ON THE GENERALIZATION OF SHANNON ENTROPY FOR SPEECH RECOGNITION

Nicolas Obin, Marco Liuni †

IRCAM-CNRS UMR 9912-STMS
Paris, France

ABSTRACT

This paper introduces an entropy-based spectral representation as a measure of the degree of noisiness in audio signals, complementary to the standard MFCCs for audio and speech recognition. The proposed representation is based on the Rényi entropy, which is a generalization of the Shannon entropy. In audio signal representation, Rényi entropy presents the advantage of focusing either on the harmonic content (prominent amplitude within a distribution) or on the noise content (equal distribution of amplitudes). The proposed representation outperforms all other noisiness measures - including Shannon and Wiener entropies - in a large-scale classification of vocal effort (whispered-soft/normal/loud-shouted) in the real scenario of multi-language massive role-playing video games. The improvement is around 10% in relative error reduction, and is particularly significant for the recognition of noisy speech - i.e., whispery/breathy speech. This confirms the role of noisiness for speech recognition, and will further be extended to the classification of voice quality for the design of an automatic voice casting system in video games.

Index Terms— information theory, spectral entropy, speech recognition, expressive speech, voice quality, video games

1. INTRODUCTION

This paper presents a preliminary study on the classification of expressive speech for the elaboration of a voice casting system in the context of video games, and exploits a generalization of the Shannon entropy as a measure of the degree of noisiness of a signal for automatic speech recognition (ASR). The context of video games raises innovative and challenging issues over conventional ASR systems.

Voice casting is commonly used in the audio processing of video games to transfer a video game into various languages with a limited amount of available voices for each language : for each role of the source language, one needs to determine the actor in the target language which is the more similar to the source role. In particular, the objective of a voice casting system differs qualitatively from standard speaker recognition applications : in standard speech recognition applications (speaker identification/verification), the objective is to determine the "vocal signature" of a speaker that is invariant to any potential sources of speech variability. Conversely, the objective of voice casting is to determine the "acting signature" that is invariant across speakers (including voice qualities, vocal effort, emotions/attitudes, and archetypes). In the former, the exclusive invariant is the identity of a speaker, any other sources of variability being considered as noise ; in the latter, the primary invariants are precisely these sources of variations, while the identity of a speaker is secondary, only. Additionally, video games also require processing a large range of speech variability of professional actors - including the use of extreme vocal registers, and some non-natural speech (e.g., cartoons, robots, extraterrestrials) ; large databases (from 20,000 to 40,000 audio files and around 500 roles for a single role-playing video game) ; and large differences in audio recordings duration - from a single filler (e.g., inspirations, screams, $\leq 0.5s.$) to complete utterances ($\simeq 10s.$).

In a previous study, evidence for the use of glottal source characteristics in speech recognition has been provided [1], in agreement with other recent studies [2, 3]. However, the description of glottal source characteristics requires the use of glottal source and vocal tract separation methods, which are still not robust to cover satisfactorily expressive speech and adverse recording conditions. Alternatively, the degree of noisiness may be a simple and reliable measure to capture most changes in the glottal source configuration and to cover a large range of voice qualities [4]. Additionally, the degree of noisiness of an audio signal (e.g., speech, music, sound events) may be advantageously used as a complementary representation to the standard MFCCs for audio recognition. A number of representations have been proposed to measure the degree of noisiness of a signal, based on entropy measures (SHANNON ENTROPY [5], WIENER ENTROPY [6, 7]) or on harmonic/noise decomposition [8, 1].

This paper exploits the RÉNYI ENTROPY as a generali-

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

zation of the SHANNON ENTROPY to measure the degree of noisiness of an audio signal (section 2). In particular, RÉNYI ENTROPY presents the advantage over other entropy measures of focusing either on the harmonic content (a prominent amplitude within a distribution) or on the noise content (equal distribution of amplitudes) - without requiring any harmonic/noise decomposition. A multi-resolution spectral entropy is determined through the integration of spectral entropy over logarithmically distributed frequency regions - similarly to MFCCS. This is adopted in order to represent the audio signal both in terms of locally defined energy and noisiness content over the frequency scale. The proposed representation is compared to other noisiness measures including SHANNON and WIENER entropies (section 3) within a large-scale classification of vocal effort (whispered-soft/normal/loud-shouted) in the real scenario of video games production of multi-language massive role-playing video games (sections 4 and 5).

2. SPECTRAL ENTROPY

With an appropriate normalization, the power spectrum of an audio signal can be interpreted as a probability density. According to this interpretation, some techniques belonging to the domains of probability and information theory can be applied to sound representation and recognition : in particular, the concept of entropy can be extended to provide a concentration measure of a time-frequency density - which can be interpreted as a measure of the degree of voicing (alternatively, noisiness) of an audio signal. The representation adopted in this study (see [9] for the original formulation) is based on the use of Rényi entropies, as a generalization of the Shannon entropy [10]. In this section, the Rényi entropy is introduced with regard to information theory, and some relevant properties for the representation of audio signals are presented. In particular, the Rényi entropy presents the advantage over the Shannon entropy of focusing on the noise or the harmonic content.

2.1. Rényi entropy

Definition 2.1 *Given a finite discrete probability density* $P = (P_1, ..., P_N)$ *and a real number* $\alpha \geq 0,\ \alpha \neq 0$*, the* Rényi entropy *of* P *is defined as follows,*

$$\mathrm{H}_\alpha[P] = \frac{1}{1-\alpha} \log_2 \sum_{n=1}^{N} P_n^\alpha \ , \qquad (1)$$

where P *is in square brackets to indicate that discrete densities are considered.*

Among the general properties of Rényi entropies (see [11], [12] and [13]), the fundamental ones are here shortly summarized.

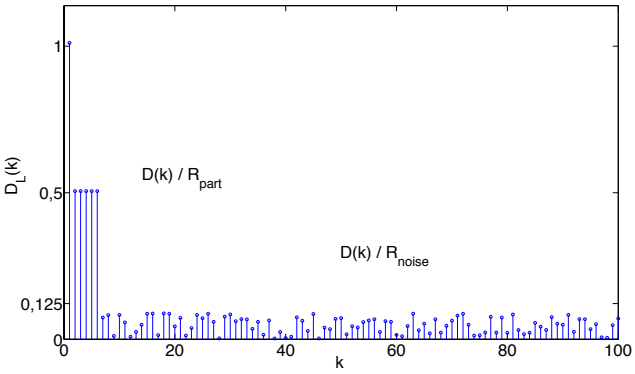

Fig. 1. An example of D_L vector, see equation (3) : here, $N_{part} = 5$, $R_{part} = 2$ and $L = 1/4$ so that $R_{noise} = 8$.

1) for every finite discrete probability density P, the entropy $\mathrm{H}_\alpha[P]$ tends to the Shannon entropy of P as the order α tends to one.

2) $\mathrm{H}_\alpha[P]$ is a non increasing function of α, so :

$$\alpha_1 < \alpha_2 \Rightarrow \mathrm{H}_{\alpha_1}[P] \geq \mathrm{H}_{\alpha_2}[P] \ . \qquad (2)$$

In case of finite discrete densities, the case $\alpha = 0$ can also be considered, which simply gives the logarithm of the number of elements in P ; as a consequence $\mathrm{H}_0[P] \geq \mathrm{H}_\alpha[P]$ for every admissible order α.

3) for every order α, the Rényi entropy H_α is maximum when P is uniformly distributed, while it is minimum and equal to zero when P has a single non-zero value.

The main advantage of the Rényi entropies is the dependence on the order α, which provides a different concept of concentration for each value of α : in particular, these measures differ by the reshaping applied to the spectrum coefficients before summing, by means of raising them to the α power. This concept is qualitatively detailed in the following section.

2.2. Some Properties on the Representation of Noise/Harmonic Content

The α parameter in equation (1) introduces a biasing on the spectral coefficients, which gives them a different relevance in the entropy evaluation of the representation ; this means that different values of α determine different concepts of noisiness. Basically, small α values tend to emphasize the noise content of signal, while large α values tend to emphasize the harmonic content of a signal.

Consider a vector D of length $N = 100$ generating numbers between 0 and 1 with a normal random distribution, and two integers $1 \leq N_{part}$, $1 \leq R_{part}$. Define D_L as follows :

$$D_L[n] = \begin{cases} 1 & \text{if } n = 1 \\ \frac{D[n]}{R_{part}} & \text{if } 1 < n \leq N_{part} \\ \frac{D[n]}{R_{noise}} & \text{if } n > N_{part} \ . \end{cases} \quad (3)$$

where $R_{noise} = \frac{R_{part}}{L}$, $L \in [\frac{1}{16}, 1]$, then we normalize to obtain a unitary sum. These vectors are a simplified model of an amplitude spectrum (figure 1) whose coefficients correspond to one main peak, N_{part} partials with amplitude reduced by R_{part}, and some noise, whose amplitude varies proportionally to the L parameter, from a negligible level to the same one of the partials. Such a class of vectors represents more general situations, as the entropy measure is permutation-invariant, so that the order of the coefficients in a vector does not modify its entropy value.

Applying Rényi entropy measures with α varying between 0 and 3, we obtain figure 2, which shows the impact of the noise level L on the evaluations with different values of α.

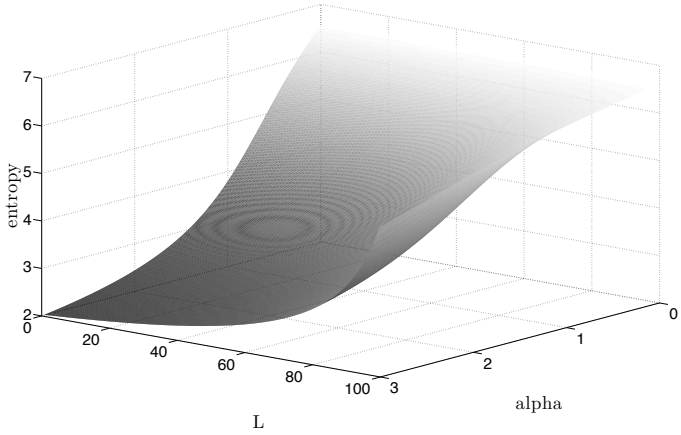

Fig. 2. Rényi entropy evaluations of the D_L vectors with varying α, $N_{part} = 5$ and $R_{part} = 2$; the entropy values rise differently as L increases, depending on α : this shows that the impact of the noise level on the entropy evaluation depends on the entropy order.

For $\alpha = 0$, $\mathrm{H}_0[D_L]$ is the logarithm of the number of non-zero coefficients and it is therefore constant. The increment of L corresponds to a strengthening of the noise coefficients, causing the rise of the entropy values for any α. The key point is the observation of how they rise, depending on the α value : the convexity of the surface in figure 2 increases as α becomes larger, and it describes the impact of the noise level on the evaluation ; the stronger convexity when α is around 3 denotes a low sensitivity of the measure to the noise coefficients, as their level needs to be high to determine a significant entropy variation. On the other hand, when α tends to 0 the entropy growth is almost linear in L, showing the significant impact of noise on the evaluation, as well as a finer response to the variation of the partials amplitude. As a consequence, the tuning of the α parameter has to be performed according to the desired trade-off between the sensitivity of the measure to the main peaks and the weak signal components to be observed, as noise in our case.

2.3. Multi-Resolution Spectral Entropy

A multi-resolution spectral entropy is adopted in order to provide a representation in which the harmonic/noise content of the signal is determined over locally defined frequency bands - similarly to MFCCS. For each frequency band, the RÉNYI ENTROPY is measured as :

$$\mathrm{H}_\alpha^{(i)} = \frac{1}{1-\alpha} \log_2 \sum_{n=1}^{N^{(i)}} \left(\frac{|A(n)|^2}{\sum_{n=1}^{N^{(i)}} |A(n)|^2} \right)^\alpha \quad (4)$$

where $A(n)$ is the amplitude of the n-th frequency bin in the considered frequency band, and the denominator on the right side accounts for the normalization of the power spectrum in the considered frequency band into a probability density function.

In this study, the RÉNYI ENTROPY has been computed over 25 Mel-frequency bands. A study-case of multi-resolution Réyni entropy for whispered and shouted speech is provided in figure 3.

3. OTHER MEASURES

3.1. MFCC

13 Mel-frequency cepstral coefficients (MFCC) are extracted after the non-linear compression of amplitude spectrum into 25 Mel-frequency bands. Short-term features were extracted with a 25-ms. hanning 5-ms. shifting window.

3.2. Spectral Flatness Measure

The spectral flatness measure (SFM) has been introduced as a measure of the noisiness contained in a signal [6, 7]. The spectral flatness is defined as the ratio of the geometric mean to the arithmetic mean of the power spectrum :

$$\mathrm{SFM}^{(i)} = \frac{\prod_{n=1}^{N^{(i)}} |A(n)^2|^{1/N^{(i)}}}{\frac{1}{N^{(i)}} \sum_{n=1}^{N^{(i)}} |A(n)|^2} \quad (5)$$

where $A(n)$ is the amplitude of the n-th frequency bin in the considered frequency band.

Hence, SFM tends to zero when the spectral distribution of the frequency band exhibits a salient peak, and is equal to one when the spectral distribution of the frequency band is flat.

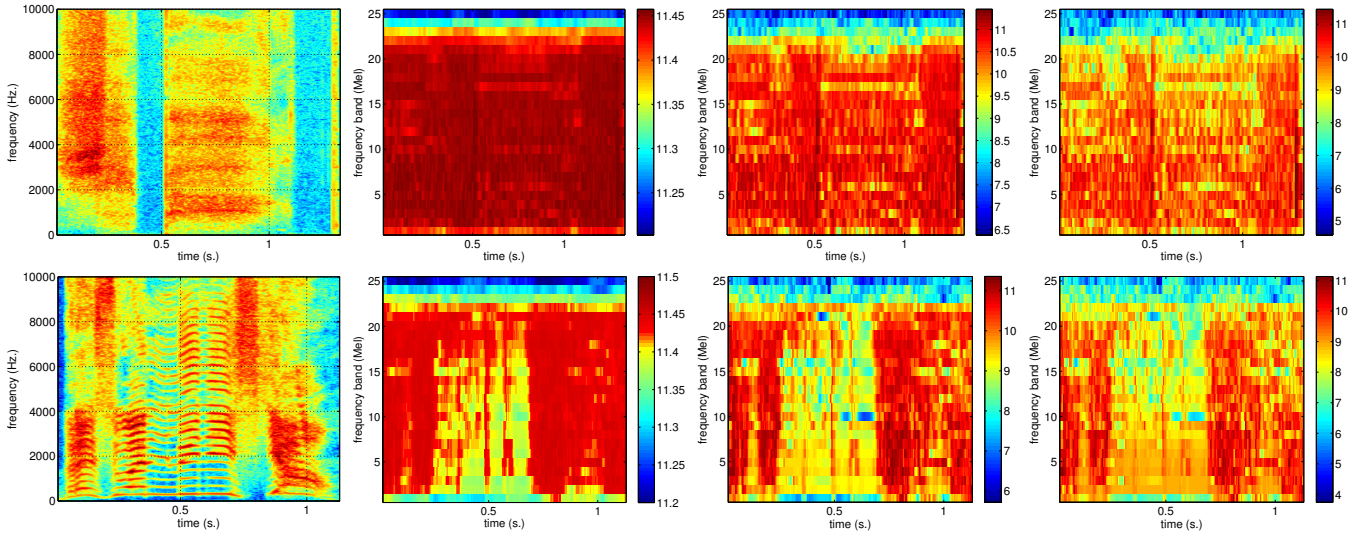

Fig. 3. Study cases of Rényi entropy measures for whispered (top) and shouted (bottom) speech. From left to right : spectrogram, $\alpha = 0.001$, $\alpha = 1$ (Shannon entropy), and $\alpha = 3$.

The spectral flatness is usually referred to and can also be interpreted as a Wiener entropy [14]. The current implementation is based on the MPEG-7 standard [15], in which the spectral flatness is computed in the following 4 frequency regions : (250-500), (500-1000), (1000-2000), (2000-4000) Hz [1].

3.3. Voiced/UnVoiced Measure

The soft voiced/unvoiced measure (VUV) [8, 1] is a measure of the degree of voicing (alternatively, noisiness) contained in a signal. For each frequency band, the VUV is measured as :

$$\mathrm{VUV}^{(i)} = \frac{\sum_{k=1}^{K^{(i)}} |A_H(k)|^2}{\sum_{n=1}^{N^{(i)}} |A(n)|^2} \qquad (6)$$

where i denotes the i-th Mel frequency band, $A_H(k)$ the amplitude of the k-th harmonic, K the number of harmonics, and $A(n)$ the amplitude of the n-th frequency bin in the i-th frequency band. Here, the harmonic/noise decomposition is obtained using the sinusoidal peak classification method presented in [16].

Hence, VUV is equal to zero when no harmonic content is present in the frequency band, and to one when only harmonic content is present in the frequency band. In this study, the VUV has been computed over 25 Mel-frequency bands.

4. SPEECH DATABASE

Speech recordings of a video game consist of the interpretation of script lines by professional actors ; they are directed by an artistic director, whose role is to control the expressive content of speech, depending on the place of the script within the overall scenario.Script lines may vary from a single sigh to a complete sentence. In role-playing games, the recording covers ten thousands of speech files that are split into hundreds of roles. The video game used for the study includes around 20,000 French (MASS EFFECT) and 40,000 German (DRAGON AGE) speech recordings, and around 500 roles (from a single to hundreds of recordings) for each video game. The duration of speech recordings varies from 0.1 seconds to 20 seconds with a mean duration of 2.5 seconds. Recordings were made in mono 48 kHz/16-24 bits uncompressed format.

Speech recordings are produced in a studio by professional actors with a varying distance and orientation to the microphone so as to compensate for the variations in acoustics due to changes in intended vocal effort (close while whispering, distant while shouting). Additionally, a sound engineer ensures that the speech level is constant through speech recordings so as to provide a homogeneous speech level through the video game. Hence, only information provided by changes in the source excitation or the vocal tract resonances can be used for the identification of changes in vocal effort. In particular, the degree and the extent of noisiness in speech may efficiently reflect significant changes in configuration and in intensity of the glottal source.

The identification of significant changes in vocal effort is critical in the production of video games for the application of specific settings that simulate the perception of proximity to the speaker by the player. For this purpose, the sound engineer is usually in charge for the manual classification of expressive speech recordings into three classes, which cover whis-

1. Contrary to the other measures, which are all computed over 25 Mel frequency bands.

MASS EFFECT FRENCH	WHISPERED	NORMAL	SHOUTED	TOTAL
MFCC	69.4	82.5	91.0	81.1
MFCC + SFM	73.1	84.4	92.6	83.3
MFCC + VUV	75.0	84.3	91.2	83.5
ENTROPY MFCC + α				
$\alpha = 0.001$	75.0	84.0	91.3	83.4
$\boldsymbol{\alpha = 0.01}$	**75.2**	**84.2**	**92.0**	**83.8**
$\alpha = 0.1$	75.0	84.3	91.3	83.5
$\alpha = 0.2$	74.6	84.0	91.6	83.4
$\alpha = 0.4$	74.5	84.0	91.5	83.4
$\alpha = 0.6$	74.2	84.0	91.3	83.2
$\alpha = 0.8$	74.2	84.0	91.2	83.0
$\alpha = 1$ (SHANNON)	74.2	83.9	91.0	83.0
$\alpha = 2$	75.0	84.3	91.3	83.5
$\alpha = 3$	74.5	84.0	91.5	83.3
$\alpha = 4$	74.6	83.8	91.5	83.2
$\alpha = 5$	74.1	83.8	91.2	83.0

DRAGON AGE GERMAN	WHISPERED	NORMAL	SHOUTED	TOTAL
MFCC	73.1	77.2	87.1	79.0
MFCC + SFM	73.4	77.9	87.5	79.6
MFCC + VUV	74.3	78.7	87.2	80.1
ENTROPY MFCC + α				
$\alpha = 0.001$	74.6	78.0	87.2	79.9
$\boldsymbol{\alpha = 0.01}$	**75.1**	**78.5**	**87.5**	**80.4**
$\alpha = 0.1$	74.5	78.0	87.5	80.1
$\alpha = 0.2$	74.5	78.2	87.3	80.0
$\alpha = 0.4$	74.2	78.0	87.3	79.8
$\alpha = 0.6$	74.0	77.9	87.0	79.6
$\alpha = 0.8$	74.0	77.7	87.0	79.6
$\alpha = 1$ (SHANNON)	73.8	77.5	87.1	79.5
$\alpha = 2$	73.6	77.8	87.4	79.6
$\alpha = 3$	72.8	77.0	87.4	79.1
$\alpha = 4$	72.5	77.1	86.7	78.8
$\alpha = 5$	72.4	77.0	86.5	78.5

Table 1. F-measure obtained for the conventional MFCC, MFCC + noisiness measures : (VUV), and MFCC + ENTROPY measures (SFM, SHANNON, RÉNYI) for the MASS EFFECT (FR) and DRAGON AGE (GR) role-playing video games.

pered/soft, normal, and loud/shouted speech. In the present study, whispered/soft speech covers sighs, true whisper, stage whisper, stressed whisper (tense whisper typically produced in a life-survival situation in which the conversation intimacy is absolutely required), soft speech, and any situation of intimacy in the speech communication. Loud/shouted speech includes orders, public announcements, exclamations, interjections, stressed-speech, and screams.

5. EVALUATION

The relevance of the noisiness measures for the classification of vocal effort into whispered/soft, normal, and loud/shouted speech has been conducted within a 5-fold cross-validation. Additional constraints on the design of the cross-validation have been adopted : well-balanced distribution of the vocal effort classes within the train and test sets ; no role overlapping across train and test sets, in order to prevent the system to be biased by speaker identification.

The classification system is based on the GMM-UBM/SVM system [17], which is a standard for speech recognition [18, 19]. A UNIVERSAL BACKGROUND MODEL (GMM-UBM) is used to model the acoustic variability in the speech database with a Gaussian Mixture Model (GMM), represented by a sequence of short-term acoustic feature vector (here, MFCCs and noisiness measure). Then, each utterance is represented as a supervector by MAP adaptation of the GMM-UBM mean vectors [18]. Then, supervectors are used to determine the parameters of a SUPPORT VECTOR MACHINE classifier (SVM) which maximize the margin of a high-dimensional separation hyperplane for the classification [19].

During the training, each feature set is considered as a separate stream for the determination of the GMM-UBM and the SVM parameters. 64 GMMS with diagonal covariance matrices have been used to determine the parameters of the GMM-UBM, and a SVM with a GMM-supervector radial basis function (RBF) kernel has been determined with various values of the radial bandwidth (from 0.1 to 5). During the classification, the decision is made by fusioning the affinity obtained for each stream using average decision fusion. The performance of the system has been measured with the F-measure metric. Finally, the performance of the system corresponds to the optimal performance obtained for each feature set.

6. DISCUSSION

Table 1 summarizes the performance of the system for the noisiness measures considered, and in particular for a large range of α values in the RÉNYI ENTROPY. The performance obtained leads to the following observations :

First, the performance tendencies are strictly similar regardless to the language of the speech database, which strongly indicates that the acoustic configurations involved in a change in vocal effort may be universal - i.e., does not depend on a specific language.

Second, the use of a noisiness representation complementary to the standard MFCCS improves the performance regardless to the noisiness measure, which confirms the relevance of noisiness measures for speech recognition. More interestingly, a comparison of the existing noisiness measures suggests the following relevance rank : 1) VUV, 2) SFM, and 3) SHANNON ENTROPY, without a significant difference for the 2 last measures.

Finally, the RÉNYI ENTROPY outperforms all of the other noisiness measures. The increase in performance is particularly significant for a small α (focus on the noise content) and for noisy speech (whispered), while a large α (focus on the harmonic content) does not provide any improvement - and may even degrade the performance of the system compared to the SHANNON ENTROPY. In all cases, the optimal performance is obtained with a α close to 0 ($\alpha = 0.01$) and represents around a 10% relative error reduction over conventional MFCCS.

7. CONCLUSION

In this paper, the Rényi entropy was introduced as a generalization of the Shannon entropy to measure the degree of noisiness in audio signals, complementary to the standard MFCCs for audio and speech recognition. In audio signal representation, Rényi entropy presents the advantage of focusing either on the harmonic content (prominent amplitude within a distribution) or on the noise content (equal distribution of amplitudes). The proposed representation outperforms all other noisiness measures - including Shannon and Wiener entropies - in the large-scale classification of vocal effort in the real scenario of video games production of multi-language massive role-playing video games. This confirms the role of noisiness for speech recognition, and will further be extended to the classification of voice quality for the design of a automatic voice casting system in video games.

8. REFERENCES

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

[2] Tomi Kinnunen and Paavo Alku, "On Separating Glottal Source and Vocal Tract Information in Telephony Speaker Verification," Taipei, Taiwan, 2009, pp. 4545–4548.

[3] Thomas Drugman, Thomas Dubuisson, and Thierry Dutoit, "On the Use of the Glottal Source for Expressive Speech Analysis," in *Pan European Voice Conference*, Marseille, France, 2011.

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

[5] Hemant Misra, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky, "Spectral Entropy based Feature for Robust ASR," 2004, vol. 1, pp. 193–196.

[6] James D. Johnston, "Transform Coding of Audio Signals using Perceptual Noise Criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–332, 1988.

[7] Shlomo Dubnov, "Generalization of Spectral Flatness Measure for Non-Gaussian Linear Processes," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 698–701, 2004.

[8] Daniel W. Griffin and Jae S. Lim, "A New Model-Based Analysis/Synthesis System," in *International Conference on Acoustics, Speech, and Signal Processing*, Tampa, Florida, 1985, pp. 513–516.

[9] Richard G. Baraniuk, Patrick Flandrin, Augustus J.E.M. Janssen, and Olivier Michel, "Measuring Time-Frequency Information Content Using the Rényi Entropies," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1391–1409, 2001.

[10] Claude E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[11] Alfréd Rényi, "On Measures of Entropy and Information," in *Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, California, 1961, pp. 547–561.

[12] Christian Beck and Friedrich Schögl, *Thermodynamics of Chaotic Systems*, Cambridge University Press, Cambridge, Massachusetts, USA, 1993.

[13] Karol Zyczkowski, "Rényi Extrapolation of Shannon Entropy," *Open Systems & Information Dynamics*, vol. 10, no. 3, pp. 297–310, Sept. 2003.

[14] Norbert Wiener, *In The human use of human beings : Cybernetics and society*, chapter Cybernetics in History, pp. 15–27, Houghton Mifflin, Boston, 1954.

[15] "Multimedia content description interface - Part 4 : Audio," *ISO/IEC FDIS 15938-4*, 2002.

[16] Miroslav Zivanovic, Axel Röbel, and Xavier Rodet, "Adaptive Threshold Determination for Spectral Peak Classification," *Computer Music Journal*, vol. 32, no. 2, pp. 57–67, 2008.

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

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

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