Feature Normalization and Selection for Robust Speaker State Recognition

Chien-Lin Huang, Yu Tsao, and Chiori Hori

Spoken Language Communication Group,
National Institute of Information and Communications Technology, Kyoto, 619-0288, Japan
chiccoCL@gmail.com, yu.tsao@nict.go.jp, chiori.hori@nict.go.jp

Abstract

Alcohol Language Corpus (ALC) and Sleepy Language Corpus (SLC) with genuine intoxicated and sleepy speech are used in INTERSPEECH 2011 Speaker State Challenge. In this paper, we report a series of experiments on the variety of feature analysis and statistical classifications. In addition, we present a feature analysis of feature normalization and selection for the speaker state recognition from speech signals. In the analysis of speech signals, acoustic features are extracted from the low-level-descriptors and their related functionals generated by openSMILE. To reduce the feature mismatch caused by the variability of speakers and channels, the histogram equalization normalization is used to normalize each feature component. In addition, an eigen feature selection is performed for the discriminative representation. In this representation, the meaningful features in a reduced feature dimension are obtained via subspace projection to eliminate the noise features. Finally, Gaussian mixture models are adopted to classify the emotional states. We conclude that the proposed feature normalization and selection can contribute to the robustness of speaker state recognition in practical application scenarios.

1. Introduction

Speech is the most convenient way for the interaction of human-to-human and human-to-machine. The applications of spoken document retrieval in entertainment, business and education are rapidly growing. The recent attempts include mixed-language speech recognition [1], news broadcasts retrieval and summarization [2], spoken dialog system [3], speaker identification and recognition [4], etc. One of important issue of human-machine interaction is to recognize user's emotional status. The affective computing ability enables the machine to understand human's emotion [2]. The acoustic, linguistic and semantic information has been widely used to recognize the speaker emotion. For example, the studies in the emotion recognition focus on the prosodic features, in particular pitch, duration and intensity etc. [3], [7]. Moreover, the voice quality features, such as Noise-to-Harmonics Ratio, jitter, shimmer, and Mel-frequency cepstral coefficients (MFCC) [8], have been recently found useful to emotion detection.

In [9], a multi-modal emotion recognition system is constructed to extract emotion information from both speech and text input. Six emotion types are classified based on 33 acoustic features and emotional keywords. There are a lot of variability in the speaker state recognition such as the channel mismatch, the variety of speakers, the different ranges of features, and so on. Normalization is the important technique to reduce variability in speech and solve the mismatch problem. For example, the speaker-specific feature warping is used as a means of normalizing acoustic features to overcome the problem of speaker dependency [10]. Speaker normalization on the functional level shows a normalization of each calculated functional feature to a mean of zero and standard deviation of one. Corpus normalization is to eliminate different recording conditions as varying room acoustics, different types of and distances to the microphones [11].

In this study, we propose to incorporate the normalization based on the speaker and feature histogram equalization (HEQ) to reduce the mismatch between speakers and features. Furthermore, an eigen feature selection is applied to reduce the high-dimensional feature vectors to a low-dimensional space, and to project the features to the orthogonal. We conduct the experiments on INTERSPEECH 2011 speaker state challenge dataset. The results show that the incorporating speaker and feature HEQ normalizations and eigen feature selection are effective to recognize the speaker emotion.

The remainder of this paper is organized as follows. Section 2 elucidates the HEQ normalization for speakers and features and the eigen feature selection for speaker state recognition. We describe the experimental setup and report a series of experiments in Section 3. Finally, Section 4 concludes this work.

2. Feature Analysis

Feature normalization and feature selection are two important issues for emotion recognition. In this study, we apply HEQ normalization to remove inter-speaker and inter-feature variability. In addition, we use the dimensionality reduction approach based on the subspace projection to eliminate the noise features and extract meaningful features. Fig. 1. shows the feature analysis process for speaker state recognition.

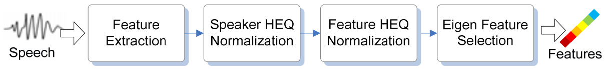


Figure 1. Feature analysis for speaker state recognition.

2.1. Feature Extraction

A set of features has been provided by Schuller et al. [12] and generated by the openSMILE tool [13], in which the feature vector is composed of pitch, energy, zero crossing rate (ZCR) and MFCC, etc. We use this set of features as well as their first order delta coefficients for speaker state recognition. Table 1 gives a summary of the generated features. The low-level-descriptors are for signal energy, loudness, spectra, MFCC, pitch, voice quality, and so on. Moreover, the related functionals are applied to each contour of the low-level-descriptors, such as means, moments, segments, peaks, percentiles, durations, onsets, DCT coefficients, linear and quadratic regression, etc.

2.2. HEQ Normalization

The low-level-descriptors and functionals are extracted from different speakers in feature analysis. In addition, the variety of feature analysis shows different ranges of feature values. Motivated by the variations between different speakers and

Table 1. Overview of low-level-descriptors and functionals.

low-level-descriptors	functionals
audspec_lengthL1norm, audspecRasta_lengthL1norm, RMSenergy, zcr, audSpec_Rfilt(1~26), fband25-650, fband1000-4000, spectralRollOff25.0, spectralRollOff50.0, spectralRollOff75.0, spectralRollOff90.0, spectralFlux, spectralEntropy, spectralVariance, spectralSkewness, spectralKurtosis, spectralSlope, mfcc(1~12), F0final, voicingFinalUnclipped, jitterLocal, jitterDDP, shimmerLocal, F0final_ff0	quartile, iqr, percentile, pctlrange, amean, qmean, stddev, centroid, skewness, kurtosis, meanPeakDist, peakMean, peakMeanMeanDist, peakDistStddev, linregc1, linregerrQ, qregc, qregerrQ, qregc, downleveltime25, upleveltime90, risetime, falltime, leftctime, rightctime, lpgain, lpc, meanSegLen, maxSegLen, minSegLen, segLenStddev, nnz, duration

features, we further the studies by proposing feature normalization using the two dimensional histogram equalizations (speakers and features). It is important to decide consecutive optimization. In this study, we normalize for speakers, subsequently normalize features for each speakers as shown in Fig. 1 and Fig. 2.

The histogram equalization (HEQ) provides a transformation mapping the histogram of each vector component onto a reference histogram for feature normalization [14]. HEQ is commonly used in image processing, speech recognition and speaker recognition. In HEQ, each dimension of the feature vector is viewed as independent. We can estimate the transformation by using the cumulative density function (CDF). In this study, the speech-by-feature matrix $\mathbf{W}=[\mathbf{v}_1,\mathbf{v}_2,...,\mathbf{v}_S]$ is derived from a dataset which is comprised an accumulation of speech. S denotes the number of speech. Each of the columns in $\mathbf{W}$ is a feature vector in K dimensions.

The transform of feature normalization is estimated by $\hat{\mathbf{w}}=HEQ(\mathbf{w}(k))$, where $\mathbf{w}(k)=\mathbf{r}_k=[r_1^k,r_2^k,...,r_s^k]$ means the k-th component of feature vector in speech s. The feature normalization is used to normalize the variety of feature values between different speakers and channels. After the feature normalization, the transform $\tilde{\mathbf{w}}=HEQ(\hat{\mathbf{w}}(s))$ is estimated to normalize the feature values in speech s. Each speech s is represented by a high-dimensional feature vector $\mathbf{w}(s)=\mathbf{c}_s=[c_1^s,c_2^s,...,c_K^s]$ derived from the variety of feature analysis such as pitch, energy, MFCCs and so on. In this study, we use $\mathbf{r}_k$ and $\mathbf{c}_s$ for row and column vectors, respectively. In general, we apply histogram normalization first to the rows (features) and then to the columns of $\mathbf{W}$.

2.3. Dimensionality Reduction

Since the dimension of feature vectors is high and not all the features are extracted robustly, it is not effective to use the original extracted features for the statistical modeling. As a result, we apply the principal component analysis (PCA) to select meaningful features and reduce the feature dimension. The basic idea of PCA is to project data onto the pairwise linear discriminants and take as features. The linear combinations of the discriminants are depicted by the R first eigenvectors and eigenvalues. PCA starts with the singular value decomposition (SVD) [15] which finds the optimal projection as shown in Fig. 2.

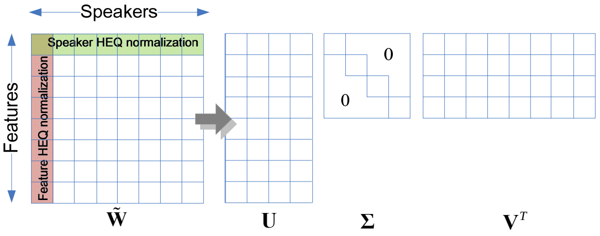


Figure 2. Singular value decomposition.

SVD is related to the eigenvector decomposition and factor analysis. We perform SVD of the matrix $\tilde{\mathbf{W}}$ as follows

$$\tilde{\mathbf{W}}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular matrix, respectively. T denotes the matrix transposition. Both $\mathbf{U}$ and $\mathbf{V}$ show the orthogonal character. $\mathbf{\Sigma}$ is the $K\times K$ diagonal matrix whose nonnegative entries are K singular values in a descending order, i.e., $\sigma_1\geq\sigma_2\geq...\geq\sigma_K>0$. SVD can be used to project all the dimensions of the feature vectors onto a latent information space with significantly reduced dimensionality. The eigenvectors $\hat{\mathbf{U}}$ is treated as a transform basis. The first R eigenvectors $\hat{\mathbf{U}}=[\hat{u}_1,\hat{u}_2,...,\hat{u}_R]$ were empirically selected, where $R\leq K$ denotes the projected dimensions of the original feature vector in eigenspace. The associated eigenvalues allow us to rank the eigenvectors according to their usefulness in characterizing the variation among the features [16]. The transformed feature vector is converted into the orthornormal space as $\hat{\mathbf{v}}=\hat{\mathbf{U}}\times\mathbf{v}$ based on the eigenvector basis $\hat{\mathbf{U}}$. This leads to a meaningful representation in a vector of low dimension space.

3. Speaker State Recognition

With the normalized features from the feature analysis, we use Gaussian mixture models (GMMs) to characterize the speaker state. A log-likelihood ratio (LLR) based evaluation function is applied for testing. Therefore, given an observation $\mathbf{v}$, it is scored against both the positive state model θ_+ and the negative state model θ_- to estimate a LLR score:

$$\Lambda(\mathbf{v})=\ln p(\mathbf{v}\,|\,\theta_+)-\ln p(\mathbf{v}\,|\,\theta_-) \tag{2}$$

If the log-likelihood score is higher than the threshold $\Lambda>\varepsilon$, the claimed speaker state will be accepted, else rejected [17].

We maximize the likelihood of observation feature vectors $\mathbf{v}$, given the GMMs $\theta_\pm$,

$$p(\mathbf{v}\,|\,\theta)=\sum_{m=1}^{M}w_m p(\mathbf{v}\,|\,\lambda_m) \tag{3}$$

where w_m is the weight of Gaussian component, satisfying $\sum_{m=1}^{M}w_m=1$. The mixture model $p(\mathbf{v}\,|\,\lambda_m)$ is a normal distribution $N(\mathbf{v};\mu_m,\Sigma_m)$, with each Gaussian component represented by the parameters $\lambda_m=\{\mu_m,\Sigma_m\}$, where μ_m is the mean vector and Σ_m is the covariance matrix. The iterative

Table 3. Results of various individual features with incorporating speaker and feature HEQ normalizations.

features	dim	ALC				SLC			
		org		norm		org		norm	
		UA	WA	UA	WA	UA	WA	UA	WA
lengthL1norm + Δ	148	55.98%	62.45%	57.57%	63.89%	51.30%	42.16%	57.36%	54.85%
RMSenergy + Δ	74	50.21%	40.15%	54.20%	61.72%	52.75%	43.12%	52.96%	52.08%
zcr + Δ	74	53.10%	41.11%	54.22%	51.44%	53.77%	54.07%	55.59%	49.33%
audSpec_Rfilt (1~26)	962	55.16%	62.27%	58.63%	61.46%	54.37%	49.26%	58.46%	55.88%
fband + Δ	148	51.91%	69.02%	52.72%	63.31%	52.27%	42.50%	53.96%	51.73%
spectralRollOff + Δ	296	52.59%	43.41%	54.71%	60.30%	55.12%	54.75%	53.86%	54.58%
spectralFlux + Δ	74	53.80%	48.89%	54.35%	47.98%	55.51%	48.82%	54.09%	53.07%
spectralEntropy + Δ	74	53.30%	65.35%	55.15%	51.49%	59.99%	56.26%	57.31%	53.76%
spectralVariance + Δ	74	54.84%	57.83%	52.15%	55.23%	53.11%	47.24%	57.67%	51.32%
spectralSkewness + Δ	74	51.93%	63.66%	54.08%	60.33%	54.25%	46.90%	58.49%	53.17%
spectralKurtosis + Δ	74	49.16%	30.15%	53.66%	52.50%	51.52%	41.99%	56.59%	55.33%
spectralSlope + Δ	74	53.96%	50.33%	56.59%	54.67%	57.37%	53.14%	59.76%	57.91%
mfcc (1~12)	444	55.12%	61.41%	53.46%	61.52%	53.81%	48.99%	54.02%	59.31%
F0final + Δ	96	51.90%	57.22%	51.64%	36.82%	51.32%	44.53%	56.73%	60.03%
jitter + Δ	144	53.74%	49.39%	52.07%	54.42%	53.15%	53.21%	53.15%	54.34%
shimmer + Δ	72	50.19%	30.30%	51.39%	42.20%	52.24%	41.78%	52.03%	42.37%
all feature set	4368	58.10%	56.77%	60.60%	63.43%	55.48%	50.53%	60.88%	60.34%
average		53.23%	52.34%	54.54%	55.45%	53.96%	48.19%	56.05%	54.08%

EM algorithm is adopted to estimate the parameters of the Gaussian mixture models, $\theta = \{w_1,..,w_M,\ \lambda_1,..,\lambda_M\}$. In this study, the initialization of GMMs $\theta^{t=0}$ is performed using the *k*-means algorithm with multiple random starting points [18]. Variances were floored to 0.01 of the global variance. A suitable number of Gaussian mixtures *M*=8 is empirically selected for the robust estimation of the speaker state GMMs.

4. Experiments

We use INTERSPEECH 2011 Speaker State Challenge to broaden the scope by addressing two less researched speaker states while focusing on the crucial application domain of security and safety: the computational analysis of intoxication and sleepiness in speech. Main applications are easily found in the medical domain and surveillance in high-risk environments such as driving, steering or controlling. For these Challenge tasks, the Alcohol Language Corpus (ALC) and the Sleepy Language Corpus (SLC) with genuine intoxicated and sleepy speech will be provided by the organizers. ALC consists of 39 hours of speech, stemming from 154 speakers in gender balance, and will serve to evaluate features and algorithms for the estimation of speaker intoxication. SLC features 21 hours of speech recordings of 99 subjects, annotated in 10 different levels of sleepiness. The details can be found in [12]. Two sub-tasks are addressed. In the Intoxication sub-task, alcoholisation of speakers has to be determined as two-class classification task: alcoholised for a blood alcohol concentration exceeding 0.5 or non-alcoholised. In the Sleepiness sub-task, sleepiness of speakers has to be determined by a suited algorithm and acoustic features. While the annotation provides sleepiness in ten levels, only two classes have to be recognized accordingly: sleepiness for a level exceeding level seven.

4.1. Evaluation Metrics

Table 2 shows the number of recording files in ALC and SLC including TRAIN, DEVEL and TEST data sets. All sets are gender balanced. The class labels of "AL" and "NAL" denote alcoholized and non-alcoholized. The class labels of "SL" and "NSL" are sleepiness and non-sleepiness, respectively. We report results obtained on the development data set by unweighted and weighted accuracy on average per class (UA/WA, weighting with respect to number of instances per class) [12]. Because the distribution among classes is not balanced, the competition measure is UA as earlier stated.

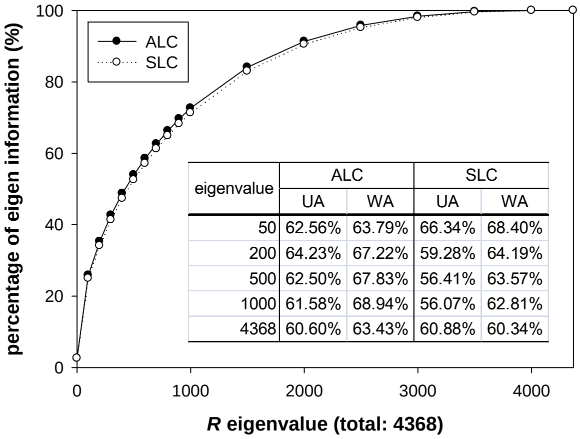
Table 2. Number of recording files in ALC and SLC.

	ALC		SLC	
	AL	NAL	SL	NSL
TRAIN	1650	3750	1241	2125
DEVEL	1170	2790	1079	1836
TEST	1380	1620	851	1957

4.2. Speaker and Feature HEQ Normalizations

Experiments were conducted on ALC and SLC to decide true or false of intoxication and sleepiness in speech. In order to evaluate the performance of individual features and the incorporating speaker and feature HEQ normalizations, we selected 16 subsets of acoustic features for experiments. The comparison results were illustrated in Table 3.

The testing is done on the development set. The term "dim" means the dimension of features. "org" indicates the original features without incorporating speaker and feature HEQ normalizations. "norm" is features with incorporating speaker and feature HEQ normalizations. We can find that the individual features of lengthL1norm, audSpec_Rfilt, spectralEntropy, spectralSkewness, and mfcc showed good recognition accuracy in both SLC and ALC evaluations. In SLC evaluation, the average results indicate there are 2.09% UA and 5.89% WA absolute improvements considering with and without incorporating speaker and feature HEQ normalizations. In ALC evaluation, the average results show



eigenvalue	ALC		SLC	
	UA	WA	UA	WA
50	62.56%	63.79%	66.34%	68.40%
200	64.23%	67.22%	59.28%	64.19%
500	62.50%	67.83%	56.41%	63.57%
1000	61.58%	68.94%	56.07%	62.81%
4368	60.60%	63.43%	60.88%	60.34%

Figure 3. Illustration of eigen feature selection with PCA.

Table 4. Comparison of different classifiers in ALC and SLC.

classifiers	ALC		SLC	
	UA	WA	UA	WA
random	50.28%	50.35%	48.85%	48.92%
SG	64.25%	66.52%	62.98%	65.59%
GMMs	64.23%	67.22%	66.34%	68.40%
K-NN (10)	59.53%	58.26%	63.03%	61.96%
K-NN (100)	62.55%	61.49%	64.09%	61.99%
K-NN (200)	62.25%	61.94%	64.33%	62.44%

there are 1.31% UA and 3.11% WA absolute improvements considering with and without norm.

Although the all feature set is a 4368 high dimensional feature vector in speech, the best result was obtained using all features. The SLC recognition accuracy of all feature set was improved from UA= 55.48% to UA= 60.88%, and from WA= 50.53% to WA= 60.34%. The ALC recognition accuracy of all feature set was improved from UA= 58.10% to UA= 60.60%, and from WA= 56.77% to WA= 63.43%. There is a significant gap between org and norm evaluations.

4.3. Eigen Feature Selection

To reduce the noise features and find a discriminative representation. The acoustic features were analyzed using eigen feature selection with PCA. Figure 3 showed the relation between the eigenvalue and the percentage of eigen information. The eigenvalue spectrum was similar in ALC and SLC. Experiments were conducted to choose the best setting in different combination eigenvalue subspace *R*. The eigenvalues decided on the important principle components were 50 and 200 for SLC and ALC evaluations, which showed a significant dimension reduction. The SLC recognition accuracy with eigen feature selection was improved from UA= 60.88% to UA= 66.34%, and from WA= 60.34% to WA= 68.40%. The ALC recognition accuracy with eigen feature selection was improved from UA= 60.60% to UA= 64.23%, and from WA= 63.43% to WA= 67.22%. The eigen feature selection reduces the redundancy and dimensionality.

4.4. Variety of Classifications

In the past, several classifiers were used for the emotional classification such as neural networks, Bayes classifier, Gaussian mixture models, hidden Markov models, decision trees, K-nearest neighbor, and support vector machines [19]. A number of standard statistical classifiers were evaluated in this study. The parametric classifiers were used to estimate the probability density function (pdf) for feature vectors of each class such as the simple Gaussian (SG) classifier and GMMs [20]. In this study, SG denotes a multidimensional Gaussian distribution.

There is no pdf assumption in the non-parametric classifier. For example, the K-nearest neighbor (K-NN) classifier was used to label each feature vector according to the majority of its K nearest neighbors. Based on the incorporating speaker and feature HEQ normalizations and eigen feature selection, the evaluation of different classifiers was conducted and results were shown in Table 4. As we know, the accuracy of two class classification is about 50% in the random condition. The accuracy of classifiers SG, GMMs and K-NN were similar using the proposed approaches.

5. Conclusion

The feature normalization and selection is studied for an efficient and effective speaker state recognition. The proposed method normalizes static acoustic feature using the transform of histogram equalization considering the variety between speakers and features. We further compute the principal component analysis for the eigen feature selection. It is a representation that allows a useful information extraction from a source. Experiments were conducted on INTERSPEECH 2011 speaker state challenge. There are about 2~10% recognition accuracy absolute improvements by using the feature normalization. With the development dataset, we further improved the accuracy of intoxication and sleepiness recognition by reducing the feature dimension using the eigen feature selection. We conclude that the incorporating speaker and feature HEQ normalizations and eigen feature selection can significantly contribute to the robustness of speaker state recognition of intoxication and sleepiness.

6. References

[1] C.-L. Huang and C.-H. Wu, “Generation of Phonetic Units for Mixed-Language Speech Recognition Based on Acoustic and Contextual Analysis,” *IEEE Trans. on Computers*, vol. 56, no. 9, pp. 1225–1233, 2007.

[2] C.-L. Huang and C.-H. Wu, “Spoken Document Retrieval Using Multi-Level Knowledge and Semantic Verification,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2551–2560, 2007.

[3] J.-R. Ding, C.-L. Huang, J.-K. Lin, J.-F. Yang and C.-H. Wu, “Interactive Multimedia Mirror System Design,” *IEEE Trans. on Consumer Electronics*, vol. 54, no. 3, pp. 972–980, 2008.

[4] C.-L. Huang, H. Su, B. Ma, H. Li, “Speaker Characterization Using Long-Term and Temporal Information,” in *Proc. Interspeech*, pp. 370–373, Makuhari, Japan, 2010.

[5] R. W. Picard, *Affective computing*, The MIT Press, Cambridge, 1997.

[6] D. Cairns and J.H. L. Hansen, “Nonlinear Analysis and Detection of Speech under Stressed Conditions,” *J. Acoustical Soc. Am.*, vol. 96, no. 6, pp. 3392–3400, 1994.

[7] B. Schuller, M. W¨ollmer, F. Eyben, and G. Rigoll, *The Role of Prosody in Affective Speech*, pp. 285–307. Peter Lan Publishing Group, 2009.

[8] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, “The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals,” in *Proc. Interspeech*, ISCA, pp. 2253–2256, 2007.

[9] Z.-J. Chuang and C.-H. Wu, “Multi-Modal Emotion Recognition from Speech and Text” *Computational Linguistic and Chinese Language Processing*, vol. 9, no. 2, pp. 45–62, 2004.

[10] V. Sethu, E. Ambikairajah, and J. Epps, “Speaker normalisation for speech-based emotion detection,” in *Proceedings of the 15th International Conference on Digital Signal Processing (DSP)*, pp. 611–614, 2007.

[11] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies", *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010 .

[12] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, “The Interspeech 2011 Speaker State Challenge,” in *Proc. Interspeech*, ISCA, Florence, Italy, 2011.

[13] F. Eyben, M. Wöllmer, and B. Schuller: “openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor,” in *Proc. ACM Multimedia (MM), ACM*, Firenze, Italy, pp. 25–29, 2010.

[14] A. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Bentez, and A. J. Rubio, “Histogram Equalization of Speech Representation for Robust Speech Recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[15] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[16] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[17] C. McCool, J. Sanchez-Riera, and S. Marcel, “Feature distribution modelling techniques for 3d face verification,” *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1324–1330, 2010.

[18] G. Tzanetakis, and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[19] B. Schuller, G. Rigoll, and M. Lang, “Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine-Belief Network Architecture,” in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 577–580, 2004.

[20] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, New York: Wiley, 2000.