On Automatic Voice Casting for Expressive Speech: Speaker Recognition vs. Speech Classification

Nicolas Obin, Axel Roebel, Grégoire Bachman

ON AUTOMATIC VOICE CASTING FOR EXPRESSIVE SPEECH: SPEAKER RECOGNITION VS. SPEECH CLASSIFICATION

Nicolas Obin[1], *Axel Roebel*[1], *Grégoire Bachman*[2] †

[1] IRCAM - STMS 9912 IRCAM-CNRS-UPMC
[2] ExeQuo - The Production & Localization Company
Paris, France

ABSTRACT

This paper presents the first large-scale automatic voice casting system, and explores the adaptation of speaker recognition techniques to measure voice similarities. The proposed system is based on the representation of a voice by classes (e.g., age/gender, voice quality, emotion). First, a multi-label system is used to classify speech into classes. Then, the output probabilities for each class are concatenated to form a vector that represents the vocal signature of a speech recording. Finally, a similarity search is performed on the vocal signatures to determine the set of target actors that are the most similar to a speech recording of a source actor. In a subjective experiment conducted in the real-context of voice casting for video games, the multi-label system clearly outperforms standard speaker recognition systems. This indicates evidence that speech classes successfully capture the principal directions that are used in the perception of voice similarity.

Index Terms : voice casting, voice similarity, speaker recognition, speech classification.

1. INTRODUCTION

Voice casting is used to transfer a video game (or a film) from a source language (typically, American-English and Japanese) to a target language (typically, French, German, Spanish, Mandarin) with a small amount of available voices for each language. In this context, the notion of VOICE SIMILARITY is central : voice casting requires defining a measure of voice similarity that reflects the perception of the similarity between voices : the smaller/larger the distance of a source voice to a target voices is measured, the closer/farther they are perceived. However, the definition of what defines voice similarity is vague. Recently, the role of voice quality in the perception of voice similarity has been suggested [1]. Also, recent research in speaker clustering (speaker content graphs [2, 3], and speech synthesis [4, 5]) addresses to some extent the measurement of speaker/voice similarities.

Accordingly, the objective of a voice casting system differs qualitatively from standard speaker recognition applications [6, 7, 8] : speaker recognition tends to determine similarity measures that are extremely accurate in the local neighbourhood of a speaker - in the sense that a speaker can be authenticated in the presence of close impostor speakers. However, there is no evidence that this similarity measure can be extended to the entire acoustic space. Voice casting raises two challenging issues over speaker recognition systems :

- voice similarity : the measure of voice similarity must reflect the perception of voice similarity (i.e., the principal directions used in the perception of voice similarity) ;
- tags : a semantic representation of speech recordings is additionally desired to tag a database of actors (e.g., in order to query a 45-year-old male with a breathy/tense voice).

Also, the variability in duration of the speech recordings (from 0.1 s to 15 s with an average duration of 5 s) differs significantly from standard speaker recognition applications, and the similarity must be determined from a single - and generally short - speech recording.

The original contribution of this paper is the representation of speech by classes (e.g., age/gender, voice quality, emotion), which will further be used to measure voice similarity - in place of the standard acoustic models used in speaker recognition. First, a multi-label classifier is used to classify the speech recordings into speech classes. Then, the output probabilities to each class are concatenated to form a vector that represents the vocal signature of the speech recording. Finally, the similarity search is performed on the vocal signature vectors in order to determine the set of target actors that are the most similar to a speech recording of a source actor. The main assumption is that the representation of speech by classes (e.g., age/gender, voice quality, emotion) captures the principal directions used in the perception of voice similarity [1]. The proposed system is subjectively compared to a speaker recognition system in the real context of voice casting in video games.

2. SPEECH ANNOTATION

The representation of speech by classes has been widely studied through the literature : from the representation of the physiological characteristics of a speaker (e.g., age and gender), to voice quality [9] and emotions [10]. Also, a large number of studies have investigated the automatic classification of speech into classes. The classification accuracy significantly depends on the class : the classification of the gender of a speaker is extremely accurate (around 90% for adult speakers [11]), the age can be reasonably determined (within 10 years, [12]), while emotion remains an open issue (from 70% for happiness to 90% and 95% for anger and sadness [13]). More recently, the classification of voice quality has been raised as a novel topic in speech classification [14]. The present study assumes that current technologies for the automatic classification of speech are sufficiently accurate to design a voice casting system that is based on the representation of speech by classes.

The representation of speech that has been retained in this study is based on the following constraints : existing research in

†This study was supported by the European FEDER project VOICE4GAMES.

GENERAL DESCRIPTION	CLASS	LABELS
PHYSIOLOGICAL	GENDER	male, female
	AGE	child, teenager, young adult, adult, old, very old
PHONATION	VOICE QUALITY	breathy, creaky, hoarse
	TENSION	relaxed, normal, tensed, pressed
	VOCAL EFFORT	whispered/soft, normal, loud/shouted
TIMBRE	TIMBRE	clear, dark
ARTICULATION	ARTICULATION	hypo, normal, hyper
PROSODY	F0 REGISTER	extreme-low, low, medium, high, extreme-high
	F0 RANGE	flat, normal, extended
	SPEECH RATE	slow, normal, fast
ACTING	ATTITUDE	affirmation, confirmation, exclamation, interrogation, order, other
	EMOTION	tender, excited, happy, neutral, sad, angry, scared, stressed, surprise, other
	SITUATION	action, conversation, information, monologue, other
	ARCHETYPE	announcer, artificial intelligence, basic soldier, brute, commander, hero neutral, old wise, rookie soldier, sensual, suffer, veteran soldier, other

Table 1. Representation of speech used for the automatic voice casting system.

speech, specific needs of professional voice casting operators, and time constraints related to the large-scale annotation of speech. The final representation is decomposed into speech dimensions (e.g., physiological, phonation, acting), speech classes (e.g., for phonation : breathy, creaky, hoarse, tension), and speech labels (e.g., for tension : relaxed, normal, tense, pressed). The final representation includes : 6 dimensions, 14 classes, and 68 labels. The exhaustive glossary for the representation of speech is presented in table 1.

The annotation consisted of the definition of guidelines and the training of a naive annotator. The guidelines were defined by 2 experts in speech technologies, which include : definitions for each class and label, and speech examples representative of each class and label. First, pilot annotations were conducted on small sets of speech recordings (around 50-100) by the naive annotator and the two expert annotators, until the naive annotator presents a sufficiently satisfactory agreement with the expert annotators. Then, the large-scale annotation was conducted on a selection of 4,000 speech recordings extracted from the 20,000 speech recordings of the French version of the MASS EFFECT 3 video game, covering 54 speakers interpreting 500 roles, with a maximum of 10 speech recordings for each role.

3. SPEAKER RECOGNITION : PARADIGMS

This section summarizes the main paradigms of speaker recognition, which will be further used for the training of the speaker recognition and the multi-label speech classification systems. All systems are based on the IRCAMCLASSIFIER [15] system developed in the context of Music Information Retrieval (MIR) [16, 17]. This system includes the ALIZÉE 3.0 speaker recognition [18] and the LIBSVM [19] SVM libraries.

3.1. Acoustic Space Modeling : Universal Background Model and GMM supervector

The Universal Background Model (UBM) is used to model the distribution of the entire acoustic space [18], which is usually achieved with a standard Gaussian Mixture Model (GMM-UBM). Then, the means parameters of the UBM are adapted to each speech recording by using maximum a posteriori (MAP) adaptation. Finally, each speech recording is represented by the mean vectors of the adapted mixture components :

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M]^T \tag{1}$$

where $\boldsymbol{\mu}$ - referred to as a GMM SUPERVECTOR - is the concatenation of all the mean vectors of the M mixture components.

3.2. Factor Analysis : Total Variability Space and i-vector

An i-vector is the compact representation of a high-dimensional speech recording into a low-dimensional space called Total Variability space [8] - assuming an affine linear model (i.e., factor analysis) :

$$\boldsymbol{\mu} = \boldsymbol{m}_{\boldsymbol{\mu}} + \mathbf{T}\mathbf{x} \tag{2}$$

where : $\boldsymbol{\mu}$ is the adapted GMM-supervector of a speech recording, $\boldsymbol{m}_{\boldsymbol{\mu}}$ is the GMM-supervector corresponding to the UBM mean parameters, $\mathbf{T}$ is the $(M \times p)$ total variability matrix, and $\mathbf{x}$ is a p normally-distributed vector - referred to as an I-VECTOR. The total variability matrix $\mathbf{T}$ is modelled by Maximum-Likelihood (ML) and Expectation-Maximization (EM). The i-vector of a speech recording is determined by MAP adaptation [8].

3.3. Inter-Session Compensation : i-vector Transformation

The i-vector transformation is used to account for the total variability of the high-dimensional acoustic space (i.e., speaker/class information and session/channel information) in a low-dimensional space - in which the i-vectors distribution is assumed to be normal for each speaker/class. In order to compensate explicitly for the session/channel information, and to constrain the i-vector distribution to be normally distributed for each speaker/class, a large number of methods have been proposed : from Linear Discriminant Analysis [8] for inter-session compensation, to Within-Class Covariance Normalization (WCCN, [20]), Length Normalization (L-norm, [21]), Eigen Factor Radial Normalization (EFR, [22]), and Sphere Nuisance Normalization (sphNorm, [22, 23]), depending on the speaker/class.

The Eigen Factor Radial Normalization and Sphere Nuisance Normalization are recursively determined as :

$$\mathbf{x}^{(i+1)} = \frac{\boldsymbol{\Sigma}^{(i)^{-\frac{1}{2}}}(\mathbf{x}^{(\mathbf{i})} - \boldsymbol{\mu}_{\mathbf{x}}^{(i)})}{||\boldsymbol{\Sigma}^{(i)^{-\frac{1}{2}}}(\mathbf{x}^{(\mathbf{i})} - \boldsymbol{\mu}_{\mathbf{x}}^{(i)})||} \tag{3}$$

where : $\boldsymbol{\mu}_{\mathbf{x}}^{(i)}$ and $\boldsymbol{\Sigma}^{(i)}$ denote the total mean vector and covariance matrix at iteration i (EFR), and the within speaker/class mean vector and covariance matrix at iteration i (sphNorm), respectively.

3.4. Scoring

3.4.1. Discriminative Model : SVM

Among the number of classifiers for speaker/class recognition, the Support Vector Machine (SVM) is historically a milestone in speaker recognition [7], and is still popular for speech classification.

For each label, the classification of a vector $\mathbf{x}$ (e.g., supervector, i-vector) corresponding to a speech recording is obtained with regard to the decision function :

$$f(\mathbf{x}) = \sum_{i=1}^{N} \omega_i K(\mathbf{x}, \mathbf{x}_i) + b \ , \qquad f \in [-1, 1] \tag{4}$$

where : $< w_i, \mathbf{x_i}, b >_{i=1}^{N}$ are the parameters of the maximum-margin hyperplane determined during training (respectively : weights, support vectors, and offset), and $K(.,.)$ the SVM kernel [24].

3.4.2. Direct Scoring : Cosine Distance

More recently, direct cosine distance scoring ([25]) has been proved to be extremely efficient for speaker recognition. The cosine distance directly measures the similarity between two speech recordings $\mathbf{x_{src}}$ and $\mathbf{x_{tgt}}$ in the i-vector acoustic space :

$$K(\mathbf{x_{src}}, \mathbf{x_{tgt}}) = \frac{\langle \mathbf{x_{src}}, \mathbf{x_{tgt}} \rangle}{||\mathbf{x_{src}}||\ ||\mathbf{x_{gt}}||} \tag{5}$$

Importantly, the cosine distance considers only the angle between the two i-vectors and not their magnitudes, which are assumed to convey non-speaker information (i.e., session, channel) only.

3.4.3. Generative Model : PLDA

The last advance is the introduction of generative models for speaker recognition [26]. Among them, the Probabilistic Linear Discriminant Analysis (PLDA) [27] is the most popular generative model currently used for speaker recognition. In the original form, PLDA linearly decomposes an i-vector in eigen-speaker and eigen-channel subspaces (respectively of rank N_{speaker} and N_{channel}). In the case where the eigen-channel is assumed to be full-rank ($N_{\text{channel}} = p$), each i-vector $\mathbf{x}$ of a speaker s can be decomposed as [28] :

$$\mathbf{x} = \mathbf{m_x} + \mathbf{\Phi y_s} + \epsilon \tag{6}$$

where : $\mathbf{m_x}$ is the overall mean of the i-vectors, $\mathbf{\Phi}$ is a $(N_{\text{speaker}} \times p)$ eigen-speaker matrix, $\mathbf{y}_s$ is the normally distributed p vector decomposition of the i-vector along the speaker basis $\mathbf{\Phi}$, and ϵ is a p residual vector with a full covariance matrix. Methods for the estimation of the PLDA parameters and scoring are described in [27] and [28].

4. VOICE CASTING : MULTI-LABEL SCORING

The originality of the contribution of this paper is to introduce a representation of speech by classes in place of the scoring directly performed in the acoustic space as for standard speaker recognition. First, the speech recordings are classified into a number of speech classes ; then the resulting classification is used to score the similarity between speech recordings.

To do so, a multi-label system is constructed by converting the classification of multiple labels into multiple binary classifications [16]. First, each label of the speech representation (e.g., the speech recording is creaky) is turned into a binary representation (i.e., yes/no). Then, a classifier is trained for each label separately, which results into C independent classifiers. Each speech recording is then represented by the affinity vector corresponding to the affinity of the speech recording to each label :

$$\mathbf{\Psi} = [\psi_1, \ldots, \psi_C]^T \tag{7}$$

where : ψ_c is the affinity of the observation vector $\mathbf{x}$ to the c-th label. This affinity vector reflects the likelihood of the speech recording to the labels, and is referred to the VOCAL SIGNATURE of the speech recording. Similarly to the GMM SUPERVECTOR and the I-VECTOR, the vector $\mathbf{\Psi}$ representing the VOCAL SIGNATURE of a speech recording is a single vector summarizing each speech recording.

Finally, the similarity of a source to a target speech recording is defined as the distance of their vocal signatures :

$$d(\mathbf{\Psi_{src}}, \mathbf{\Psi_{tgt}}) = \langle \mathbf{\Psi_{src}}, \mathbf{\Psi_{tgt}} \rangle \tag{8}$$

In the context of voice casting, the advantage of the multi-label scoring is double : first, the multi-label scoring can be used to automatically tag speech databases ; secondly, the main difference to speaker recognition systems is that the representation of speakers in the acoustic space is replaced by a representation of speech by a number of classes that are assumed to reflect explicitly the perception of voice similarity.

5. EXPERIMENTS

Two experiments were conducted to compare speaker recognition and multi-label speech classification in the context of voice casting. First, an objective experiment was conducted to determine the parameters of the optimal configurations of the speaker recognition and multi-label systems. Then, a subjective experiment was conducted to compare the optimal speaker recognition and multi-label systems in the real context of voice casting for video games.

5.1. Objective Experiment

The aim of the objective experiment is to determine the optimal configurations for speaker recognition and speech classification that will be further used for the subjective comparison. At this point, no comparison is conducted - but separate optimization, only (one for speaker recognition, one for speech classification).

The objective experiment was conducted on the French version of the MASS EFFECT 3 video game containing 20,000 speech recordings, around 500 roles, around 50 speakers, and around 20 hours of speech of professional actors. A subset of 4,000 speech recordings was used for the annotation of speech classes. Each speech recording were recorded in professional conditions, and encoded into a 48 kHz-16 bits format. The duration of speech recording varies from 0.1 s to 15 s. Speech recordings shorter than 1 s were removed from the speech database.

The front-end processing consisted in the extraction of short-term acoustic features (20 ms. Hanning window with 50% overlapping) : Mel-Frequency Cepstral Coefficients (MFCC, 13 cepstral coefficient determined on 25 Mel-frequency bands). The system setups were defined as follows : N_{GMM} = 8 to 2048 (GMM-UBM), p = 10 to 800 (i-vector), and shared among the speaker recognition and multi-label systems. For the speaker recognition system : N_{LDA} = 10 to 200 (LDA), N_{it} = 1 for EFR (length normalization), N_{it}=3 for sphNorm, N_{speaker} = 10 to 400 and N_{channel} = p (PLDA). For the cosine and PLDA scoring, the scoring was performed by using the mean i-vector of the speaker [23]. For the speech classification, a standard SVM system with a Gaussian kernel [29] was used - in the absence of further studies on the use of cosine and PLDA for speech classification -, and trained

on the subset of manually annotated speech recordings. For EFR and SphNorm : two versions were compared, with (norm) and without (noNorm) the length normalization performed in the equation 3. The experiment was conducted in a form of a 2-fold cross validation. For speaker recognition, the standard Equal Error Rate (EER) was used to measure the performance. For speech classification system, the Balanced Accuracy (B-ACC) - which manages unbalanced classes - was used to measure the performance.

The performance obtained for speaker recognition is presented in table 2. The optimal performance was obtained with the i-vector + sphNorm + PLDA method with the following configuration : 512 GMM (UBM), p = 400 (i-vector), N_{speaker} = 50 and N_{channel} = 400 (full-rank) (PLDA). The speaker recognition performance is proved to be robust to expressive variability of the speaker, and to variability in duration of the speech recordings.

METHOD	EER (%)
i-vector + cosine	4.04
i-vector + LDA/WCCN + cosine	3.02
i-vector + PLDA	2.80
i-vector + EFR + PLDA	2.73
i-vector + sphNorm + PLDA	**2.50**

Table 2. Performance of speaker recognition systems.

The performance obtained for speech classification - for clarity, averaged over all labels and all classes - is presented in table 3. The optimal performance was obtained with the i-vector + EFR (noNorm) + SVM method with the following configuration : 512 GMM (UBM), and p = 50 (i-vector), which is the only transformation method that outperforms the standard i-vector + SVM. In all cases, the i-vector classification significantly outperforms the super-vector classification. Also, the noNorm outperforms the norm method in all cases for speech classification.

METHOD	B-ACC (%)
super-vector + SVM	67.60
i-vector + sphNorm (norm) + SVM	71.96
i-vector + WCCN + SVM	72.29
i-vector + sphNorm (noNorm) + SVM	72.49
i-vector + EFR (norm) + SVM	72.59
i-vector + SVM	72.62
i-vector + EFR (noNorm) + SVM	**73.05**

Table 3. Average performance of speech classification systems.

The optimal configurations were further retained for the subjective comparison of speaker recognition and speech classification systems for voice casting in the real study-case of video games.

5.2. Subjective Experiment

The subjective experiment consisted of the comparison of the 2 optimal systems previously determined in the real condition of voice casting from English-American to French. The English-American (source language) and the French (target language) versions of the MASS EFFECT 3 video game were used for the experiment. First, 50 speech samples were selected from the English-American version : one speech recording for each of the 50 speakers (around 5 sec. in duration, and representative of the speaker). For each source speech sample, the 3 most similar samples were determined in the target speech database for each system. Then, the source speech sample and the 3 target speech samples determined by the 2 systems were presented to the listener. For each source speech sample, the listener was asked to rate the overall similarity of the target speech samples to the source speech sample on a 5 degree scale : very dissimilar, fairly dissimilar, slightly similar, fairly similar, very similar. 30 French native individuals participated in the experiment (20 males/ 10 females, 20-35 years old, same headphones, same professional listening room, paid experiment). The comparison of the 2 methods is presented in figure 1.

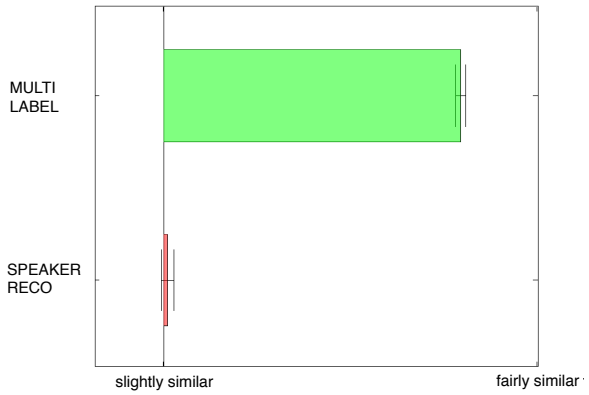

Fig. 1. Mean similarity score and 95% confidence interval for the 2 systems. The score is focused on the +0 (slightly similar) to +1 (fairly similar) interval, on a similarity scale that ranges from -2 (very dissimilar) to +2 (very similar).

The multi-label system significantly outperforms standard speech recognition systems for voice casting. For comparison, the target speech samples determined by the speaker recognition (i-vector + sphNorm + PLDA) and multi-label (i-vector + EFR (noNorm) + SVM) systems are considered as *slightly similar* and *fairly similar* to the source sample in average, respectively. This almost constitutes a one degree difference on the 5 degree scale. These observations support evidence that speech classes successfully capture the principal directions that are used in the perception of voice similarity.

6. CONCLUSION

In this paper, the first large-scale automatic voice casting system was presented, with a main focus on the measurement of voice similarity. The originality of the contribution is to introduce a representation of speech by classes in place of the measurement of voice similarity directly in the acoustic space as performed in standard speaker recognition systems. In a subjective experiment conducted in the real-context of voice casting, the multi-label system clearly outperformed standard speaker recognition systems. Further studies will investigate the used of short-term glottal source [30, 31] and long-term characteristics (prosody) [32, 33] for speaker recognition and speech classification. Also, research will focus on the determination of the principal speech classes used for the measurement of voice similarity, in order to reduce the number of classes required to perform voice casting.

7. REFERENCES

[1] F. Nolan, P. French, K. McDougall, L. Stevens, , and T. Hudson, "The Role of Voice Quality 'Settings' in Perceived Voice

Similarity," in *International Association for Forensic Phonetics and Acoustics*, Vienna, Austria, 2011.

[2] Z. Karam, W. M. Campbell, and N. Dehak, "Graph Relational Features for Speaker Recognition and Mining," in *IEEE Statistical Signal Processing Workshop*, 2011, p. 525–528.

[3] W. M. Campbell and E. Singer, "Query-by-Example using Speaker Content Graphs," in *Interspeech*, Portland, USA, 2012.

[4] R. Dall, C. Veaux, J. Yamagishi, and S. King, "Analysis of Speaker Clustering Strategies for HMM-based Speech Synthesis," in *Interspeech*, Portland, USA, 2012.

[5] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical Parametric Speech Synthesis Based on Speaker and Language Factorization," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1713–1724, 2012.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[7] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, , and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[9] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge : Cambridge University Press, 1980.

[10] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal Cues in Emotion Encoding and Decoding," *Motivation and Emotion*, vol. 15, p. 123–148, 1991.

[11] M. Kockmann, L. Burget, and J. ernocky, "Brno University of Technology System for Interspeech 2010 Paralinguistic Challenge," in *Interspeech*, Makuhari, Japan, 2010, pp. 2822–2825.

[12] M. H. Bahari, M. McLaren, H. V. hamme, and D. V. Leeuwen, "Age Estimation from Telephone Speech using i-vectors," in *Interspeech*, Portland, USA, 2012.

[13] A. Hassan and R. I. Damper, "Multi-Class and Hierarchical SVMs for Emotion Recognition," in *Interspeech*, Makuhari, Japan, 2010, pp. 2354–2357.

[14] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating Fuzzy-Input Fuzzy-Output Support Vector Machines for Robust Voice Quality Classification," *Speech Communication*, vol. 27, no. 1, p. 263–287, 2013.

[15] G. Peeters, "A Generic System for Audio Indexing : Application to Speech/Music Segmentation and Music Genre," in *International Conference on Digital Audio Effects*, Bordeaux, France, 2007.

[16] J.-J. Burred and G. Peeters, "An Adaptive System for Music Classification and Tagging," in *International Workshop on Learning the Semantics of Audio Signals*, Graz, Austria, 2009.

[17] C. Charbuillet, D. Tardieu, and G. Peeters, "GMM-Supervector for Content based Music Similarity," in *International Conference on Digital Audio Effects*, Paris, France, 2011, pp. 425–428.

[18] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, "ALIZE/SpkDet : a State-of-the-Art Open Source Software for Speaker Recognition," *Odyssey : The Speaker and Language Recognition Workshop*, 2008.

[19] C.-C. Chang and C.-J. Lin, "LIBSVM : a Library for Support Vector Machines," 2001.

[20] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-Class Covariance Normalization for SVM-based Speaker Recognition," in *International Conference on Spoken Language Processing*, Pittsburgh, USA, 2006, pp. 1471–1474.

[21] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech 2011*, Florence, Italy, 2011, p. 249–252.

[22] P.-M. Bousquet, A. Larcher, D. Matrouf, J.-F. Bonastre, and O. Plchot, "Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis," in *Odyssey : The Speaker and Language Recognition Workshop*, Singapore, Singapore, 2012, pp. 157–164.

[23] A. Larcher, K. A. Lee, and H. L. Bin Ma, "Phonetically-Constrained PLDA modeling for Text-Dependent Speaker Verification with Multiple Short Utterances," in *International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013, pp. 7673–7677.

[24] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[25] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification," in *Interspeech*, 2009, pp. 4237–4240.

[26] N. Dehak, "Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling : Application to Speaker Verification," PhD. Thesis, Ecole de Technologie Supérieure, 2009.

[27] S. J. D. Prince and J. H. Elder, "Probabilistic Linear Discriminant Analysis for Inferences about Identity," in *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, 2007, pp. 1751–1758.

[28] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey : The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

[29] N. Dehak and G. Chollet, "Support Vector GMMs for Speaker Verification," in *Odyssey : The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, pp. 1–4.

[30] N. Obin, "Cries and Whispers - Classification of Vocal Effort in Expressive Speech," in *Interspeech*, Portland, USA, 2012.

[31] N. Obin and M. Liuni, "On the Generalization of Shannon Entropy for Speech Recognition," in *IEEE workshop on Spoken Language Technology*, Miami, USA, 2012.

[32] N. Obin, "MeLos : Analysis and Modelling of Speech Prosody and Speaking Style," PhD. Thesis, Ircam - UPMC, 2011.

[33] N.Obin, F. Lamare, and A. Roebel, "Syll-O-Matic : an Adaptive Time-Frequency Representation for the Segmentation of Speech into Syllables," in *International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013.