeidi@let.ru.nl

ABSTRACT

Text-independent speaker identification is studied using neutral and shouted speech in Finnish to analyze the effect of vocal mode mismatch between training and test utterances. Standard mel-frequency cepstral coefficient (MFCC) features with Gaussian mixture model (GMM) recognizer are used for speaker identification. The results indicate that speaker identification accuracy reduces from perfect (100 %) to 8.71 % under vocal mode mismatch. Because of this dramatic degradation in recognition accuracy, we propose to use a joint density GMM mapping technique for compensating the MFCC features. This mapping is trained on a disjoint emotional speech corpus to create a completely speaker- and speech mode independent emotion-neutralizing mapping. As a result of the compensation, the 8.71 % identification accuracy increases to 32.00 % without degrading the non-mismatched train-test conditions much.

Index Terms— speaker identification, shouted speech

1. INTRODUCTION

Research in both speech and speaker recognition has largely focused on normalizing out undesirable variations caused by transmission channel and acoustic environment. Combating for these *technical* nuisance factors has lead to many successful normalization techniques in feature [1], model [2] and match score domains [3]. A much less studied problem, however, is that of intra-person variations caused by changes in the vocal production process itself. Of particular interest is variation in speaker's *vocal effort*. Vocal effort has a communicative purpose, such as an attempt to conceal the speech content (whispering), increasing intelligibility in noisy environments (loud speech) or indicating emergency or other type of urgency (shouting). Speech in forensic speaker recognition and accident investigation is likely to have been produced under stress and is therefore combined with high vocal effort.

Even though differences between neutral and shouted/loud speech in traditional acoustic parameters - formants, fundamental frequency and intensity - are well studied (e.g. [4, 5, 6]), the effect of vocal effort on automatic speech and speaker recognition [4, 6, 7, 8] has received much less attention. The question of how shouting affects between-speaker and within-speaker differences is not only relevant for forensic speaker recognition, but of fundamental nature that has implications to other recognition applications as well.

In the NIST 2010 speaker recognition evaluation (SRE) campaign, the effect of vocal effort on speaker recognition was analyzed [7]. In that study, speakers produced soft and loud utterances in a controlled set-up. It was reported that mismatched vocal effort between training and test (training with normal vocal effort and testing with high vocal effort) cause degradation of recognition accuracy. The authors of [9] have found that features extracted from nasal syllables are relatively robust to high vocal effort. They have reported that, in context of a GMM-JFA recognizer on the NIST 2010 SRE corpus, the nasal constrained cepstral coefficients tend to bring advantage over using all cepstral coefficients. In [4], whispered speech was found to give the lowest identification rate and it was reported that 98.8 % identification accuracy obtained in neutral training-neutral test condition whereas in neutral training-shouted test case identification accuracy reduced to 56.3 %. An HMM based text-dependent speaker identification method for shouted speech was proposed in [8] and it was reported that identification accuracy decreases from 96 % to 73 % when shouted speech is used for testing. In that study, the speech samples used in the experiments were collected in different sessions and speaker models were trained using neutral speech.

In this study, we consider idealized speaker identification conditions where the typically included effects of channel mismatch, environmental noise and reverberation are completely excluded. To this end, we consider closed-set speaker identification using Finnish utterances recorded in an anechoic chamber. This approach, importantly, enables studying speaker identifiability solely under varying vocal modes; if one cannot correctly identify speakers *even* under such idealized setting, one should not expect accurate recognition under additional nuisance factors due to channel or environment.

In order to study speaker identification in mismatch conditions between neutral and shouted speech, a method based on joint density GMM mapping is proposed to compensate the effect of shouting. To this end, we adopt methods from voice conversion [10] – typically used for speaker identity conversion – to train a speaker-independent joint density Gaussian mixture model mapping on the MFCC feature space. This mapping, intended to remove any expressive factors from a given stream of MFCCs, is independently trained on a disjoint emotional German speech corpus including parallel recordings of neutral and emotionally colored speech samples.

2. NEUTRAL VS. SHOUTED SPEECH

The authors of [4] categorize speech as having five different modes: *whispered*, *soft*, *neutral*, *loud* and *shouted*. Vocal intensity is lowest in whispered speech which is acoustically generated by an aperiodic weak excitation waveform in the absence of the vocal fold vibration. Due to the lack of vocal fold vibration, whispered speech is the lowest vocal mode. Shouted speech, in turn, is the highest vocal mode.

The work was supported by Academy of Finland.

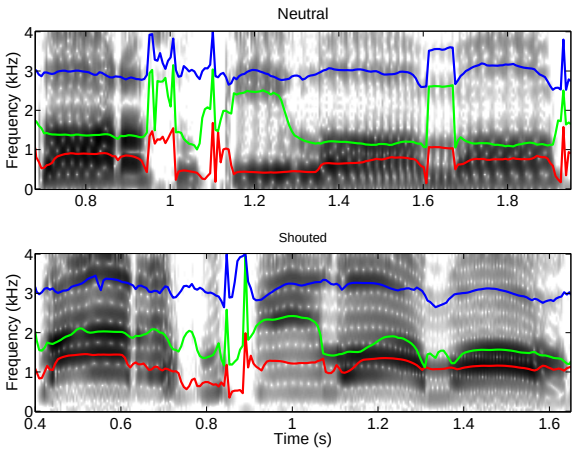

Fig. 1. Spectrograms and first three formants of neutral and shouted versions of the same utterance spoken by a female speaker.

It calls for increased lung effort generating rapid period fluctuation of the vocal folds and a prominent voice excitation, which result in maximal vocal intensity. Current speaker recognition studies mostly focus on neutral, normally spoken speech.

A number of authors have analyzed acoustic differences of neutral and shouted speech. In [5], acoustic differences between normal and shouted speech were analyzed in forensic settings. In that study, it was found that the fundamental frequency (F_0) and the first formant frequency (F_1) increase in shouting whereas the second and the third formants (F_2 and F_3) were less affected by shouting. In [4], different speech modes were analyzed in terms of the sound intensity level, duration and frame energies. It was found that the average sound intensity level of shouted speech is higher than that of neutral speech and sentence duration of shouted sentence is longer than neutral. Number of low energy frames, on the other hand, is smaller in shouting than in neutral speech. This is in line with [7] where statistically significant differences between the average energy levels of normal and high vocal effort utterances in NIST SRE 2010 were reported. In [6], emergency situation detection was studied for an indoor acoustic-based security system and it was found that both F_1 and F_2 and their standard deviations increase in shouting. Recognition of shouted speech was also considered and it was reported that word recognition accuracy decreases for shouted speech.

Acoustic differences between normal and shouted speech can easily be seen from spectrograms. Fig. 1 displays the wideband spectrograms and the first three formants (F_1-F_3) calculated using *Praat*[1] for neutral and shouted version of the same utterance. As seen from the figure, the formants (especially F_1) are shifted to higher frequencies in shouted speech. Differences in neutral and shouted speech are further described in Fig. 2, which show spectra of these two vocal modes computed by fast Fourier transform (FFT) and linear prediction (LP). Clearly, the shouted speech is characterized by sharper peaks in the spectral envelope. The spectral dissimilarities between neutral and shouted speech will affect MFCCs utilized in the feature extraction of speaker recognition resulting in a speech mode mismatch between training and test. In the following, we propose a feature compensation to mitigate for such mismatches.

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

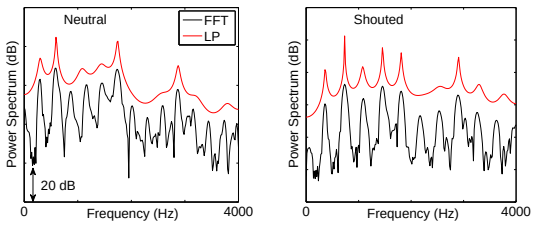

Fig. 2. Power spectra of a voiced speech frame in neutral ($F_0 = 297$ Hz) and shouted mode ($F_0 = 375$ Hz).

3. SHOUT COMPENSATION

To compensate the effect of different speech modes, voice conversion can be utilized to convert an utterance from one mode to another. *Voice conversion* refers to methodologies for converting one speaker's (*source*) utterances to given an impression that they are spoken by another speaker (*target*) [10]. A voice conversion system consists of two main components, signal parameterization and feature mapping function. Signal parametrization model such as STRAIGHT [11] is used for analyzing (and synthesizing) utterances, whereas mapping is used for learning a regression function between the vocal spaces of the source and the target speakers. As we do recognition rather than synthesis or conversion, we only consider the feature mapping part. We directly plug-in our feature mapping function to our recognizer MFCC front-end as will be detailed below.

A generic feature mapping function is denoted here by $f_{\mathbf{\Theta}}(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^d$, where $\mathbf{\Theta}$ denotes the model parameters and d is the dimensionality of the acoustic vectors. In the training phase, the parameters $\mathbf{\Theta}$ are learnt from a training set consisting of frame-aligned feature vector pairs $\{(\mathbf{x}_t, \mathbf{y}_t)|t = 1, 2, \ldots, T\}$. To ensure that training utterances are phonetically aligned, they are usually taken to be *parallel* so that both the source and the target speakers read the same sentences. Alignment of the feature vectors is achieved using dynamic time warping (DTW). In the conversion phase – which is completely text-independent – one applies $\hat{\mathbf{y}}_t = f_{\mathbf{\Theta}}(\mathbf{x}_t)$ for each source vector $\mathbf{x}_t$ to find predicted target speaker vector $\mathbf{y}_t$ for that observation. In this study, we adopt feature mapping techniques from voice conversion to compensate for shouted speech. To this end, now $\mathcal{X}$ and $\mathcal{Y}$ represent non-neutral and neutral vocal spaces of the *same speaker* rather than two different speakers. We compensate non-neutral speech using Gaussian mixture model (GMM) conversion [12]. In particular, we adopt the *joint density GMM* originally proposed in [13]. In this model, the joint distribution of the source (non-neutral) and the target (neutral) features is modeled by GMMs trained using the stacked feature vectors $\mathbf{z}_t = [\mathbf{x}_t^\top, \mathbf{y}_t^\top]^\top$ of dimensionality $2d$. The joint probability density function is given by,

$$p(\mathbf{z}_t|\mathbf{\Theta}^{(z)}) = \sum_{m=1}^{M} P_m^{(z)} \mathcal{N}(\mathbf{z}_t|\boldsymbol{\mu}_m^{(z)}, \mathbf{\Sigma}_m^{(z)}),$$

where $\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}$ and $\mathbf{\Sigma}_m^{(z)} = \begin{bmatrix} \mathbf{\Sigma}_m^{(xx)} & \mathbf{\Sigma}_m^{(xy)} \\ \mathbf{\Sigma}_m^{(yx)} & \mathbf{\Sigma}_m^{(yy)} \end{bmatrix}$ are the mean vector and covariance matrix of the multivariate Gaussian density $\mathcal{N}(\mathbf{z}_t|\boldsymbol{\mu}_m^{(z)}, \mathbf{\Sigma}_m^{(z)})$, respectively, and $P_m^{(z)}$ are the prior probabilities constrained by $P_m^{(z)} \geq 0$ and $\sum_m P_m^{(z)} = 1$. The joint model parameters are estimated to maximize likelihood for the training data set using the conventional expectation-maximization (EM) algorithm [14]. In our implementation, we use full covariance matrices and 40 EM iterations starting from randomized initial solution. Even though speaker recognition systems typically use di-

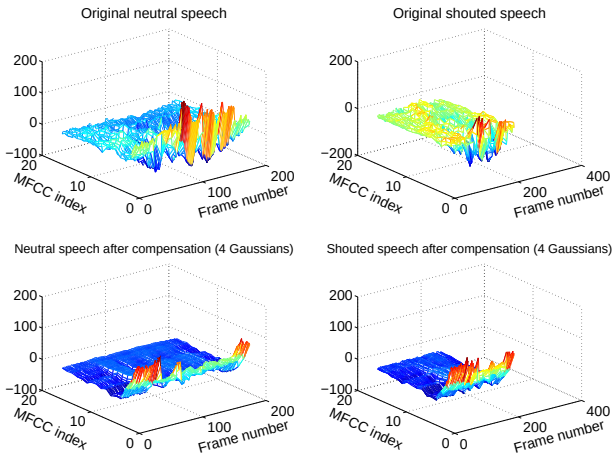

Fig. 3. MFCCs of neutral, shouted and their compensated counterparts.

agonal covariances, full covariances are common in voice conversion. They capture cross-correlations across the source and the target spaces, while diagonal covariance (for all the four submatrices $\mathbf{\Sigma}_m^{(xx)}$, $\mathbf{\Sigma}_m^{(xy)}$, $\mathbf{\Sigma}_m^{(yx)}$ and $\mathbf{\Sigma}_m^{(yy)}$) implies independent conversion of each cepstral coefficient. In preliminary tests, we implemented both variants and, despite small amount of training data, full covariance with less Gaussians outperformed systematically all trialed diagonal conversions (up to 256 Gaussians). To reduce sensitivity to parameter initialization, we repeat training 20 times, each starting from a different random guess, and pick the GMM which yields largest likelihood. Given the trained joint density model, the predictor for future data points is,

$$\hat{\mathbf{y}} = f(\mathbf{x}) = \sum_{m=1}^{M} p_m(\mathbf{x})(\boldsymbol{\mu}_m^{(y)} + \mathbf{\Sigma}_m^{(yx)}(\mathbf{\Sigma}_m^{(xx)})^{-1}(\mathbf{x} - \boldsymbol{\mu}_m^{(x)})),$$

where $p_m(\mathbf{x}) = P_m\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m^x, \mathbf{\Sigma}_m^{xx}) / \sum_k P_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k^x, \mathbf{\Sigma}_k^{xx})$ denotes the posterior probability of $\mathbf{x}$ originating from the m^{th} Gaussian.

Fig. 3 shows the MFCCs of the same utterance spoken with neutral and shouted speech modes and their compensated versions using 4 Gaussians, as an example. It can be seen that the variations between neutral and shouted speech modes of the same utterance are highly reduced after compensation.

4. EXPERIMENTAL SETUP

The speech corpus used in experiments consists of 11 male and 11 female speakers. Each speaker produced 24 Finnish utterances using neutral speech mode. The same 24 utterances were also produced with shouting. The sentences were recorded using a high-quality microphone in an anechoic chamber so that the device, environmental and channel effects are completely excluded. The average duration of utterances is approximately 3 seconds. Half of the sentences are in imperative and half in indicative mood. For more details about the database, refer to [15].

In training the joint density GMM feature mapping, we utilize the *Berlin database of emotional speech* [16][2]. This corpus consists of German speech samples from ten speakers (5 males and 5 females) recorded also in an anechoic chamber. Each speaker produces 5 short and 5 longer sentences in seven different emotional modes: *neutral*, *anger*, *happiness*, *fear*, *boredom*, *disgust* and *sadness*. Using this corpus, we train a speaker-independent feature mapping that attempts to normalize out any emotional effects of a given speech utterance. To this end, we consider all the non-neutral utterances of a given speaker as our source utterances and the corresponding neutral utterance *of that speaker* as the target utterance. The DTW alignment is first computed to the MFCC vectors *within each speaker* by using cosine similarity as a vector similarity measure so that all the non-neutral utterances utilize the corresponding neutral utterance as a target. Additional care is taken to exclude many-to-one and one-to-many assignments of the training vectors [17]. The aligned vector pairs from all the 10 speakers are then pooled and used for training a speaker-independent joint density GMM as detailed above. This feature mapping is then applied to all training and test utterances in our evaluation set.

Compensating for emotions rather than shouting is naturally a more general problem setup. In fact, in preliminary experiments, we trained only *angry-to-neutral* mapping on the same corpus as angriness is among the seven emotions of [16] the one which corresponds best with shouting. However, since we have a rather limited training set with full covariance GMM modeling, including the other source emotions helped preventing numerical problems in GMM training. For the same reason (small training set relative to the dimensionality of the joint feature space), we also experiment with two alternative feature mappings. In the first approach, we train mapping on base MFCC coefficients only and add the delta and double delta coefficients *after* feature mapping. In the second approach, we train the mapping function directly on the higher dimensional MFCC + Δ + Δ^2 features (see below).

In the speaker identification experiments, we use standard MFCCs extracted from 20 ms Hamming windowed speech frames every 10 ms. We use two standard spectrum estimation methods, FFT and LP with prediction order of $p = 20$, to compute spectra of windowed frames. The power spectra are processed through a 27-channel triangular filterbank. The logarithmic filterbank outputs are converted into MFCCs by discrete cosine transform. The first and second time derivatives (Δ and Δ^2) are appended to the first 16 MFCCs which leads to 48 dimensional feature vectors. Finally, cepstral mean and variance normalization (CMVN) are applied to the features. Gaussian mixture model (GMM) is used as the classifier. We use GMMs with 32 Gaussians trained by maximum likelihood (ML) criterion [14] using 5 EM iterations.

We consider text-independent speaker identification in the experiments. Due to relatively small amount of data, the speaker identification experiments are carried out using leave-one-out cross validation to maximize the number of test trials. That is, each speaker model is trained using his/her 23 sentences and the held-out utterance is used for testing. Rotating over all 24 utterances and 22 speakers, this yields $24 \times 22 = 528$ identification trials. In the experiments we consider four different training and test conditions:

- Neutral - Neutral (**N-N**): Training and test utterances are both in neutral speech mode.
- Shouted - Shouted (**S-S**): Shouted speech is used in both training and test.
- Neutral - Shouted (**N-S**): Each speaker model is trained using neutral speech and tested with shouted speech.
- Shouted - Neutral (**S-N**): Each speaker model is trained using shouted speech and tested with neutral speech

As the performance criterion, we use identification accuracy, which is the ratio of the correctly identified trials to the total number of trials.

[2] `http://pascal.kgw.tu-berlin.de/emodb/`

Table 1. Identification accuracy (%) for different speech modes using feature mapping

Training-Test condition	Baseline: no compensation		Compensation applied to MFCCs		MFCCs+Δ+Δ^2	
	FFT	LP	FFT	LP	FFT	LP
N-N	100.00	99.81	86.55	89.96	94.50	75.37
S-S	99.43	99.24	91.47	92.61	96.96	89.58
N-S	8.71	18.56	25.37	26.32	32.00	28.40
S-N	22.15	27.65	24.43	29.35	30.87	33.90

5. EXPERIMENTAL RESULTS

We first analyze the performance of the baseline speaker identification system without any feature compensations. The identification accuracy for different scenarios and with different features are provided as the first two columns of Table 1. In the matched vocal mode cases (N-N and S-S), both the FFT and LP spectrum estimators yield high identification accuracies. In the case of the mismatched vocal mode cases (N-S and S-N), both methods degrade to unusable levels which confirms the general observation on previous studies on the topic. In the mismatched cases, LP outperforms FFT.

We next evaluated the shout compensation technique described in Section 3 using different number of Gaussian components. Fig. 4 shows the identification rates for the N-S and the S-N conditions using FFT spectrum estimator. Feature mapping improves the identification rates considerably in comparison to the uncompensated baseline system. Comparing the two types of feature mappings, mapping the full front-end (MFCC + Δ + Δ^2) works generally slightly better. This might be because the full front-end presents richer feature space and directly compensates also for the cepstral dynamics. Regarding the number of Gaussians, single Gaussian is not enough as expected. Using 32 Gaussians yields the highest identification accuracy for both the N-S and the S-N conditions.

Identification rates using feature mapping are given in Table 1. Feature mapping improves recognition accuracies for mismatched modes (N-S and S-N) by a wide margin whereas identification rates decreases in comparison to the uncompensated baseline on the matched conditions (N-N and S-S). However, these relative degradations on N-N (5.5 %) and S-S conditions (2.48 %) are acceptable, given that the mismatched vocal modes experience impressive improvements (for instance, around 4-fold increase for FFT in the N-S condition). In contrast to baseline performances, now FFT outperforms LP in most cases.

Finally, the number of misidentified trials are given in Fig. 5 for FFT features before and after compensation. In the case of no compensation (baseline MFCC) the errors are uniformly distributed and for most speakers all the 24 trials are misidentified. However, the compensation reduces the number of errors almost for every speaker. Fig. 5 reveals that, while shout compensation is successful for some speakers (e.g., 11 and 12), it makes no difference for some speakers (e.g. 8, 10, 14 and 19). The rest of the speakers fall in between these two extremes. There are two possible reasons for such behavior. Firstly, the shout compensation mapping training set is both small and language-mismatched with our evaluation data, as no additional parallel Finnish shouted speech corpus was available. Secondly, being trained from a pool of many speakers, the mapping function does statistical averaging that may remove speaker cues in addition to compensating shouting.

6. DISCUSSION

We evaluated text-independent speaker identification using shouted speech and proposed a first step towards explicit shout compensation using joint density GMM mapping. Identification accuracy is reasonable when the training and test conditions are matched but large degradation on the recognition accuracy occurs in the case of mismatched vocal modes. It was shown that this degradation on recognition accuracy can partly be compensated by training feature mapping on the MFCCs. It is important to note that the proposed compensation mapping is speaker-independent and was trained on a different set of speakers – actually even different spoken language, due to the lack of Finnish data to train the mapping function. While the authors of [4] uses a small database which consists of 12 male speakers with total of 48 identification trials, our results are in reasonable agreement with the results of that study. However, the author in [8] reported smaller degradation in shouted case and this is probably because text-dependent speaker identification were considered using a database of 50 speakers (25 male and 25 female speakers) and each speaker trained using 40 utterances (almost two times more than our training data) and identification experiments carried out with 1600 neutral and 3600 shouted identification trials whereas in this study we have 528 identification trials.

The results for N-N and S-S in Table 1 after applying the transformation reveals that the proposed transformation is smoothing out some speaker specific information from MFCCs. This is also seen from Fig. 3 where, by applying the transformation, most of the MFCC fluctuations are softened for both neutral and shouted speech. On the other hand, the reason for improved recognition accuracy in N-S and S-N condition after applying the proposed transformation could be also found in reduced mismatch between neutral and shouted MFCCs as can be seen from the second row of Fig. 3.

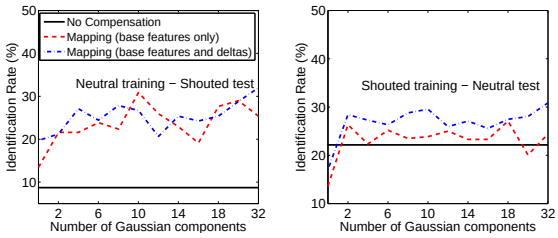

Fig. 4. Identification rates for different number of Gaussians used for compensation.

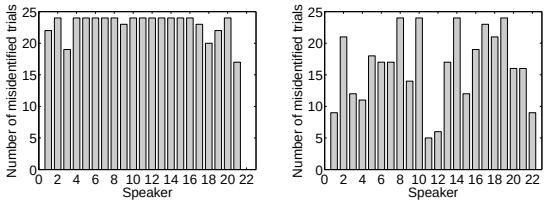

Fig. 5. Number of misidentified trials per each speaker for no compensation (left) and after compensation (right).

7. CONCLUSION

In this paper we evaluated the text-independent speaker identification using shouted speech. Four different training/test conditions have been analyzed and it has been found that recognition performance of speaker identification is quite reasonable when the training and test conditions are matched but large degradation on the recognition accuracy occurs in the case of vocal effort mismatch between training and testing. It was shown that this degradation on recognition accuracy can be partly compensated by applying the feature mapping on the MFCCs. Future work should address how such mapping could be trained ensuring that speaker features are retained.

8. REFERENCES

[1] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey 2001*, 2001, pp. 213–218.

[2] P. Kenny, "Joint factor analysis of speaker and session variability : theory and algorithms," *Technical Report*, 2005.

[3] R. Auckenthaler, "Score normalization for text-independent speaker verification systems," *Dig. Sig. Proc.*, vol. 10, no. 1-3, pp. 42–54, 2000.

[4] C. Zhang and J. H. L. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *Proc. Interspeech 2007*, 2007, pp. 2289–2292.

[5] J. Elliot, "Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics," in *Proc. Aust. Int. Conf. Speech Sci. and Tech.*, 2000, pp. 154–159.

[6] H. Nanjo, T. Nishiura, and H. Kawano, "Acoustic-based security system:towards robust understanding of emergency shout," in *Proc. Int. Conf. Inf. Assurance and Sec.*, 2009, pp. 725–728.

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

[8] I. Shahin, "Employing second-order circular suprasegmental hidden Markov models to enhance speaker identification performance in shouted talking environments," *EURASIP J. on Audio, Speech and Music Proc.*, p. 10 pages, 2010.

[9] 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.

[10] Y. Stylianou, "Voice transformation: A survey," in *ICASSP-2009*, 2009, pp. 3585–3588.

[11] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[12] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE T. Speech and Audio Process*, vol. 6, no. 2, pp. 131–142, 1998.

[13] A. Kain and M.W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP' 98*, 1998, vol. 1, pp. 285–288.

[14] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE T. Speech and Audio Process.*, pp. 72–83, 1995.

[15] J. Pohjalainen, P. Alku, and T. Kinnunen, "Shout detection in noise," in *Proc. ICASSP 2011*, 2011, pp. 4968–4971.

[16] F. Burkhard, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech 2005*, 2005, pp. 1517–1520.

[17] E. Helander, J. Schwarz, J. Nurminen, H. Silen, and M. Gabbouj, "On the impact of alignment on voice conversion performance," in *Proc. Interspeech 2008*, 2008, pp. 1453–1456.