Similarity Search of Acted Voices for Automatic Voice Casting

Nicolas Obin, Axel Roebel

Similarity Search of Acted Voices for Automatic Voice Casting

Nicolas Obin, *Member, IEEE*, and Axel Roebel, *Member, IEEE*

***Abstract*—This paper presents a large-scale similarity search of professionally acted voices for computer-aided voice casting. The proposed voice casting system explores Gaussian mixture model-based acoustic models and multilabel recognition of perceived paralinguistic content (speaker states and speaker traits, e.g., age/gender, voice quality, emotion) for the voice casting of professionally acted voices. First, acoustic models (universal background model, super-vector, i-vector) are constructed to model the acoustic space of voices, from which the similarity between voices can be measured directly in the acoustic space. Second, multiple binary classification of speaker traits and states is added to the acoustic models in order to represent the vocal signature of a voice, which is then used to measure the similarity between voices in the paralinguistic space. Finally, a similarity search is processed in order to determine the set of target actors that are the most similar to the voice of a source actor. In a subjective experiment conducted in the real-context of cross-language voice casting, the multilabel scoring system significantly outperforms the acoustic scoring system. This constitutes a proof of concept for the role of perceived para-linguistic categories in the perception of voice similarity.**

***Index Terms*—Multi-label classification, para-linguistics, speaker recognition, speaker traits and states, voice casting, voice similarity.**

I. Introduction

THE production of multi-media content (films, series, video-games) available to various countries requires the translation of the speech content from a source language (typically, English) to a set of target languages (typically, French, German, Spanish, Japanese, Mandarin). Translation of the speech content can be simply obtained by subtitles, but very often the original speech content is totally replaced by the corresponding speech content in the target language. This process, referred to as dubbing, is obtained by first translating the text from the source to the target language, then selecting actors in the target language, and finally recording the actors synced to the original speech content. These actors must be selected so as to preserve as much as possible the voice and the acting of the original actors. Voice casting denotes the selection of a voice in a target language that is the most similar to a voice in a source language, and is usually performed by human experts who manually select actors according to a database of available voices in the target language. Beyond voice casting stands the open scientific issue on the perception and the measurement of voice similarity: the closer/farther a source voice is perceived from a target voice, the smaller/larger the distance should be measured. What defines the perception of voice similarity remains vague: common expressions (i.e., gender: male/female, age: young/old) are generally used to describe the main traits of a voice/speaker [1], and the role of voice quality in the perception of voice similarity has been recently addressed [2]. Also, some recent research in speaker clustering (e.g., speech retrieval [3], [4], and speech synthesis [5], [6]) have addressed to some extent the measurement of speaker/voice similarities.

To the best of our knowledge, this paper is the first scientific investigation into the measurement of voice similarity for the voice casting of professionally acted voices. Two alternative solutions are investigated and compared:

- Intuitively, the use of speaker recognition techniques [7]–[9] for voice casting appears seducing: the scoring used in speaker recognition system can be interpreted straightforwardly as a similarity measure between voices, and this similarity can be directly measured in the acoustic space. In particular, the similarity measure as determined for speaker recognition has been proven to be extremely accurate in the local acoustic neighbourhood of a speaker: a speaker can be authenticated in the presence of close impostor speakers. However, there is no evidence that this similarity measure remains valid in the entire acoustic space, and actually reflects the perception of the similarity between voices.
- Alternatively, the description of a voice by perceived paralinguistic categories (speaker states and traits [10], [11] e.g., age/gender, voice quality, emotions) may efficiently capture the perception of a voice, and then serve to measure the similarity between voices. Furthermore, common expressions are widely used by human experts in voice casting to stereotype a role. For instance, Albus Dumbledore from the movie "Harry Potter" can be described as a male, old, wise, and breathy voice. Accordingly, the stereotype may be more important for voice casting than the actual acoustic similarities.

This paper explores and compares the use of GMM-based acoustic models and multi-label classification of perceived paralinguistic categories (e.g., age/gender, voice quality, emotion) for the voice casting of professionally acted voices. This extends the preliminary work presented in [12] by 1) presenting the complete details of the two contributions proposed for voice casting, one is based on acoustic models derived from speaker recognition and one is based on multi-label recognition of

Manuscript received January 19, 2016; revised May 04, 2016 and June 09, 2016; accepted June 09, 2016. Date of publication; date of current version. This work was supported by the European FEDER project VOICE4GAMES (2011–2014). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomi Kinnunen.

The authors are with the Centre National de la Recherche Scientifique, Institute for Research and Coordination in Acoustics and Music, University of Pierre and Marie Curie-Sorbonne Universités, Paris 75004, France (e-mail: nobin@ircam.fr; Axel.Roebel@ircam.fr).

Digital Object Identifier 10.1109/TASLP.2016.2580302

perceived para-linguistic categories (speaker traits and speaker states); and 2) providing a detailed evaluation of para-linguistic recognition for all considered categories, including a comparison of MFCC, super-vector, and i-vector acoustic representations for some classic and novel para-linguistic categories. First, GMM-based acoustic models and multi-label classification of speaker traits and states are presented in order to score the similarity between voices in the acoustic and para-linguistic spaces (Section II). Second, the set of categories used to describe the traits and states of a speaker for the multi-label scoring is presented (Section III). The performance of the acoustic and the multi-label scoring is first evaluated in an objective experiment, and then compared in a subjective experiment in the real context of professional voice casting (Section IV).

II. Similarity Scoring for Voice Casting

This section presents the details of the acoustic models and scoring derived from speaker recognition, and the multi-label classification and scoring from perceived para-linguistic categories for voice casting. The implementation is based on ircamClassifier [13], a system developed in the context of Music Information Retrieval [14], [15]. This system includes the Alizée 3.0 speaker recognition [16] and the LibSvm [17] SVM libraries.

A. Acoustic Space Modeling: Universal Background Model (UBM) and GMM supervector

The UBM is used to model the distribution of the entire acoustic space [7]. This modelling is usually achieved with a standard Gaussian mixture model (GMM-UBM). The likelihood of the $(D \times 1)$ feature vector $\mathbf{o}$ describing the acoustic characteristics of speech is defined as

$$p(\mathbf{o}|\boldsymbol{\lambda}) = \sum_{i=1}^{M} \alpha_i p_i(\mathbf{o}) \tag{1}$$

where M is the number of mixture components, $\boldsymbol{\lambda} = \{\boldsymbol{\alpha}_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i\in[1,M]}$ represents the weights, means, and variances of the i-th Gaussian, and $p_i(\mathbf{o}) = \mathcal{N}(\mathbf{o}|\boldsymbol{\mu}_\mathbf{i}, \boldsymbol{\Sigma}_\mathbf{i})$ denotes the likelihood of the ith mixture component, where $\mathcal{N}$ denotes a Gaussian distribution.

Then, the mean parameters $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M\}$ of the UBM are adapted to each speech recording by using relevance maximum a posteriori (MAP) adaptation [7]. This is achieved by updating the means of the mixture components to a sequence of acoustic observations $\mathbf{o} = [\mathbf{o_1}, \ldots, \mathbf{o_T}]$, of length T. Finally, each speech recording is represented by the mean vectors of the adapted mixture components:

$$\boldsymbol{\mu}^{\text{adapt}} = [\boldsymbol{\mu}_1^{\text{adapt}\top}, \ldots, \boldsymbol{\mu}_M^{\text{adapt}\top}]^\top \tag{2}$$

where $\top$ denotes the transposition operator, and $\boldsymbol{\mu}^{\text{adapt}}$, referred to as a GMM-supervector, is the concatenation of all the mean vectors of the adapted mean parameters of the UBM.

B. 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 [9], assuming an affine linear model (i.e., factor analysis):

$$\boldsymbol{\mu}' = \boldsymbol{\mu} + \mathbf{T}\mathbf{x} \tag{3}$$

where $\boldsymbol{\mu}'$ is the GMM-supervector of a speech recording, $\boldsymbol{\mu}$ is the GMM-supervector corresponding to the UBM mean parameters, $\mathbf{T}$ is the $(DM \times q)$ total variability matrix, and $\mathbf{x}$ is a q-dimensional vector assuming a prior normal distribution, referred to as an i-vector. The total variability matrix $\mathbf{T}$ is estimated by expectation-maximization [9]. The i-vector of a speech recording is determined as a MAP point estimate of the latent variable $\mathbf{x}$ [9].

C. Inter-Session Compensation: i-vector Transformation

The i-vector transformation is used to project 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 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 [9] (LDA) for inter-session compensation, to within-class covariance normalization (WCCN, [18]), length normalization (LN, [19]), eigen factor radial normalization (EFR, [20]), and sphere nuisance normalization (SN, [16], [20]) for speaker/class normalization.

The LN is a simple normalization:

$$\mathbf{x} = \frac{\mathbf{x}}{||\mathbf{x}||} \tag{4}$$

where $||.||$ denotes the L-2 norm.

The WCCN whitens the covariance matrix of each class:

$$\widehat{\mathbf{x}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{x} \tag{5}$$

$$\mathbf{W} = \sum_{k=1}^{K} p(k)\boldsymbol{\Sigma}_\mathbf{x}^{(k)} \tag{6}$$

where $\mathbf{W}$ is the covariance matrix defined as the weighted sum of the within-class covariance matrices $\boldsymbol{\Sigma}_\mathbf{x}^{(k)}$, where K is the number of classes, and $p(k)$ is the prior probability of class k.

The EFR processes recursively standardization and normalization:

$$\mathbf{x}^{(i+1)} = \frac{\boldsymbol{\Sigma}_\mathbf{x}^{(i)^{-\frac{1}{2}}}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_\mathbf{x}^{(i)})}{||\boldsymbol{\Sigma}_\mathbf{x}^{(i)^{-\frac{1}{2}}}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_\mathbf{x}^{(i)})||} \tag{7}$$

where $\boldsymbol{\mu}_\mathbf{x}^{(i)}$ and $\boldsymbol{\Sigma}_\mathbf{x}^{(i)}$ denote the mean vector and covariance matrix of all i-vectors at iteration i.

The SN is similar to the EFR, except that the covariance matrix is replaced by the within-class covariance matrix $\mathbf{W}$.

D. Acoustic Scoring

The first contribution of this paper is to investigate the acoustic scoring derived from speaker recognition to measure the similarity between voices for voice casting (Fig. 1). While the

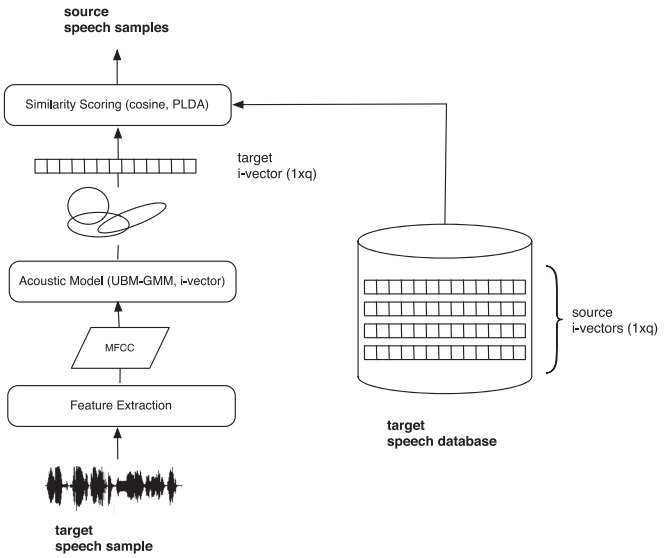

Fig. 1. Architecture of the acoustic voice casting system. On bottom, unsupervised acoustic extraction and modelling; on top, supervised/unsupervised acoustic scoring.

support vector machine (SVM) is historically a milestone in speaker recognition [8], some recent advances have been proposed to score the similarity between speakers.

1) Direct Scoring: Cosine Similarity: First, direct cosine similarity [21] has been proven to be extremely efficient for speaker recognition. The cosine similarity measures the similarity between two speech recordings $\mathbf{x}_{\mathrm{src}}$ and $\mathbf{x}_{\mathrm{tgt}}$ in the i-vector acoustic space:

$$s(\mathbf{x}_{\mathrm{src}}, \mathbf{x}_{\mathrm{tgt}}) = \frac{\langle \mathbf{x}_{\mathrm{src}}, \mathbf{x}_{\mathrm{tgt}} \rangle}{||\mathbf{x}_{\mathrm{src}}||\ ||\mathbf{x}_{\mathrm{gt}}||} \tag{8}$$

where $< ., . >$ is the scalar product operator.

Importantly, the cosine similarity assumes that only the angle between two i-vectors provides information about the similarity between speech recordings. Furthermore, the cosine similarity can be computed directly in the acoustic space, without any prior training.

2) Generative Model: PLDA: One of the last advances is the introduction of generative models for speaker recognition [22]. Among them, the probabilistic linear discriminant analysis (PLDA) [23] 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 ($N_{\mathrm{channel}} = q$) (Gaussian PLDA [24] or simplified PLDA [19]), each i-vector $\mathbf{x}_s$ of a speaker s can be expressed as

$$\mathbf{x}_s = \mu_{\mathbf{x}} + \mathbf{S}\mathbf{h}_s + \epsilon \tag{9}$$

where $\boldsymbol{\mu}_{\mathbf{x}}$ is the total i-vectors mean vector, $\mathbf{S}$ is the $(N_{\mathrm{speaker}} \times q)$ eigen-speaker matrix, $\mathbf{h}_s$ is the position of the i-vector within the eigen-speaker space $\mathbf{S}$ (the latent speaker vector, assumed to be normally distributed), and ϵ is the q residual vector with a full covariance matrix. Maximum-Likelihood (ML) estimation of the PLDA parameters is described in [23].

Then, the similarity between two speech recordings $\mathbf{x}_{\mathrm{src}}$ and $\mathbf{x}_{\mathrm{tgt}}$ can be computed as the likelihood ratio [24]:

$$s(\mathbf{x}_{\mathrm{src}}, \mathbf{x}_{\mathrm{tgt}}) = \frac{p(\mathbf{x}_{\mathrm{src}}, \mathbf{x}_{\mathrm{tgt}}|\mathcal{H}_1)}{p(\mathbf{x}_{\mathrm{src}}|\mathcal{H}_0)\ p(\mathbf{x}_{\mathrm{tgt}}|\mathcal{H}_0)} \tag{10}$$

where the hypothesis $\mathcal{H}_1$ indicates that both vectors come from the same latent speaker (respectively, class), and $\mathcal{H}_0$ indicates they come from different latent speakers (respectively, classes). A closed form solution can be computed as detailed in [24], [25].

The acoustic scores derived from speaker recognition can be straightforwardly turned into a similarity measure between two speech recordings for voice casting, by ignoring the identification, recognition, verification hard decision of a speaker identity. The main advantages of the acoustic scoring for voice casting is that the scoring can be performed directly in the acoustic space, and the similarity measure has been proven to be extremely efficient for speaker recognition. Also, the similarity metric may be strictly unsupervised (e.g., cosine distance), or supervised by available information (e.g., PLDA). Since the speaker's identity is generally the only available information in the context of voice casting and assuming the accuracy of speaker recognition systems, this first contribution will address the use of a similarity metric supervised by speaker's identity for voice casting.

E. Multi-Label Scoring

The second contribution of this paper is the use of a multi-label scoring based on the "semantic" description of a voice with perceived para-linguistic categories (speaker states and speaker traits). This multi-label scoring is presented as an alternative to the acoustic scoring as used for speaker recognition. First, a multi-label classifier is added on top of GMM-based acoustic models to assign the labels corresponding to a speech recording. Then, the posterior probabilities of each label are concatenated to form a vector that represents the signature of a voice, which is used to measure the similarity between voices (Fig. 2).

Multi-label classification [26] is commonly used for the indexing, retrieval, and similarity search of multi-media content (e.g., [27], [28] for text, [29], [30] for music, [31] for image, and [32] for video). Multi-label classification assumes that a media content (text, image, video, sound) can be described with a set of labels that are independent to each other. Multi-label classification is opposed to the multi-class classification commonly used in para-linguistic classification: the classification does not result to a single label, but to a vector of multiple co-occurring and non-exclusive labels (Fig. 3). In consequence, multiple labels can be assigned to a media content. Typically, the emotion content of a speech recording can be a mix of "sadness" and "fear" and even of "sadness" and "joy," whereas the speech recording would be classified only as "sadness" with multi-class classification. This provides an extended description of a speech content that can be further used for similarity search.

For voice casting, a multi-label scoring is constructed by converting the classification of multiple labels into multiple binary classifications [14], [26]. First, each label of the speech description (e.g., the speech recording is creaky) is turned into a

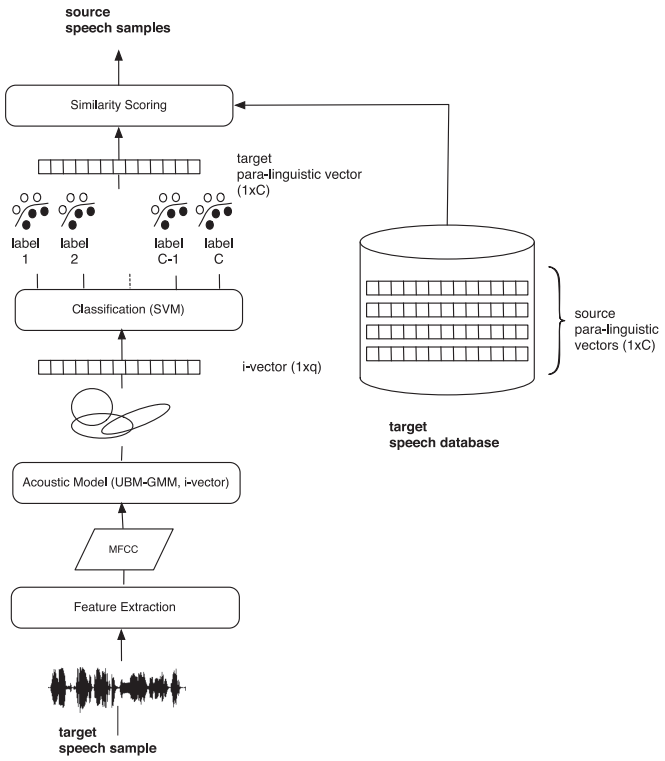

Fig. 2. Architecture of the multi-label voice casting system. On bottom, unsupervised acoustic extraction and modelling; on top, supervised multi-label recognition and scoring.

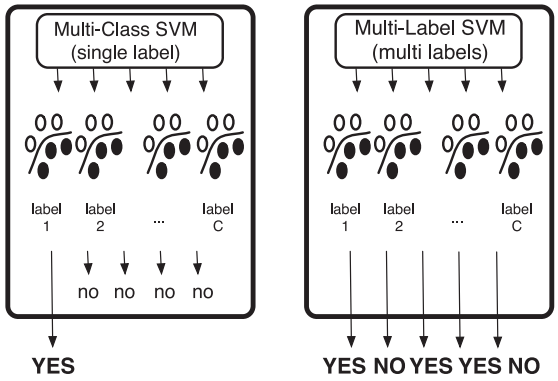

Fig. 3. Multi-class vs. multi-label classification.

binary representation (i.e., yes/no). Then, a classifier is trained for each label separately, which results into C independent one-versus-all classifiers [33]. A complete description of the labels used in this paper is provided in Section III.

Here, the SVM ([17]) is used for multi-label classification. For each label c, 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:

$$f_c(\mathbf{x}) = \sum_{i=1}^{N} \omega_c^i K(\mathbf{x}, \mathbf{x}_c^i) + b_c \tag{11}$$

where $\Theta_c = \{w_c^i, \mathbf{x}_{c^i}, b_c\}_{i=1}^N$ are the parameters of the maximum-margin hyperplane determined during training (respectively weights, support vectors, and offset), and $K(.,.)$ the SVM kernel [34].

In a standard SVM, the binary label corresponding to the observation vector x is assigned with regard to the sign of the decision function:

$$\hat{y}_c = \text{sign}(f_c(\mathbf{x})) \tag{12}$$

where $\text{sign}(x) = +1$ for $x \geq 0$ and -1 otherwise, so that y_c is 1 when the label is positive, and 0 when the label is negative.

Here, the decision function is converted into a posterior probability estimate for each label c, as detailed in [35]:

$$\psi_c = p_c(y = 1|\mathbf{x}) = p(y = 1|\mathbf{x}, \Theta_c), \quad c \in [1, \ldots, C]. \tag{13}$$

Then, the posterior probabilities of each label are concatenated to form a vector that represents the vocal signature signature of a speech recording:

$$\mathbf{\Psi} = [\psi_1, \ldots, \psi_C]^\top \tag{14}$$

where ψ_c is the posterior probability of the c-th label conditionally to the observation vector $\mathbf{x}$. 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 d of their vocal signatures:

$$s(\mathbf{x}_{\text{src}}, \mathbf{x}_{\text{tgt}}) = d(\mathbf{\Psi}_{\text{src}}, \mathbf{\Psi}_{\text{tgt}}) \tag{15}$$

where $\mathbf{\Psi}_{\text{src}}$ and $\mathbf{\Psi}_{\text{src}}$ denotes the vocal signature of the source and target speech recordings, respectively. Here, $d(.,.)$ is defined as the Kullback–Leibler divergence $KL(.||.)$ which is a natural distance measure between posterior probabilities [36], [37].

The main advantage of the multi-label scoring for voice casting lies on the assumption that the para-linguistic content of a voice (speaker traits and states) may reflect more explicitly the *perceived* similarity between voices. Furthermore, the multi-label classification system can be used to automatically tag and search voices based on their perceived para-linguistic content within large speech databases.

III. Description of Speaker Traits and States

This section presents the details of the para-linguistic categories used for the multi-label scoring system presented in Section II-E. A short review on standard description and recognition of a voice content is first provided, followed by a specification of the selected description for voice casting.

Common linguistic expressions can be associated with a voice to describe the perceived "quality" of the voice [1]. Among them, the age (young/old), sex (male/female), and emotions are the most widely used expressions to describe a voice. These expressions can be directly associated with speaker traits and states: speaker traits denote persistent/external traits of a speaker (e.g., personality), and speaker states denote the temporary/internal states of a speaker (e.g., emotions). The definition and the description of speaker traits and states has been widely studied in the literature, from biological speaker traits primitives (e.g., age and gender), individual, social and cultural

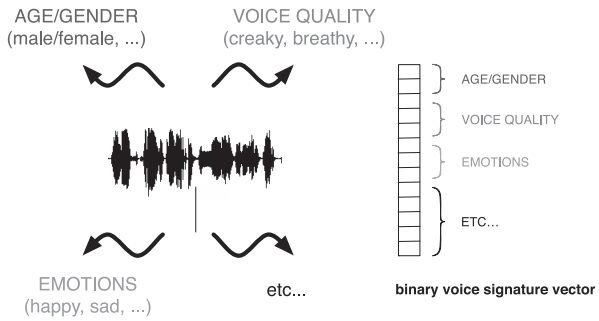

Fig. 4. Multi-label tagging of a speech recording.

speaker traits (e.g., voice quality, [38]), to speaker states (e.g., emotions [39]). Also, research on the automatic recognition of perceived para-linguistics categories in speech has considerably increased over the past few years, pushed by the emergence and needs for human-robot interaction and multi-media retrieval applications (computational paralinguistics challenges: speaker age and gender [40], speaker state [10], [41], speaker traits [11]). The recognition scores obtained for perceived para-linguistics recognition significantly varies depending on the task: the classification of the gender of a speaker is accurate (around 90% for adult speakers [42]), the age can be reasonably determined (within 10 years on telephone speech, [43]), while emotion remains an open issue (from around 80% [44] for acted speech [45], to only 60% for spontaneous speech [46], [47]). More recently, the recognition and modeling of voice quality has raised as a novel topic in para-linguistics recognition [48]–[51].

For voice casting, one must first define a comprehensive set of para-linguistics categories that can be used by expert voice casting operators and exploited for voice similarity search. This set must cover the main traits and states of a speaker, and fulfill specific ad-hoc needs of expert voice casting operators (mostly related to acting and stereotypes). The first set of speaker traits and states comprises standard para-linguistic categories:

- biological speaker traits: sex (male, female), and age (child, teenager, young adult, adult, old, very old);
- speaker state: emotion (tender, excited, happy, neutral, sad, angry, fear, stressed, surprise, other);

A second set comprises categories associated with perceived acoustic characteristics of the voice:

- phonation: voice quality (breathy, creaky, hoarse), tension (relaxed, normal, tensed, pressed), vocal effort (whispered/soft, normal, loud/shouted);
- articulation: articulation (hypo, normal, hyper);
- timbre: timbre (clear, dark);
- prosody: F0 register (extreme-low, low, medium, high, extreme-high), F0 range (flat, normal, extended), and speech rate (slow, normal, fast);

A last set comprises categories associated with the role and the situation of acting:

- attitude/modality: affirmation, confirmation, exclamation, interrogation, order, 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.

The selected description of a speaker includes 14 classes (e.g., gender, age, emotion voice quality), and 68 labels (e.g., for voice quality: breathy, creaky, hoarse). For clarity, the terms "class" and "label" are here used by analogy to multi-class and multi-label classification (see Table II): a class denotes a group containing multiple instances (e.g., the emotion class contains multiple instances: angry, happy, neutral, sad), and a label denote each particular instance (e.g., angry, happy, neutral, sad are labels). The multiple labeling of a speech recording results in a binary vector which represents the voice signature of the speech recording (Fig. 4).

A preliminary phase of manual labeling was conducted in order to train multi-label classifiers for the recognition of speaker traits and speaker states, and to process multi-label scoring for voice casting. Beforehand, a guideline was created to define each class and each label, accompanied by a set of representative speech samples, and a PHP web interface was designed to allow easy and fast on-line annotation of a speech database. The manual labeling was produced by a non-expert individual, preliminary trained by two speech experts (the author, and an expert voice casting operator). First, pilot campaigns were conducted on small sets of speech recordings (around 50-100) by the non-expert annotator and the two expert annotators, until the non-expert annotator presents a sufficiently satisfactory agreement with the expert annotators. The final inter-annotator agreement for coding speaker traits and states has an average Krippendorff's alpha of $\alpha = 0.52$ [52], which represents a fairly reliable agreement regarding the ambiguity and the diversity of the classes considered for labeling. Then, a large-scale annotation was conducted on a selection of 4000 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 (see Section IV for a detailed description of the speech database).

IV. Experiments

Two experiments were conducted to compare acoustic and multi-label similarity scoring in the context of professional voice casting. First, an objective experiment was conducted to determine the parameters of the optimal configurations of the acoustic and multi-label scoring systems. Then, a subjective experiment was conducted to compare the optimal acoustic and multi-label similarity scoring systems in the real context of professional voice casting. For all comparisons, the acoustic and multi-label similarity scoring systems share the same unsupervised acoustic space representation (MFCC, super-vector, i-vector). The systems differ only by the way the similarity measure is constructed: for the acoustic scoring, the similarity metric is defined in the acoustic space, and supervised with respect to the speaker's identity; for the multi-label scoring, the similarity metric is defined in the para-linguistic space, and supervised with respect to each para-linguistic label.

TABLE I
PERFORMANCE OF SPEAKER RECOGNITION SYSTEMS (EER (%))

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**

A. Objective Experiment

The purpose of the objective experiment is to determine the optimal configurations of the acoustic and multi-label scoring systems, in order to select the configurations that will be used for the subjective comparison. Accordingly, the objective experiment is only concerned with separate optimization of the acoustic scoring and multi-label scoring systems. The acoustic scoring system is optimized with respect to a speaker recognition experiment, and the multi-label system is optimized with respect to a perceived para-linguistic classification experiment. Besides optimization, the objective experiment explores the use of advanced acoustic modeling (super-vector, i-vector) for the recognition of a large set of para-linguistic speech categories, which extends preliminary research (for age recognition, [43]), and includes novel para-linguistic categories (e.g., attitude/modality, situation, archetypes).

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 4000 speech recordings was used for the manual annotation of perceived para-linguistic categories. All speech recordings were recorded in professional conditions (professional studio recordings, same recording material, same supervision), and encoded into a 48 kHz-16 bits high-quality 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 (20 ms. Hanning window with 50% overlapping) Mel-frequency cepstral coefficients (MFCC, 13 cepstral coefficient determined with 40 Mel-frequency bands), without delta and delta-delta. The system setups were defined as follows: $N_{\mathrm{GMM}} = 8$ to 2048 (number of GMM-UBM mixture components), $q = 10$ to 800 (dimension of i-vector), and shared among the acoustic scoring and multi-label scoring systems. For the acoustic scoring system, $N_{\mathrm{LDA}} = 10$ to 200 (dimension of LDA reduction), $N_{it} = 1$ for EFR (LN), N_{it}=3 for sphNorm (number of iterations), $N_{\mathrm{speaker}} = 10$ to 400 and $N_{\mathrm{channel}} =$ q (dimension of the speaker and channel spaces for PLDA). For the cosine and PLDA scoring, the scoring was performed by using the mean i-vector of the speaker [16]. For the multi-label scoring, a SVM classifier with a Gaussian kernel [53] was used for binary classification of each label (Fig. 2), each trained on the subset of manually annotated speech recordings.

The experiment was conducted in the form of a 2-fold cross-validation for speaker recognition and 5-fold cross-validation for para-linguistic classification. In k-fold cross-validation, the dataset is first partitioned into k subsets of equal size, then k-1 subsets are used for training the model parameters, and the remaining subset is used for testing the model. This process is repeated for the k folds. The main advantage of cross-validation is the explicit consideration of the performance variability, which can be then be used to assess the statistical difference between different model configurations (acoustic space modeling, inter-session compensation, and scoring). Here, the subsets are constructed by randomly partitioning the available speech recordings regardless of the speakers. For speaker recognition, the standard equal error rate (EER) was used to measure the performance, as determined from the detection error trade-off curve [54] by following the NIST SRE 2012 guidelines [55]. For para-linguistic classification, the balanced accuracy (BA%) is used to measure the recognition performance [56]. The balanced accuracy is the equivalent for binary classification of the unweighted average recall (WA%, [57]) for multi-class classification, a well-established measure for emotion and other para-linguistics recognition ([10], [58], [59]). Indeed, the balanced accuracy is simply defined as the unweighted average of true and false recalls of a binary classification. For one label, it is computed as

$$BA = \frac{R_P + R_N}{2} = \frac{1}{2}\left(\frac{T_P}{T_P + F_N} + \frac{T_N}{F_P + T_N}\right) \quad (16)$$

where R_P and R_N are the positive and negative recalls, and T_P, T_N, F_P, and F_N are the true positive, true negative, false positive, and true negative counts.

The main idea of these measures is to compensate for imbalanced datasets when computing the accuracy score. This is particularly true for para-linguistic binary classification, where the class of interest (the positive one) is highly under-represented as compared to the other (e.g., creaky = yes vs. creaky = no, emotion = sad vs. emotion = not sad, etc...). For each model configuration, the average balanced accuracy is obtained by averaging the balanced accuracy over all folds to compute the per label score, and over all labels to compute the overall score. A grid search is processed to determine the optimal model configuration. First, cross-validation is processed for each model configuration: model parameters are estimated on the training set and the corresponding performance is evaluated on the test set. Then, the hyper-parameters are tuned for each configuration as the one maximizing the overall score.

The performance obtained for speaker recognition is presented in Table I. The optimal performance was obtained with the i-vector + sphNorm + PLDA scoring method with the following configuration, 512 GMM (UBM), $q = 400$ (i-vector), $N_{\mathrm{speaker}} = 50$ and $N_{\mathrm{channel}} = 400$ (full-rank) (PLDA). The speaker recognition performance (EER $= 2.50\%$) indicates the robustness of GMM-based acoustic models to the expressive variability of the speaker, and to the variability in duration of the speech recordings.

The performance obtained for para-linguistic classification is presented in details in Table II (with the exception of 4 "other" labels and 4 other minor labels, only for the sake of space purpose), and in summary with 95% confidence

TABLE II
AVERAGE PERFORMANCE (BA%) OF THE MULTI-LABEL CLASSIFICATION

CLASS	LABEL	MFCC	SUPER-VECTOR					I-VECTOR				
			w/o	WCCN	LN	EFR	SN	w/o	WCCN	LN	EFR	SN
GENDER	MALE	92.77	92.93	93.74	93.63	93.69	93.35	94.04	94.99	94.92	**95.62**	94.69
	FEMALE	92.60	93.52	93.95	93.50	93.47	93.61	94.30	94.96	94.82	**95.24**	94.46
AGE	CHILD	96.40	93.81	95.09	94.94	94.88	94.79	95.78	**96.40**	96.14	96.30	95.50
	TEENAGER	87.03	85.46	87.12	82.24	86.93	86.57	90.08	93.48	91.93	87.56	**94.62**
	YOUNG ADULT	68.95	72.25	73.44	73.16	72.87	73.20	75.10	76.62	75.95	**77.50**	75.48
	ADULT	60.63	64.54	65.61	64.13	63.97	64.84	68.16	69.05	68.49	**69.82**	68.03
	OLD	67.63	70.55	71.93	70.70	70.97	71.22	73.75	74.53	74.74	**75.46**	73.69
	VERY OLD	68.76	67.25	69.41	66.42	67.36	69.14	72.29	73.25	72.59	**75.60**	72.21
VOICE QUALITY	BREATHY	70.03	70.67	72.12	72.44	72.55	71.14	73.61	73.94	74.60	**74.86**	74.33
	CREAKY	73.63	74.22	75.50	75.66	75.40	75.67	76.14	76.55	**78.42**	77.71	75.81
	HOARSE	71.53	73.21	74.44	73.87	74.50	73.65	77.11	76.91	77.53	**77.88**	77.26
TENSION	RELAXED	72.50	71.89	73.13	73.01	72.26	73.34	73.62	74.36	75.57	**76.05**	74.82
	NORMAL	64.92	67.65	68.33	67.78	67.58	67.54	68.56	68.91	69.23	**69.48**	68.60
	TENSED	62.05	62.53	63.78	62.55	63.36	62.52	63.32	64.05	64.37	**64.58**	63.81
	PRESSED	80.57	82.16	83.91	83.43	83.33	83.59	83.60	83.95	84.19	**84.44**	84.03
VOCAL EFFORT	WHISPERED/SOFT	80.24	82.01	83.00	82.64	83.33	82.54	83.27	83.31	83.61	83.28	**83.73**
	NORMAL	68.34	72.50	74.23	72.12	72.08	72.07	73.04	74.47	75.21	**76.24**	73.87
	LOUD/SHOUTED	78.44	77.82	78.35	77.85	77.89	78.63	79.39	80.02	80.95	**81.37**	79.29
ARTICULATION	HYPO	58.76	58.02	60.26	56.51	56.50	58.41	59.92	**61.09**	58.78	59.71	59.82
	NORMAL	57.29	59.75	58.62	58.66	58.04	58.16	58.87	59.48	59.48	**59.78**	59.22
	HYPER	65.11	68.90	**69.99**	67.99	68.71	68.07	68.34	68.53	69.17	69.20	68.95
TIMBRE	CLEAR	69.07	70.58	71.91	70.31	70.41	71.00	72.87	73.16	73.45	73.36	**73.62**
	DARK	69.07	70.46	72.13	70.56	70.45	71.14	72.73	73.18	**73.85**	73.22	73.35
F0 REGISTER	EXTREME-LOW	91.53	90.41	91.17	91.20	91.20	91.04	91.64	92.72	**93.04**	92.50	92.07
	LOW	83.72	85.59	86.25	85.59	86.00	86.14	86.39	86.15	**87.10**	86.50	86.22
	MEDIUM	67.67	70.92	71.87	71.19	71.03	71.28	70.49	71.24	71.50	**72.93**	71.14
	HIGH	72.39	73.86	74.72	74.49	73.95	74.84	73.85	74.13	74.56	**75.23**	74.12
	EXTREME-HIGH	85.83	86.72	88.14	87.57	87.96	88.21	87.57	87.69	**88.43**	87.64	88.08
F0 RANGE	FLAT	65.22	66.72	69.53	67.03	66.62	67.84	68.88	69.45	**70.09**	69.40	68.80
	NORMAL	57.97	57.83	60.44	61.56	59.96	59.83	61.46	60.36	62.44	**62.69**	61.09
	EXTENDED	56.48	61.42	64.44	64.84	63.80	64.65	**68.29**	65.42	66.62	66.82	64.85
SPEECH RATE	SLOW	65.72	68.95	69.04	68.32	69.00	68.35	70.32	71.02	**72.68**	70.27	71.49
	NORMAL	58.31	59.00	59.29	59.38	58.93	58.79	59.09	**61.11**	60.29	60.20	60.05
	FAST	63.67	64.99	64.37	64.69	64.22	65.09	69.49	67.96	**69.94**	68.38	69.57
ATTITUDE/MODALITY	AFFIRMATION	66.89	69.57	69.40	68.93	69.46	69.66	69.64	69.97	70.05	**70.22**	69.78
	CONFIRMATION	60.51	59.27	61.42	61.42	61.60	62.99	61.81	64.45	63.58	**65.49**	61.05
	EXCLAMATION	67.28	68.04	68.99	67.95	68.27	67.95	68.63	68.69	68.83	**69.35**	68.90
	INTERROGATION	58.08	62.31	**62.90**	62.14	61.92	62.86	59.80	58.47	61.67	59.07	59.47
	ORDER	64.95	66.77	68.26	66.66	67.20	65.98	68.71	68.45	**69.60**	68.70	68.87
EMOTION	ANGRY	61.98	62.38	63.36	62.85	62.69	62.58	63.40	64.99	64.56	**65.11**	64.44
	EXCITED	66.18	66.74	67.25	66.79	68.10	67.67	67.04	68.18	68.20	**68.22**	67.41
	HAPPY	53.89	55.81	56.45	55.82	56.71	55.42	58.40	59.24	58.96	**60.72**	59.61
	NEUTRAL	61.42	62.99	64.65	64.05	63.45	64.24	65.75	65.23	65.84	65.69	**66.01**
	SAD	60.65	61.99	63.92	63.05	64.05	63.28	62.68	63.78	64.06	**64.42**	63.15
	FEAR	63.59	63.68	64.98	64.04	64.31	63.78	64.45	63.97	64.93	**66.85**	65.96
	STRESSED	80.22	78.71	80.06	79.52	79.89	79.65	79.59	79.56	80.42	**81.12**	80.09
	SURPRISE	57.74	58.00	59.09	57.74	60.58	59.40	58.01	60.46	60.58	60.27	**60.97**
	TENDER	62.33	62.42	64.51	63.25	63.19	63.87	64.31	64.22	64.64	64.87	**64.87**
SITUATION	ACTION	83.24	81.76	83.26	83.23	82.97	82.63	82.91	82.83	**84.28**	82.60	82.79
	DIALOGUE	73.69	75.79	76.00	75.77	75.60	75.81	77.25	77.03	**78.45**	77.30	77.80
	INFORMATION	69.53	76.22	78.94	77.80	76.90	77.02	78.48	80.22	78.89	**81.73**	80.07
	MONOLOGUE	64.25	65.45	66.06	65.22	66.34	67.99	69.13	66.57	72.54	**73.76**	68.20
ARCHETYPE	ANNOUNCER	82.63	74.67	81.29	83.72	82.74	79.00	87.99	**90.20**	89.65	89.60	87.99
	ART. INTELLIGENCE	87.78	89.34	87.62	85.19	88.52	87.29	90.86	91.56	89.96	**93.24**	91.10
	BASIC SOLDIER	68.71	70.37	71.02	69.22	69.23	70.56	71.73	72.31	**73.29**	72.80	71.53
	BRUTE	73.39	75.10	76.75	76.98	76.36	76.70	78.00	78.94	79.62	**79.74**	78.59
	COMMANDER	64.96	63.99	66.47	66.00	65.31	66.30	68.49	69.58	70.26	**70.58**	70.00
	HERO	68.58	70.67	75.54	70.90	70.90	73.36	77.52	76.40	76.52	**78.11**	76.53
	ROOKIE SOLDIER	69.23	72.54	73.97	74.26	73.78	73.78	76.87	77.36	**79.31**	78.98	76.15
	VETERAN SOLDIER	70.76	70.89	73.15	71.32	71.56	72.08	72.59	73.81	74.14	**74.51**	73.40
TOTAL		70.09	71.71	72.32	71.59	71.69	71.08	73.45	73.75	74.36	**74.62**	73.79

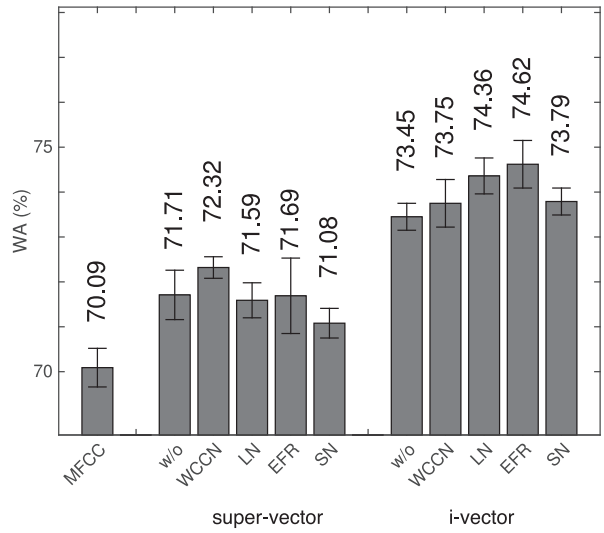

Fig. 5. Overall average recognition score (BA%) and 95% confidence interval obtained for the multi-label classification.

intervals in Fig. 5. The 95% confidence interval is computed by assuming a normal distribution of the cross-validation scores, and is equal to 1.96 times the standard deviation of the cross-validation scores, divided by the square root of the number of folds. The optimal performance was obtained with the i-vector + EFR + SVM method with the following configuration: 512 GMM (UBM), and $q = 50$ (i-vector). In all cases, the i-vector recognition (from 73.45% to 74.62%) has a greater recognition rate than the super-vector recognition (from 71.08% to 72.32%), and the MFCC recognition (70.09%). Also, the inter-session compensation improves the recognition performance, from 71.08% to 72.32% for super-vectors, and from 73.45% to 74.62% for i-vectors. A statistical comparison (one-way ANOVA [60]) shows that the super-vector recognition rate is significantly higher than the MFCC recognition rate ($F(1,18) = 20.52,\ p - value \leq 10^{-4}$), and that the i-vector recognition rate is significantly higher than the super-vector recognition rate ($F(1,18) = 35.20,\ p - value \leq 10^{-5}$). Also, the i-vector EFR recognition rate is significantly higher than the average i-vector recognition rate ($F(1,58) = 6.78,\ p - value \leq 10^{-2}$), and is higher but not significantly with the LN and WCCN recognition rates (respectively, $F(1,18) = 2.52,\ p - value = 0.12$ and $F(1,18) = 4.18,\ p - value = 0.05$). For details, the optimal configuration corresponds to 95.4% for gender, 80.4% for age, 76.8% for voice quality, 73.6% for tension, 80.3% for vocal effort, 73.3% for timbre, 62.9% for articulation, 76.7% for F0 (range and register), 65.6% for speech rate, 66.7% for attitude/modality, 66.4% for emotion, 78.5% for situation, and 78.2% for archetypes. These scores correspond to one single and globally optimal configuration for super-vectors, i-vectors, inter-session compensation, and SVM hyper-parameters in order to figure out a computationally realistic scenario, though all individual performances might be improved through dedicated optimizations. From these observations, some para-linguistic categories can be consistently recognized (age, gender, voice quality, tension, vocal effort, timbre, situation, archetype) while some others remain an open issue (articulation, F0, speech rate, attitude/modality, and emotion). In particular, some novel para-linguistic categories specific to multi-media applications (situation and archetype) are more recognized than some standard para-linguistic categories (attitude/modality, emotions). Also, extreme para-linguistic labels are generally more recognized than standard ones (e.g., normal, medium, neutral) which are more ambiguous. As a conclusion, this generalizes the role of advanced acoustic modeling (i-vector and inter-session compensation) for para-linguistic recognition, as preliminarily reported for age estimation in [43]. Moreover, these constitute encouraging performances for further similarity search for voice casting.

The optimal configurations were further retained for the subjective comparison of acoustic scoring and multi-label scoring systems in the real context of professional voice casting.

B. Subjective Experiment

The real context of voice casting consists in selecting the actors of a target language (e.g., French, German, Spanish, Japanese, Mandarin) whose voice is the most similar to actors of a source language (typically, English). Accordingly, a subjective experiment was conducted in order to address the ability of acoustic and multi-label similarity scores to estimate the perceived similarity between voices for cross-language voice casting. The objective is to compare the role of acoustic and para-linguistic information in the human perception of voice similarity.

The subjective experiment consisted in the comparison of the two optimal configurations previously determined for a voice casting from American-English to French. The American-English (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 American-English version, one speech recording for each of 50 speakers (50% male, 50% female, around 5 sec. in duration). For each source speech sample, the 3 most similar samples were determined in the target speech database for each scoring system. Then, the source speech sample and the 3 target speech samples determined by the 2 scoring 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 (-2), fairly dissimilar (-1), slightly similar (0), fairly similar (+1), very similar (+2). 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 scoring systems is presented in Fig. 6. The multi-label scoring system significantly outperforms the acoustic scoring system in the similarity judgement for voice casting. For comparison, the target speech samples determined by the acoustic scoring (i-vector + sphNorm + PLDA) and the multi-label scoring (i-vector + EFR (noNorm) + SVM) systems are considered as *slightly similar* and *fairly similar* to the source sample in average, respectively. This constitutes almost a one degree difference on the 5 degree scale. A statistical comparison (one-way ANOVA [60]) shows that the multi-label scoring

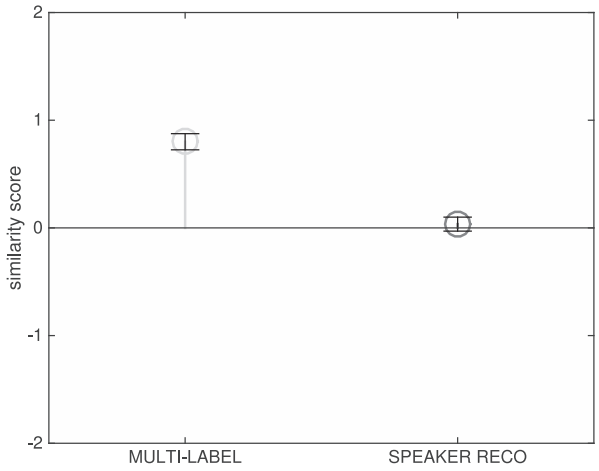

Fig. 6. Mean similarity score and 95% confidence interval for the 2 systems. The similarity scale is very dissimilar (−2), fairly dissimilar (−1), slightly similar (0), fairly similar (+1), very similar (+2).

system is judged as significantly more similar than the acoustic scoring system ($F(1, 298) = 10.86,\ p - value \leq 10^{-3}$).

This experiment provides instructive information about the role of acoustic and para-linguistic information in the perception of voice similarity. First, acoustic information appears necessary but not sufficient to fully capture the perceived similarity between voices. Second, para-linguistic information, as abstractions extracted from the speech content, provides some valuable information about the perception of voice similarity. These observations suggest that the abstraction of a voice into categories (speaker traits and states) play an important role in the human perception of voice similarity, which may prevail over pure acoustic similarity. Beyond, this highlights the role of stereotypes in the human perception of voice similarity, which might be particularly true for professionally acted voices that are generally more stereotyped than everyday speech.

V. Conclusion

In this paper, a large-scale similarity search of voices was presented to measure the perceived similarity between voices for computer-aided voice casting. The proposed voice casting system explored and compared GMM-based acoustic models and multi-label recognition of perceived para-linguistic content (e.g., age/gender, voice quality, emotion) to measure the perceived similarity between voices. In a subjective experiment, the multi-label scoring significantly outperformed acoustic scoring in the real-context of voice casting, which constitutes evidence for the role of perceived para-linguistic content in the perception of voice similarity. This constitutes a preliminary research on voice similarity search for the voice casting of professionally acted voices. Further research will investigate the use of short- and long-term speech characteristics (glottal source [48], [49], prosody [61]) during acoustic modeling, the elaboration of acoustic scoring more specific to voice casting and less constrained by speaker's identity, and the construction of a similarity scoring that covers the entire expressive range of actors instead on being based on the particular expression of a single speech recording. Finally, human experts in voice casting will be added into the subjective evaluation procedure in order to compare the judgements of naive and expert listeners, and to define guidelines for the validation of a voice casting system.

References

[1] H. Kido and H. Kasuya, "Everyday expressions associated with voice quality of normal utterance—Extraction by perceptual evaluation," *J. Acoust. Soc. Jpn.*, vol. 57, no. 5, pp. 337–344, 2001.

[2] F. Nolan, P. French, K. McDougall, L. Stevens, and T. Hudson, "The role of voice quality 'settings' in perceived voice similarity," presented at the International Association Forensic Phonetics Acoustics, Vienna, Austria, 2011.

[3] Z. Karam, W. M. Campbell, and N. Dehak, "Graph relational features for speaker recognition and mining," in *Proc. IEEE Statist. Signal Process. Workshop*, 2011, pp. 525–528.

[4] W. M. Campbell and E. Singer, "Query-by-example using speaker content graphs," presented at the Interspeech, Portland, OR, USA, 2012.

[5] R. Dall, C. Veaux, J. Yamagishi, and S. King, "Analysis of speaker clustering strategies for HMM-based speech synthesis," presented at the Interspeech, Portland, OR, USA, 2012.

[6] H. Zen *et al.*, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1713–1724, Aug. 2012.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, 2000.

[8] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[9] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[10] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," presented at the Interspeech, Florence, Italy, 2012.

[11] B. Schuller *et al.*, "The INTERSPEECH 2012 speaker trait challenge," presented at the Interspeech, Portland, Oregon, USA, 2012.

[12] N. Obin, A. Roebel, and G. Bachman, "On automatic voice casting for expressive speech: Speaker recognition vs. speech classification," presented at the International Conf. Acoustics, Speech, Signal Processing., Florence, Italy, 2014.

[13] G. Peeters, "A generic system for audio indexing: Application to speech/music segmentation and music genre," presented at the Int. Conf. Digital Audio Effects, Bordeaux, France, 2007.

[14] J.-J. Burred and G. Peeters, "An adaptive system for music classification and tagging," in *Proc. Int. Workshop Learn. Semantics Audio Signals*, Graz, Austria, 2009.

[15] C. Charbuillet, D. Tardieu, and G. Peeters, "GMM-supervector for content based music similarity," in *Proc. Int. Conf. Digital Audio Effects*, Paris, France, 2011, pp. 425–428.

[16] A. Larcher *et al.*, "ALIZE 3.0—Open source toolkit for state-of-the-art speaker recognition," presented at the Interspeech, Lyon, France, 2013.

[17] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Sust. Technol.*, vol. 2, 2001, Art. no. 27.

[18] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Pittsburgh, PA, USA, 2006, pp. 1471–1474.

[19] D. Garcia-Romero and C. Espy-Wilson, "Analysis of I-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[20] 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 *Proc. Odyssey: Speaker Lang. Recog. Workshop*, Singapore, Singapore, 2012, pp. 157–164.

[21] 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 *Proc. Interspeech*, 2009, pp. 4237–4240.

[22] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling: Application to speaker verification," Ph.D. dissertation, Ecole de Technologie Supérieure, Montreal, QC, Canada, 2009.

[23] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE Int. Conf. Comput. Vision*, Rio de Janeiro, Brazil, 2007, pp. 1751–1758.
[24] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Bayesian speaker verification with heavy-tailed priors," presented at the Odyssey: Speaker Language Recognition Workshop, Brno, Czech Republic, 2010.
[25] S. Prince, *Computer Vision: Models Learning and Inference*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[26] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowledge Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.
[27] R. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Mach. Learn.*, vol. 39, pp. 135–168, 2000.
[28] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Multilabel text classification for automated tag suggestion," in *Proc. ECML/PKDD Discovery Challenge*, Antwerp, Belgium, 2008, pp. 75–83.
[29] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas, "Multilabel classification of music into emotions," presented at the International Conf. Music Information Retrieval, Philadelphia, PA, USA, 2008.
[30] D. Tardieu, C. Charbuillet, F. Cornu, and G. Peeters, "Mirex-2011 single-label and multi-label classification tasks: Ircamclassification2011 submission," presented at the Music Information Retrieval Evaluation eXchange, Miami, FL, USA, 2011.
[31] M. Boutell, J. Luo, X. Shen, and C. Brown, "Learning multi-label scene classification," *Pattern Recog.*, vol. 37, no. 9, p. 1757–1771, 2004.
[32] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proc. Int. Conf. Multimedia*, New York, NY, USA, 2007, pp. 17–26.
[33] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, 2004.
[34] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[35] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.
[36] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991.
[37] J. V. G. Aradilla and H. Bourlard, "Using posterior-based features in template matching for speech recognition," presented at the Interspeech, Pittsburgh, PA, USA, 2006.
[38] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge, U.K.: Cambridge Univ. Press, 1980.
[39] K. R. Scherer, R. Banse, H. G. Wallbott, and T. Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation Emotion*, vol. 15, pp. 123–148, 1991.
[40] B. Schuller *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," presented at the Interspeech, Makuhari, Japan, 2010.
[41] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," presented at the Interspeech, Brighton, U.K., 2009.
[42] M. Kockmann, L. Burget, and J. H. Cernocky, "Brno University of Technology System for interspeech 2010 paralinguistic challenge," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 2822–2825.
[43] M. H. Bahari, M. McLaren, H. Van Hamme, and D. Van Leeuwen, "Age estimation from telephone speech using i-vectors," presented at the Interspeech, Portland, OR, USA, 2012.
[44] A. Hassan and R. I. Damper, "Multi-class and hierarchical SVMs for emotion recognition," in *Proc. Interspeech*, Makuhari, Japan, 2010, pp. 2354–2357.
[45] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," presented at the Interspeech, Lisbon, Portugal, 2005.
[46] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *J. Lang. Resources Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[47] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 – The First International Audio/Visual Emotion Challenge," in *Proc. Int. HUMAINE Assoc. Conf. Affective Comput. Intell. Interaction*, Memphis, TN, USA, 2011, pp. 415–424.
[48] N. Obin, "Cries and whispers—Classification of vocal effort in expressive speech," presented at the Interspeech, Portland, OR, USA, 2012.
[49] N. Obin and M. Liuni, "On the generalization of Shannon entropy for speech recognition," in *Proc. IEEE Workshop Spoken Lang. Technol.*, Miami, FL, USA, 2012, pp. 97–102.
[50] S. Scherer, J. Kane, C. Gobl, and F. Schwenker, "Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification," *Speech Commun.*, vol. 27, no. 1, pp. 263–287, 2013.
[51] G. Degottex and N. Obin, "Phase distortion statistics as a representation of the glottal source: Application to the Classification of voice qualities," presented at the Interspeech, Singapore, Singapore, 2014.
[52] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Commun. Methods Measures*, vol. 1, pp. 77–89, 2007.
[53] N. Dehak and G. Chollet, "Support vector GMMs for speaker verification," in *Proc. Odyssey: Speaker Language Recog. Workshop*, San Juan, Puerto Rico, 2006, pp. 1–4.
[54] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1997, pp. 1895–1898.
[55] "The NIST Year 2012 Speaker Recognition Evaluation Plan," Tech. Rep., 2012. [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf
[56] D. R. Velez *et al.*, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction," *Genetic Epidemiol.*, vol. 31, no. 4, p. 306–315, 2007.
[57] H. Wanger, "Measuring performance in category judgment studies on nonverbal behavior," *J. Nonverbal Behavior*, vol. 17, no. 1, p. 3–28, 1993.
[58] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion Recognition using a Hierarchical Binary Decision Tree Approach," *Speech Commun.*, vol. 53, pp. 1162–1171, 2011.
[59] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. New York, NY, USA: Wiley, 2013.
[60] J. Hair, R. Anderson, M. Tatham, and W. Black, *Multivariate Data Analysis*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1995.
[61] N. Obin, "MeLos: Analysis and modelling of speech prosody and speaking style," Ph.D. dissertation, Inst. Res. Coordination Acoust. Music—Univ. Pierre and Marie Curie, Paris, France, 2011.

Nicolas Obin (M'13) received the M.Sc. degree in acoustics, signal processing, and computer science applied to music and the Ph.D. degree in computer sciences from the University of Paris VI, Paris, France, in 2006 and 2011, respectively. He also received the M.Sc. degree in musicology, arts, and aesthetic from the University of Paris VIII, Saint-Denis, France, in 2007. He is an Associate Professor at the Institute for Research and Coordination in Acoustics & Music (IRCAM), and the University of Pierre and Marie Curie—Sorbonne Universités. In 2006, he was a Visiting Researcher at the Center for New Music and Audio Technologies), University of California, Berkeley, CA, USA. He conducted his Ph.D. at IRCAM on the modeling of speech prosody and speaking style for text-to-speech synthesis, for which he received the award for the best French Ph.D. thesis in computational sciences from "La Fondation Des Treilles" in 2011. His primary research interests include audio signal processing and machine learning with applications to voice conversion, speech synthesis, computational linguistics/para-linguistics, and computational auditory scene analysis. His research is devoted to the elaboration of audio technologies for creation, art, and culture.

Axel Roebel (M'08) received the Diploma degree in electrical engineering from Hanover University, Hanover, Germany, and the Ph.D. degree (summa cum laude) in computer science from the Technical University of Berlin, Berlin, Germany, in 1991 and 1993, respectively. In 1994, he joined the German National Research Center for Information Technology (GMD- First) in Berlin where he continued his research on adaptive modeling of time series of nonlinear dynamical systems. In 1996, he became an Assistant Professor for digital signal processing in the Communication Science Department, Technical University of Berlin. In 2000, he was a Visiting Researcher at Center for Computer Research in Music and Acoustics, Standford University, where he worked on adaptive sinusoidal modeling. In the same year, he joined the Institute for Research and Coordination in Acoustics & Music (IRCAM) to work on sound analysis, synthesis and transformation algorithms. In summer 2006, he was Edgar-Varese Guest Professor for computer music at the Electronic Studio of the Technical University of Berlin and currently he is the head of the Sound Analysis and Synthesis Team at IRCAM. His current research interests include music and speech signal analysis and transformation.