Using pitch features for the characterization of intermediate vocal productions

Lionel Feugère, Boris Doval, Marie-France Mifune

USING PITCH FEATURES FOR THE CHARACTERIZATION OF INTERMEDIATE VOCAL PRODUCTIONS

Lionel Feugère, Boris Doval
LAM-Institut Jean Le Rond d'Alembert
CNRS UMR 7190, Sorbonne Universités
UPMC Univ Paris 06, F-75005, Paris, France
lionel.feugere,boris.doval@upmc.fr

Marie-France Mifune
UMR 7206 Eco-anthropologie et Ethnobiologie
CNRS-MNHN-Université Paris Diderot-
Sorbonne Universités, Paris, France
mifune@mnhn.fr

ABSTRACT

This paper presents some pitch features for the characterization of intermediate vocal productions from the CNRS - Musée de l'Homme sound archives, in the context of the DIADEMS interdisciplinary project gathering researchers from ethnomusicology and speech signal processing. Different categories – chanting, singing, recitation, storytelling, talking, lament – have been identified and characterized by ethnomusicologists and are confronted by acoustic analysis. A database totalizing 79 utterances from 25 countries spread around the world is used. Among the tested features, the note duration distribution has proved to be a relevant measure. Categories are mostly characterized by the proportion of 100-ms notes and the duration of the longest note. An evaluation of these features has been realized through a supervised classification using the different vocal categories. Classification results show that these two features allow a good discrimination between "speech", "chanting" and "singing", but are not suited for discriminating between the "speech" subcategories "recitation", "storytelling" and "talking".

1. INTRODUCTION

The context of this study [1] is the DIADEMS project [2], which gather together ethnomusicologists, linguists, acousticians, archivists and specialists in speech processing and music information retrieval, around a sound archive web platform, Telemeta (Fillon et al., 2014). This software platform allows the user to browse, listen, watch and annotate multimedia files.

The aim of the DIADEMS project is to develop computational tools to automatically index the audio content of the CNRS - Musée de l'Homme sound archives (49,300 audio items from 5,800 collections including 28,000 items already uploaded). This ethnomusicological archive includes published and unpublished recordings of music and oral traditions from around the world, spanning a wide variety of cultural contexts worldwide starting in the 1900s until today as well as a wide variety of contents (musical practice, speech, dance, ritual, interview and so on) and various settings (inside, outside, rarely in studio settings). Many sound archives in the collection include very little contextual information. The use of automatic indexation tools will help archivists to index such sound items and add new content information. It will also facilitate searches, analyses and comparison of large corpus by ethnomusicologists.

The core of the project deals with automatic indexing of a large variety of sound productions directly from audio signals including musical instruments, environmental sounds and vocal productions. In the context of the DIADEMS project, this study aims at helping ethnomusicologist to characterize speech, song and intermediate vocal productions in terms of acoustical parameters.

2. ETHNOMUSICOLOGICAL APPROACH

Several studies in ethnomusicology have characterized speech, song and intermediate forms through musicological and acoustic parameters in relation to the social and cultural context (Seeger, 1986; Amy de la Bretèque, 2010; Rappoport, 2005).

While the classic ethnomusicological approach considers that it is not possible to define vocal categories independently of the cultural context (Picard, 2008), some ethnomusicologists proposed attempts of classification of vocal productions among the different cultures.

The ethnomusicologist List (1963) proposed a method of classification by which distinctions and relations can be made between speech, song and intermediate forms. This classification system is based on two divergent modifications of speech intonation: 1) the negation or the leveling out of speech intonation into monotone, 2) the amplification or exaggeration of speech intonation (such as Sprechstimme). The next step to song is either 1) the stability of pitch or 2) the expansion of scalar structure. In his classification model, List didn't directly take into account the pitch duration. According to him, the comparative length or shortness of sustained pitch would be a useful criterion to incorporate in this system of classification but the assumption that song exhibits pitches of greater duration than speech would not be valid in all culture.

For particular productions like lament, Urban (1988) have found that some common features could be brought out among different cultures.

It is extremely challenging for ethnomusicologists to define an efficient categorization of vocal productions based only on acoustic criteria and equally efficient in all cultural practices worldwide for two main reasons: 1) procedures and techniques of vocal productions worldwide are insuf-

[1] This work is partly supported by a grant from the French National Research Agency (ANR) with reference ANR-12-CORD-0022.

[2] http://www.irit.fr/recherches/SAMOVA/DIADEMS/fr/welcome/

ficiently described and the inventory is incomplete; 2) ethnomusicological method and acoustical terminology often lack consensus and are somewhat approximate.

The first attempt of classification of vocal productions worldwide was a typology with audio examples based on several techniques (Zemp et al., 1996): calls, cries, clamours, voice and breath, spoken, declaimed, sung, compass and register, colours and timbres, disguised voices, ornamentation, voices and musical instruments, employ of harmonics. The main issue is that these several categories are not systematically based on evident and explicit acoustic criteria. Then, based on this classification, Léothaud (2007) put forward a universal typology of vocal techniques based on acoustic criteria such as timbre, register, tessitura and intensity.

As Giannattasio (2007) suggested, realizing a typology implies an analysis based on common acoustical parameters (tempo, intonation, timbre, etc.). By exploring the continuum between speech and song, we must define the different modalities of expression based on the several parameters without defining a predetermined order among them.

This study intends both to describe acoustical features that characterize the vocal categories and that would apply to different voice production excerpts from all over the world, and to draw up definitions of each vocal category from an ethnomusicological point of view.

We chose to classify vocal productions in two general categories: speech and song for two main reasons: 1) every culture distinguishes between talking and singing among all vocal productions; 2) we consider multiple vocal categories with acoustic characteristics ranging from speech to song.

Then, according to the database, we subsequently identified and characterized subcategories such as talking, storytelling, recitation, chanting, singing and lament, not based on style or genre, but on acoustic features only. We defined these subcategories without establishing a predetermined order among them. We define *speech* as a vocal production with a significant proportion of unvoiced sounds. Alternatively, *song* is defined as a vocal production with a significant proportion of lengthened syllables and voiced sounds. We consider *talking*, *recitation* and *storytelling* close to speech, since they are also characterized with a significant proportion of unvoiced sounds. We distinguish *talking* from *storytelling* based only on the mode of realization: *talking* is characterized by dialogue and *storytelling* by monologue with or without back-channel signal (i.e., an expression or word used by a listener to indicate that he or she is paying attention to the speaker). *Recitation* is characterized by more regular breath rate and rhythmic flow than *talking*, and a monotonous statement with low frequency range variations. We consider *singing* and *chanting* close to song, since they are characterized with a significant proportion of lengthened syllables and voiced sounds. *Singing* is characterized by ordered pitches and relative stability of fundamental frequencies while *chanting* is characterized by a very limited vocal range, close to recto-tono. We define *lament* by the presence of several of the four common icons of crying (the cry break, the voice inhalation, the creaky voice and the falsetto vowel) proposed by Urban (1988).

These definitions are a first attempt based on the ethnomusicological archives with which researchers in the DIADEMS project are most familiarized. Ultimately, the goal is to refine these categorical definitions while expanding the corpus of recordings considered for the automatic indexation. In particular, we are aware that some of the terminology used here can be inappropriate for some specific practices or can be confusing for some ethnomusicologists in the community. The aim of the definition and characterization of the ethnomusicological categories is not to replace the endogenous categories but to provide scientific tools to better analyze vocal productions. This work focuses more on the characterizations of descriptors and acoustic parameters rather than in the definition of the category themselves.

In the information retrieval community, few have addressed the issue of intermediate vocal categories, focusing more on cultural style (Liu et al., 2009), singing voice timbre (Fujihara & Goto, 2007), speech style (Goldman et al., 2009), or singing versus speech classification on homogeneous and/or good quality recordings (Gärtner, 2010).

Section 3 presents the corpus and the features that have been studied, section 4 presents the results in term of characterization and classification rates, and section 5 presents a discussion and some perspectives.

3. METHODS

3.1 Corpus

Ethnomusicologists from the Diadems project manually annotated the audio contents of a representative sample of sound items from the CNRS - Musée de l'Homme sound archives, in order to give to acousticians a data set of each subcategory mentioned above.

This resulted in 79 items from different contexts (rituals, enquiries, tales, etc.) and various cultures as shown in table 1. They were selected for their non-ambiguous category during 10 sec minimum. In each category, most of the utterances are from a different field. Except if indicated, the utterances from a same country and a same category are from a different speaker (but also often from a different context).

3.2 Previous study on intermediate vocal categories from the Telemeta corpus

In the context of the DIADEMS project, Sotiropoulos (2014) proposed a decision tree to classify utterances to four categories (chanting, singing, recitation and speaking/storytelling) from a small subset of the Telemeta database, composed of 6 utterances by category (including sometimes the same speaker). Mean voicing duration discriminates song and speech categories, while in the secondary nodes of the decision tree, mean duration of non-voice units divides talking+storytelling and recitation, and pitch range allows to

Class	Number	Origin (number)
Chanting	22	Cambodia (2), South India, Indonesia (7), Iran, Ladakh, Mexico (4), Nepal (3 including 2 same speakers), Tibet, Vietnam (2)
Singing	19	Albania, Armenia (2), Azerbadjan, Egypt, Gabon (2), Indonesia (3), Macedonia, Madagascar, Morocco, Nepal, the Philippines, Yemen (3), Turkey
Storytelling	8	Central African Republic, Gabon (5 including 2 same speakers), Mali, Paraguay
Recitation	8	Madagascar, Mali (3 including 2 same speakers), Mexico, Paraguay, Tibet, Yemen
Talking	12	France (2), Gabon (8), Mexico, Madagascar
Lament	10	Albania, Armenia, Azerbaijan, Ethiopia, Gabon, Paraguay (2 same speak.), Turkey(2), Vietnam

Table 1: Number and origin of the sound utterances used in this work.

separate singing and chanting. These features will be gathered with the present article ones in the future final system.

3.3 Features

3.3.1 Chroma spectrum

The audio recordings are from all over the world and many cultures, so no pitch reference is there and we are not necessarily in an equal tempered 12-semitones scale. Then instead of using chroma vector (i.e. 12 discrete notes) as it is mostly done (Harte & Sandler, 2005; Lartillot et al., 2007), we use chroma spectrum (Dannenberg & Goto, 2008).

The frequency axis of a spectrum is interpolated to get semitones units and the octaves are summed in order to have all the frequency axis information inside one octave. In the resulting chroma spectrum, the number of peaks (with an amplitude above a given threshold) and their width are computed.

3.3.2 Note distribution

Besides the audio recordings are not studio recorded, they are very often noisy, polyphonic and accompanied with other instruments. So we decided not to use standard F0 detection, but rather to design a quite simple but robust algorithm adapted for ethnomusicological voice recordings. We chose to detect all the partials (named note) with a sufficient energy in the spectrogram. For each note, its duration is determined and the resulting note duration distribution is computed for each 10-sec audio utterance.

First a modified spectrogram is computed as follows:

- If the audio is stereo, a mono signal is computed
- The spectrogram is built by computing the magnitude spectra every 50ms on 100 ms windows, over 10 sec of the audio file, and on a 5 octaves scale from 110 Hz to 3520 Hz.
- An interpolation is done in order to transform the frequency axis in log2 scale with 110 Hz as reference.
- Two thresholds are applied: a dynamic energy threshold equal to 1.5 times the mean energy computed on a local 500-ms window, and an absolute energy threshold computed from the noise level.

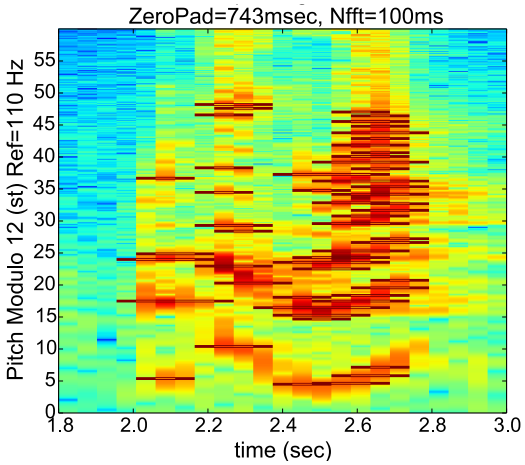

Figure 1: Spectrogram (frequency axis in semitone) annotated with the detected notes (in brown).

Then notes are detected, defined as energy on a constant bin frequency ± 0.175 semitones. The interval of ± 0.175 semitones is related to the tremor frequency range (Stables et al., 2012). For each frequency bin ± 0.175 semitones, energy greater than 0 is searched for from the initial bin window. A note is considered as finished if no energy is found for one time sample (50 ms) inside the frequency interval around the bin. Then, the note duration is computed. As an analysis step of 50 ms is used, the note duration can take values from 50 ms to 5 sec by steps of 50 ms. Finally, adjacent notes along the frequency axis are grouped together in order to avoid multiple note detection for a single energy peak. Figure 1 gives the detection result on an audio signal of 1.2 sec. The notes are surrounded by brown thin rectangles.

A distribution of the note durations is computed, normalized by the total duration of the detected notes. This normalization allows the distribution to be independent from similar repetitions of any vocal production.

Then each note duration proportion is multiplied by its note length, so that the resulted distribution can be thought of as a proportion in duration rather than a proportion in number of notes.

Figure 2 gives an example of two distributions: an audio file labelled as talking and another one labelled as chanting.

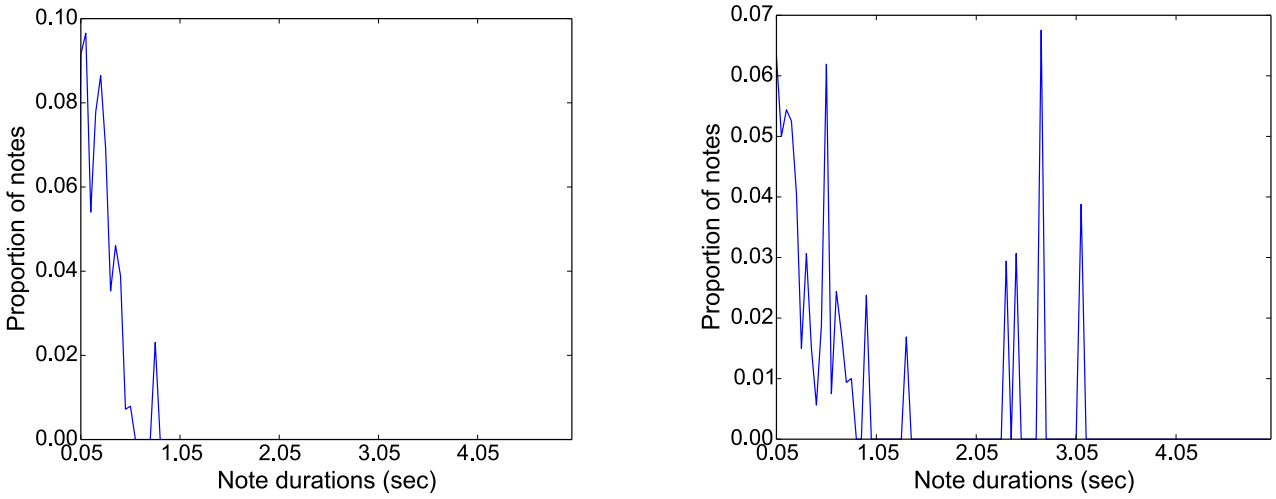

Figure 2: Note duration distributions for two excerpts. Left: talking. Right: Chanting.

Notice that no other vocal feature is taken into account to detect notes, like vowel or consonant articulation. As our note detection system focuses on pitch and energy only, the consecutive syllables on a same pitch are grouped into one single note, which occurs quite often in chanting (and spread in several utterances along the frequency axis). In this case, the number of long notes, as we defined it, is increased.

From this distribution, we chose to compute the note duration range (NDR), which is the longest note duration. On figure 2, the left distribution has a NDR of 0.9 sec while the right distribution has a NDR of 3.05 sec.

Other parameters are computed from the note detection. The normalized total duration of detected notes (named TotDurNote) is the total duration of detected note divided by the audio file duration. The mean instantaneous note number (named InstNoteNum) is computed by dividing the total duration of the detected notes by the segment duration where notes have been detected. The voicing proportion (named VoicProp) is the segment duration where notes have been detected divided by the audio file duration.

4. RESULTS

4.1 Characterization from the note duration distribution

Statistics on note duration distributions are displayed figure 3. Each figure is related to one category and displays the note duration distribution for all sound utterances of this category. For each note duration proportion, statistics over the utterances from a same category are represented by the help of the following tools: a vertical box for which borders correspond to the first and third quartile (i.e. quarter of the utterances are above the box while quarter of the utterances are below the box, so half of the data is included in this box); an horizontal red line showing the median (i.e. half of the data are spread above the line, the other half below the line); vertical whiskers representing 50% of the data plus 3 times the interquartile range (third quartile minus first quartile); fliers representing data that extend beyond the whiskers (outliers).

The main difference in the note duration distribution between categories is the proportion of very short and long notes.

For each category, table 2 gives the note duration range NDR (calculated either from the median or from the outliers) and the coordinate of the median maximum of the note duration distribution. Notice that the different dispersions between the categories are related to their number of utterances, which are quite different in our database.

4.2 Characterization from 100-ms note proportion and note duration range

Several features computed from the note detection are discriminant but the result is different from one category to another and it depends on the number of considered categories. One of the most discriminant couple of features for the whole categories are the 100-ms note proportion (second value of the note duration distribution) and the note duration range (NDR), using cross validation protocol and k-Nearest Neighbour classifier, so they were chosen to display the utterances.

Displayed in the 2D space formed by the values of these two parameters, the utterances are expected to group into characteristic areas related to their category.

As shown on figure 4:

- singing category is characterized by a small value of note duration range and proportion of 100-ms notes;
- chanting category is rather characterized by a small value of proportion of 100-ms notes and a large value of note duration range;
- speech categories (talking, recitation and storytelling) are characterized by a small value of note duration range and a large proportion of 100-ms notes;
- lament category is characterized by a small proportion of 100-ms notes.

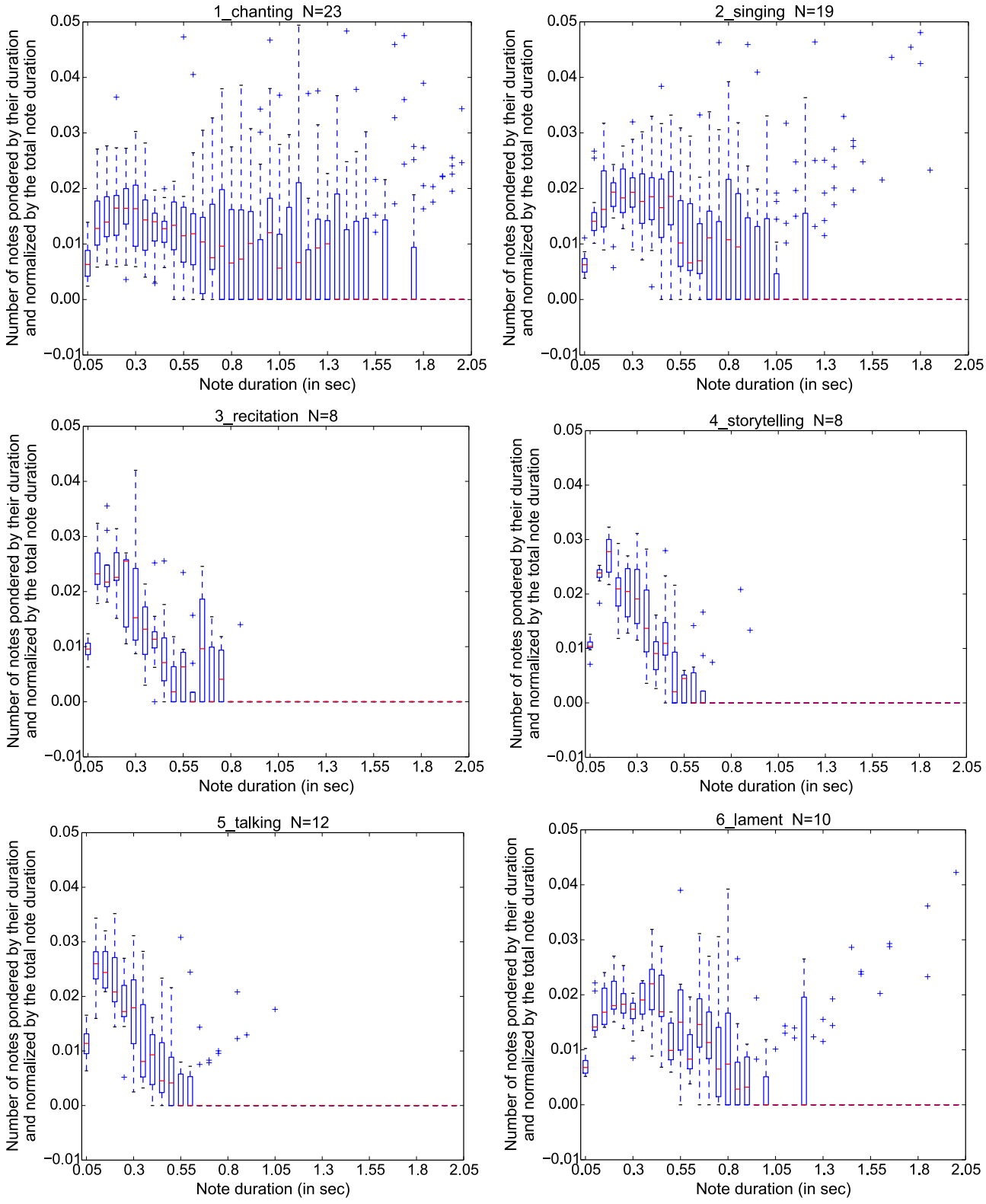

Figure 3: Median values and dispersion of the note duration distribution for the 6 categories chanting, singing, recitation, storytelling, talking and lament (see the text for more information).

Parameter	Chanting	Singing	Recitation	Storytelling	Talking	Lament
NDR (median)	1.30 sec	0.85 sec	0.75 sec	0.55 sec	0.50 sec	0.90 sec
NDR (all points except outliers)	1.75 sec	1.20 sec	0.75 sec	0.65 sec	0.60 sec	1.20 sec
Distrib. maximum (median)	1.64%	1.93%	2.55%	2.78%	2.60%	2.20%
and duration	0.20 sec	0.20 sec	0.25 sec	0.15 sec	0.10 sec	0.40 sec

Table 2: First line: note duration beyond which the median is 0. Second line: note duration beyond which non-0-values come from outliers only. Third line: coordinate of the median maximum of the note duration distribution.

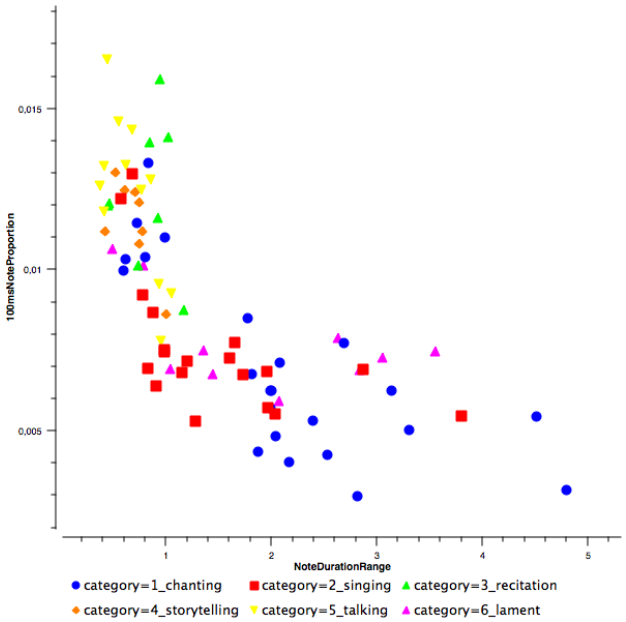

Figure 4: Utterance position in a 2D space according to their values of NDR and to 100-ms note proportion.

As can be seen on figure 4, the items from talking, storytelling and recitation are mostly overlapped, the items from chanting and singing are partially overlapped and lament items are mostly overlapped with singing or chanting ones. Then it seemed interesting to study the results when grouping speech categories from one side (talking, storytelling and recitation), and song categories from the other side (singing and chanting), as displayed in figure 5. Notice that lament was not considered as a distinct category by ethnomusicologists.

These overlappings are quite easily explained. First, on the 3-classes figure, the two singing utterances [3] which have a large value of NDR have long notes like in chanting, while the three singing utterances [4] with high value of proportion of 100-ms notes are very articulated like in speech. Second, on the same figure, the 6 chanting utterances [5] with a low value of NDR are the most articulated among the chanting instances. Third, the 4 chanting utterances [6] with the smallest values of NDR and 100-ms note proportion are somehow quite close from their neighbour utterances in term of note durations. Lastly, a long syllable is found in the bottom-right speech utterance [7] with the greatest note duration range.

4.3 Evaluation by supervised classification

In order to evaluate quantitatively our features, supervised classification were performed with the database and the best classifier was selected [8]. Results are reported below in the form of confusion matrix with different groupings among the 6 vocal categories. The rows correspond to true categories while the columns are the predicted categories. For instance, in table 3, the utterances labelled as chanting by the ethnomusicologists were correctly classified by the algorithm as 57%, while 22% were classified in the singing category and none was classified in the lament (0%).

Classification was done using the following features: note duration distribution, note duration range, normalized total duration of detected notes (TotDurNote), mean instantaneous note number (InstNoteNum), voicing proportion (VoicProp), peak number and peak width of the chroma spectrum.

First, classification with all the 6 categories gives some quite bad results (see table 3), including the storytelling class with a true prediction score less than with a random classification (for 6 classes: 17%) and the others between 25% and 57%.

As the speech classes are not well discriminated between each other, we grouped the talking, storytelling and recitation together to form the speech class, which is detected with a 82% rate (for a total of 28 utterances) against the chanting and singing classes (see top table 4). The chanting class gets a 65% score of true prediction while singing class performs 74% of true detection.

If grouping singing and chanting categories as one, named song, then about 9/10 of the sound examples are well classified (bottom table 4).

If considering song, speech and lament as three categories to be classified, lament are classified as song for 90% of their utterances, 10% as speech, and 0% as lament

[3] 10 first seconds of CNRSMH_I_2012_004_001_002 and CNRSMH_E_1992_007_002_001_11

[4] 10 first seconds of CNRSMH_I_1971_025_003_020_12s-39s, CNRSMH_I_2003_010_001_04, CNRSMH_E_1959_002_004_003_02_1mn14-end

[5] 10 first seconds of CNRSMH_I_1970_068_016_01_30s-end, CNRSMH_I_1975_015_017_05, CNRSMH_I_1972_013_025_02_1s-end, CNRSMH_I_1975_015_017_02, CNRSMH_I_1972_012_015_05_6s-end, and CNRSMH_I_1972_012_015_03

[6] 10 first seconds of CNRSMH_I_1970_068_010_02, CNRSMH_I_2011_016_008_04_37s-end, CNRSMH_I_1970_071_009_04_55s-end, CNRSMH_I_1959_006_001_01

[7] 10 first seconds of CNRSMH_I_2007_005_044_01_30s-end, CNRSMH_I_1974_003_006_01_2s-end

[8] Among Naive Bayes, Classification Tree, SVM, kNN, Neural Network, Random Forest, CN2 rules.

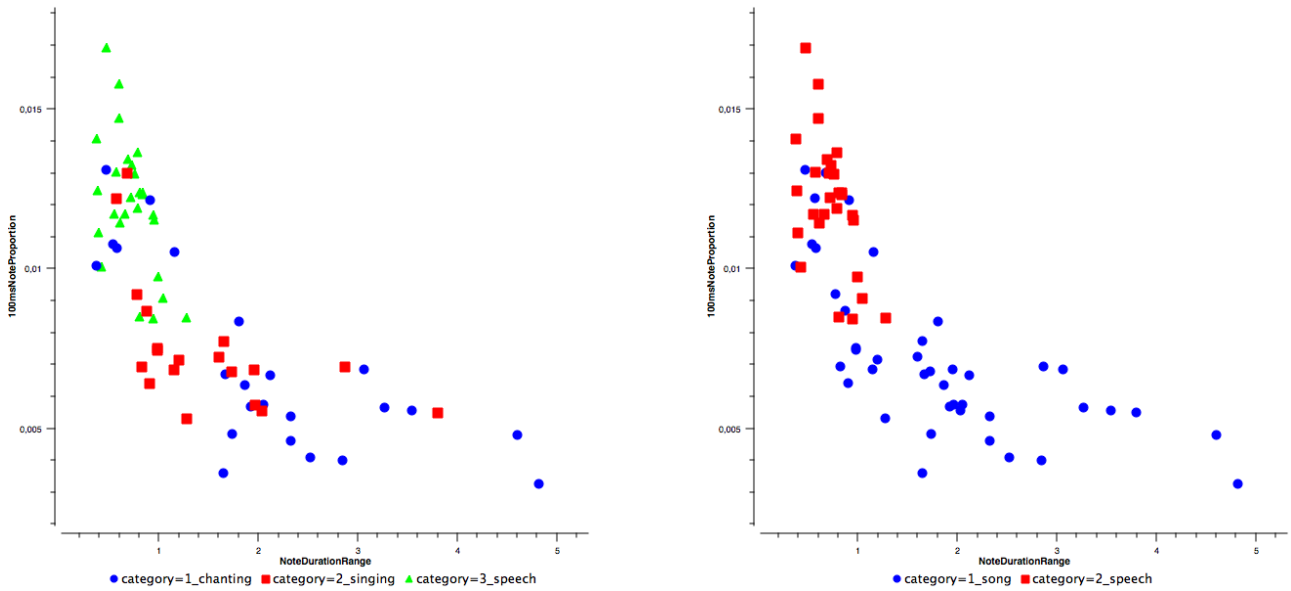

Figure 5: Utterance position in a 2D space according to their values of NDR and 100-ms note proportion. Left: Chanting VS Singing VS Speech (i.e. recitation+storytelling+talking). Right: Song (chanting+singing) VS Speech.

	Chanting	Singing	Recitation	Storytelling	Talking	Lament
Chanting	57%	22%	13%	4%	4%	0%
Singing	5%	53%	0%	0%	16%	26%
Recitation	50%	12%	25%	0%	13%	0%
Storytelling	12%	0%	25%	13%	50%	0%
Talking	25%	17%	0%	17%	33%	8%
Lament	20%	30%	0%	0%	10%	40%

Table 3: Confusion Matrix of the 6-classes classification (proportions of true) using a Classification Tree and a cross validation procedure. Columns represent predictions, rows represent true classes.

while song and speech respectively reach a score of 79% and 89%. On the whole, grouping song and lament categories improves the true prediction score: 90% for the category song+lament and 86% for speech. It means that our lament category is closer to song than speech.

	Chanting	Singing	Speech
Chanting	65%	18%	17%
Singing	16%	74%	10%
Speech	0%	18%	82%

	Song	Speech
Song	86%	14%
Speech	7%	93%

Table 4: Confusion Matrix (proportions of true) of the 3-classes (resp. 2-classes) classification using a Naive Bayes (resp. Random Forest) learner and cross validation. Columns represent predictions, rows represent true classes.

5. DISCUSSION AND PERSPECTIVES

We studied some features computed mostly from note duration distribution, and we discussed the characterization of categories by the proportion of 100-ms note and the note duration range. Nevertheless, the discrimination is reached with the help of the whole features, including all the note duration distribution and additional pitch features like voicing proportion, normalized total duration of detected notes, and mean instantaneous note number. Chroma features were found relatively non-discriminating.

The results show that these features are useful for classifying singing, chanting and speech but not for discriminating speech categories (storytelling, recitation, talking). The lament category, chosen for the presence of icons of crying, was found to be closer to song than to speech. Our study confirm the previous work by Sotiropoulos (2014) that it is possible to group vocal productions in acoustically coherent categories.

From an ethnomusicological point of view, these results bring new perspective on the definitions of vocal categories. Classic ethnomusicological approach focuses on endogenous categorizations of musical practices, thus specific to each culture and never solely based on acoustic criteria. The DIADEMS project ambitioned to build a transver-

sal characterization of vocal categories sampled from different cultures and based only on acoustic parameters.

The aim of this study is to test acoustic parameters on several scientific categories defined and characterized by ethnomusicologist. This work focuses more on the characterizations of descriptors and acoustic parameters, rather than on the definition of the categories themselves.

Results show that talking, storytelling and recitation categories can not be distinguished based only on the acoustic features here tested. Results support the definitions given by ethnomusicologists: talking and storytelling only differ by the mode of realization (monologue/dialogue), which is not embedded in pitch features. Furthermore, results corroborate the distinction made by ethnomusicologists between speech and song and their respective subcategories (talking, storytelling, recitation concerning speech and singing, chanting concerning song).

The results show that the lament category seems closer to song based on pitch features. This classification does not support the ethnomusicological definition proposed by Urban but brings new acoustic characterization. To test the lament category and the icons of crying proposed by Urban, other parameters than note duration distribution should be tested. Results may be also biased by our test dataset and must be further tested using other datasets.

Concerning implementation, this system will be improved and completed by other features from the literature (Sotiropoulos, 2014) and other type of pitch features, especially taking into account the time evolution of the detected notes. Our work is intended to be included in the timeside library [9], an open source plugin used in the Telemeta interface as a graphical help for ethnomusicologists. However, the final user interface in Telemeta should not give predicted category names only, but rather more subtle raw informations to help the ethnomusicologists with the data indexation.

6. REFERENCES

Amy de la Bretèque, E. (2010). Des affects entre guillemets. mélodisation de la parole chez les Yézidis d'Arménie. *Cahiers d'ethnomusicologie*, *23*, 131–145.

Dannenberg, R. B. & Goto, M. (2008). *Handbook of Signal Processing in Acoustics*, chapter Music Structure Analysis from Acoustic Signals, (pp. 305–331). Springer New York.

Fillon, T., Simonnot, J., Mifune, M.-F., Khoury, S., Pellerin, G., Le Coz, M., Amy de la Bretèque, E., Doukhan, D., & Fourer, D. (2014). Telemata: An open-source web framework for ethnomusicological audio archives management and automatic analysis. In *Proc. 1st Digital Libraries for Musicology workshop (DLfM 2014)*, London, UK.

Fujihara, H. & Goto, M. (2007). A music information retrieval system based on singing voice timbre. In *8th International Society for Music Information Retrieval Conference*.

Gärtner, D. (2010). Singing / rap classification of isolated vocal tracks. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, (pp. 519–524).

Giannattasio, F. (2007). *Musiques. Une encyclopédie pour le XXIe siècle. L'unité de la musique* (Actes Sud/ Cité de la musique ed.)., volume 5, chapter Du parlé au chanté: typologie des relations entre musique et texte, (pp. 1050–1087). Arles, Paris.

Goldman, J.-P., Auchlin, A., & Simon, A. C. (2009). Discrimination de styles de parole par analyse prosodique semi-automatique. In *Actes d'IDP*, (pp. 2114–7612)., Paris.

Harte, C. & Sandler, M. (2005). Automatic chord identification using a quantised chromagram. In *Audio Engineering Society Convention*, volume 118.

Lartillot, O., Toiviainen, P., & Eerola, T. (2007). A matlab toolbox for music information retrieval. In *Proc. of International Conference on Digital Audio Effects*, Bordeaux.

Léothaud, G. (2007). *Musiques. Une encyclopédie pour le XXIe siècle. L'unité de la musique* (Actes Sud/ Cité de la musique ed.)., volume 5, chapter Classification universelle des types de techniques vocales, (pp. 803–832). Arles, Paris.

List, G. (1963). The boundaries of speech and song. *Ethnomusicology*, *7*(1), 1–16.

Liu, Y., Xiang, Q., Wang, Y., & Cai, L. (2009). Cultural style based music classification of audio signals. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (pp. 57–60).

Picard, F. (2008). Parole, déclamation, récitation, cantillation, psalmodie, chant. *Revue des Traditions Musicales des Mondes Arabe et Mediterraneen (RTMMAM)*, 8–24.

Rappoport, D. (2005). Les langues frétillantes. Modalités de profération de la parole rituelle chez les Toraja d'Indonésie. In *Second Congress of Asia Network*, Paris. EHESS.

Seeger, A. (1986). *Native South American Discourse* (Mouton de Gruyter ed.)., chapter Oratory is Spoken, Myth is Told, and Song is Sung: But They Are All Music to My Ears, (pp. 59–82). Berlin.

Sotiropoulos, T. (2014). Caractérisation des voix intermédiaires : de la taxinomie des voix dans un contexte ethnomusicologique. Master's thesis, Université Paul Sabatier, équipe SAMoVA.

Stables, R., Athwal, C., & Bullock, J. (2012). *Speech, Sound and Music Processing: Embracing Research in India*, volume 7172 of *Lecture Notes in Computer Science*, chapter Fundamental Frequency Modulation in Singing Voice Synthesis, (pp. 104–119). Springer Berlin Heidelberg.

Urban, G. (1988). Ritual Wailing in Amerindian Brazil. *American Anthropologist*, *90*(2), 385–400.

Zemp, H., Léothaud, G., Lortat-Jacob, B., Tran, Q. H., & Schwarz, J. (1996). *Voices of the World. An Anthology of Vocal Expression*. Collection CNRS-Musée de l'Homme / Le Chant du Monde CMX 374 1010.12.

[9] https://github.com/yomguy/TimeSide/