SPEECH, MUSIC AND SONGS DISCRIMINATION IN THE CONTEXT OF HANDSETS VARIABILITY

Hassan Ezzaidi and Jean Rouat

ERMETIS, DSA, Université du Québec à Chicoutimi,
555 boul. de l'Université, Chicoutimi, Québec, Canada G7H 2B1.
http://wwwdsa.uqac.uquebec.ca/ermetis

1. ABSTRACT

The problem of speech, music and music with songs discrimination in telephony with handsets variability is addressed in this paper. Two systems are proposed. The first system uses three Gaussian Mixture Models (GMM) for speech, music and songs respectively. Each GMM comprises 8 Gaussians trained on very short sessions. Twenty six speakers (13 females, 13 males) have been randomly chosen from the SPIDRE corpus. The music were obtained from a large set of data and comprises various styles. For 138 minutes of testing time, a speech discrimination score of 97.9% is obtained when no channel normalization is used. These performance are obtained for a relatively short analysis frame (32ms sliding window, buffering of 100 ms). When using channel normalization, an important score reduction (on the order of 10 to 20%) is observed. The second system has been designed for applications requiring shorter processing times along with shorter training sessions. It is based on an empirical transformation of the ΔMFCC that enhances the dynamical evolution of tonality. It yields in average an acceptable discrimination rate of 90% (speech-/music) and 84% (speech, music and songs with music).

2. INTRODUCTION

As the digital storage and processing of audio signals are increasing and become popular, speech and music discrimination systems are crucial to extend the functionality of various information and communication systems. In Multimedia applications, such systems can be useful to achieve automatic classification, indexation, archiving and retrieving of information from large multimedia databases [5]. Speech and music discrimination can also play a significant role in speaker or speech recognition systems, by rejecting non speech segments. The new generation of low bit rate coders and compression technologies need an estimation of the signal nature in order to achieve a better compression. Therefore, a fast and efficient speech/music discriminator is crucial for that type of coders as the performance is strongly related to the accuracy of the speech/music discriminator [4].

To retrieve and achieve a good discrimination, the dynamical time evolution of the vocal tract, during the speech production (vowels, consonants, coarticulation, etc.), is one salient property that has been strongly exploited in many ways in the literature. In brief, the vocal signal is characterized by a formantic and dynamical structure contrary to the musical signal that is rather characterized by an harmonic and regular structure over longer durations. Generally, the proposed techniques are simple in order to be applicable to a vast form of music style and to be independent of speaker and speech for most cases. Saunders [6] proposes four features derived principally from the zero crossing rate and the energy contour. A set of 13 features related to the amplitude, fine spectrum and signal frequency also are studied by Scheirer and Stanley [7]. Carey et *al.* [2] examine the discrimination achieved by several different features (filterbank energy, cepstre, pitch and zero-crossing) using common training and test sets and the same classifier. Samouelian and *al.* [5] perform the automatic labeling and classification of TV broadcast material into speech, music, silence and noise segments. El-Maleh and *al.* [3] suggest the use of line spectral frequencies (LSFs) and zero-crossing-based features for frame-level narrowband speech/music discrimination. Recently, Ajmera and *al.* [1] use a posterior probability based entropy and dynamism features that are integrated over time through a 2 state HMM. Generally, the authors use features that are estimated over a long time range (0.5 to 5 seconds). One exception to this, is the work by El-Maleh et *al.* [3], where duration tests as short as 20 ms are used. Their scores with such a short window length are still interesting and average to 82.5% for the speech segments and 79.2% in the music case.

In the present work we are interested in a system that performs indexation and retrieving of telephone speech information from a corpus that comprises conversations in alternance with music or songs. The indexation system performs first the speech/music discrimination and then, the speaker identification and speech recognition tasks. The design of the discrimination system is therefore constrained to the use of the MFCC vectors like the speaker/speech recognition systems. Furthermore, performance of the discrimination system should be independent on the recording conditions and more specifically of the telephone handset variability. We propose two discrimination systems. The first system uses a parametric multi-Gaussian modeling that comprises 8 mixtures of Gaussians per model. The second system is based on the comparison of distances between ΔMFCC derived features and a preset threshold.

3. DATABASE AND EVALUATION CRITERION

3.1. Database

The data were downloaded from various broadcasting Internet sites and from the university telephone switchboard. The audio quality ranges from narrow-bandwidth of AM radios (3 kHz) to high fidelity music (16 kHz bandwidth). The music files where then played trough a loudspeaker. A telephone handset was placed in front of the loudspeaker to transmit the music through the tele-

This work has been partially supported by NSERC and CSE.

phone network (digital in the university, analogue elsewhere). This operation limits the bandwidth to that of the telephone lines and introduces to a certain extend the channel effect and handset variability. The recordings were carried out during several sessions and comprises various types of music and songs.

Table 1. Music database with style description and durations respectively for the training (GMM system) and the testing sessions.

Styles for Music only	Training time (mn)	Testing time (mn)
Classic	1.77	11.6146
Standard	0.5	5.4354
Jazz	0.5	5.5942
Rock	0.04	0.3517
Metal	0.5	2.4975
Western	0.9	3.6269
Country	0.54	5.4231
Total for Music	4.75	34.5434
Styles for Music with Songs	Training time (mn)	Testing time (mn)
Blues	0.5	1.7808
Country	1.5	3.5712
Rap	1.5	2.4879
Rock	1.0	2.4985
Films	0.62	9.4087
Various	1.16	8.7456
Reggae	0.33	1.2098
Total for Music with Songs	6.61	29.702
Total for Speech	6.00	74.2169
Total Duration	17.3	138.4628

Table 1 gives the different music style and time durations used during training and testing.

A subset of the speech SPIDRE–Swichboard Corpus is used to extract randomly speech data. Each speaker has 4 conversations originating from 3 different telephone handsets. 74 minutes of speech that comprises all conversations taken from ten males and ten females have been used for testing. In addition, we randomly choose one conversation of 3 women and 3 men from which we extracted 1s when training was necessary. The total speech duration in the training session is therefore of 6 minutes. The training speakers are not presented during the test. All the database is sampled at 8 Khz.

3.2. Discrimination score

The discrimination score $S = \frac{C}{T}$ is defined as the ratio of the number C of correctly classified frames over the number T of tested frames.The same score S is used for all the reported experiments.

4. ML SPEECH MUSIC DISCRIMINATION

In this section we design a Gaussian Mixture Models speech/music discriminator to study the influence of channel normalization – commonly used to improve speech recognition or identification rates in the context of telephone handsets variability – and the benefit of using a model for music with songs.

4.1. Perspectives and models

In a first set of experiments, one model for speech and another for music (with or without songs) are used. In that situation, – in the context of music with singer(s) – depending on their relative dominance the system will decide whether it is speech or music. A second set of experiments is designed where 3 models are used (speech, music only, music with voice or singer(s)).

4.2. Training session

For each style of music (Classic, Western, Pop, Rock, Jazz, Soft music, Instrumental and Rap), only a few seconds have been used to train the music GMM. Also, for each style of songs, a few seconds of signal, are extracted from the following categories: Pop, Rock, Blues and Rap to train the GMM songs. The speech model is trained on 6 minutes taken from 6 conversations of three men and three women and originating from different telephone handsets. Each conversation corresponds to one minute duration. None of the training data are used for the testing.

4.3. Testing session

A sliding window of 32ms length and 10ms shift is placed on the signal to estimate every 10ms the log–likelihood of the parameter observation for each model. Results are presented for averages over 100ms (10 frames) or 200ms (20 frames) of the log–likelihood. Finally, the model with the maximum averaged log–likelihood is retained.

4.4. Results

4.4.1. Influence of the channel normalization

Table 2. Parametric models and discrimination scores. 2 models: one GMM for Speech, one for Music **and** Songs; 3 models: one GMM for Speech, one for Music, one for Songs; CN: Channel Normalization. Sg in the music styles indicates presence of songs or singer with the music.

	2 models, no CN		2 models, CN		3 models, no CN	
Styles	100ms	200ms	100ms	200ms	100ms	200ms
classique	93.35	94.52	86.03	88.51	97.79	98.50
Standard	96.84	97.16	90.03	91.48	98.96	98.98
Jazz	91.66	94.13	85.38	88.85	98.53	99.45
Rock	91.31	93.13	84.73	85.76	98.01	98.88
Metal	98.37	99.33	93.98	97.22	99.28	99.62
Western	87.65	88.92	69.19	67.51	95.52	96.74
Country	80.76	80.89	75.18	75.40	95.46	97.68
$\overline{Music}$	91.42	92.58	83.50	84.96	97.65	98.55
BluesSg	54.80	52.33	55.75	51.34	72.06	71.36
CountrySg	78.19	79.33	61.29	57.72	94.37	96.11
RapSg	71.80	72.04	42.54	32.62	95.05	98.11
RockSg	61.84	55.18	37.06	25.88	85.04	91.26
Films	80.07	80.17	74.23	74.02	88.3111	88.37
VariousSg	87.20	88.23	74.24	74.38	95.0686	96.19
ReggaeSg	63.08	57.89	36.82	25.86	92.27	94.01
$\overline{MusicSg}$	71.00	69.31	54.56	48.83	88.88	90.77
$\overline{Speech}$	92.46	96.62	89.96	96.52	94.28	97.93
$\overline{Total}$	86.84	88.78	79.49	81.71	93.77	96.29

Channel Normalization (CN) – subtraction of the mean feature vector from each of the feature vectors – is usually used to reduce the influence of the channel variability over the speaker or speech recognition systems. We study the influence of CN by comparing the discrimination rates for classification achieved with or without channel normalization. In each case, two GMM models are used (one for Speech, another for Music **and** Songs). Averaged scores are given respectively for 100ms and 200ms test time durations (Table 2 (columns 2 to 5). The first column is the name category.

By comparing the columns, we notice that the scores discrimination of speech remains almost similar (92.46% without normalization, 89.96% with normalization for 10 frames averaging; 96.62% without normalization, 96.52% with normalization for 20 frames averaging). However, the discrimination is much better on music and songs when no normalization is used. Particulary, the discrimination for music is reduced from 92.58% to 84.96% when normalization is used (20 frames averaging). In fact, the normalization, unavoidably removes some features that are supposed to characterize principally the regularity of music tonality. Therefore the normalization is inappropriate for music signal. When the music is mixed with singers the difference is greater (from 69.31% to 48.83% with 20 frames averaging). This situation is probably related to the fact that the music regularity is perturbed and speech singer is improved when the normalization is taken into account. Here we can retain that the normalization is inappropriate for parameters deriver from music and song signals.

4.4.2. *Integration of a model for Songs*

In order to increase the discriminability, we introduce a third GMM model that is dedicated to music mixed with songs. An experiment is performed with three GMM models. The first is trained with speech, the second with music only and the third with songs mixed with music. The same data and durations than in previous subsections have been presented to the system.

It is known that during all training sessions, exhaustive data should be presented to the system in order to better adapt the model parameters (various speaker conversations and music styles are required). For this purpose, the computational time is expensive and can even exceeds the hardware limits. Also, the number of mixtures can increase considerably to accurately take into account all the situations and styles. Another alternative would be in modelling separately each music style with a different model. However the execution time would be considerably longer. We think that such training technique is intended to characterize the intra-speaker and intra-music styles variabilities. We do not need such accurate models but prefer models that loosely estimate the distributions of speech, music and music with songs, as we want general characteristics for each classes while achieving a good discrimination. So, we propose here to limit the number of GMMs to 8 and to use short duration data for training. Even if in previous works, up to 64 number of mixture were used for the same problem, we show that good discrimination is obtained by using 8 mixtures of Gaussians per model of classes.

Discrimination scores are reported in table 2 (columns 6 and 7). No normalization has been performed. An averaged discrimination score of 93.77% (10 frames averaging) or of 96.29% (20 frames) is obtained in comparison to 86.83% and 88.78% for the same conditions but with 2 GMMs (one for speech, one for music and songs). Therefore, the use of 2 models to differentiate music and music with songs (instead of one model for both) is a good strategy.

5. A FASTER SPEECH-MUSIC DISCRIMINATOR: TRACKING THE SHORT–TERM VARIABILITY BASED ON ΔMFCC

5.1. Feature extraction

We extract features that exploit the short–time variability and irregular duration of speech tonalities. In a preliminary study we observed that ΔMFCC yield a better discrimination than MFCC. Therefore, in the following, ΔMFCC are used as basic features. For each 32 ms frame, 12 delta coefficients are obtained. The 32 ms frame is shifted every 10 ms. The dynamic tonality indirectly is captured by different measurements. We first compute the Euclidian distance between $\Delta MFCC$ parameter vectors from 3 adjacent segments. Let us write $\Delta MFCC_i(n)$ the $i\ th$ dimension of vector $\Delta MFCC(n)$, where $\Delta MFCC(n)$ is the vector of 12 delta Mel Cepstrum Coefficients computed at frame n. We define $d_1(n)$ and $d_2(n)$ as:

$$d_1(n) = \sum_{i=1}^{12} (\Delta MFCC_i(n) - \Delta MFCC_i(n-1))^2 \qquad (1)$$

$$d_2(n) = \sum_{i=1}^{12} (\Delta MFCC_i(n) - \Delta MFCC_i(n-2))^2 \qquad (2)$$

It is observed that the time evolution of the two distances between segments is almost stable with weak fluctuations for all the music segments, while significant fluctuations are observed for speech segments. Then, the standard deviation of these distances over a duration of 100 ms is calculated.

$$\sigma_1(n) = \frac{\sqrt{\sum_{i=n-L}^{n} (d_1(n) - \overline{d_1})^2}}{L} \qquad (3)$$

$$\sigma_2(n) = \frac{\sqrt{\sum_{i=n-L}^{n} (d_2(n) - \overline{d_2})^2}}{L} \qquad (4)$$

In preliminary experiments, we found that speech is always characterized by strong transitions and strong amplitude of $\sigma_i(n)$, which is not the case for the music. This motivates us to define a parameter $P_\sigma(n)$ that takes into consideration this observation.

$$P_\sigma(n) = (\sigma_1(n) - \sigma_2(n))^2 + max(\sigma_1(n), \sigma_2(n)) \qquad (5)$$

Based on this parameter, we proceed now in the classification experiments between *Speech*, *Music* and *Music with Songs*.

5.2. Decision boundaries

In this section, we study the possibility of discriminating speech from any kind of music or songs by finding the boundary between the classes. As the parameter $P_\sigma(n)$ is scalar, the Bayesian classification scheme is reduced to the problem of finding the hresholds that yield the optimal results. As we suppose the apriori equiprobability of the classes, minimizing the Bayesian cost function is equivalent in maximizing the recognition rate when a lossless decision matrix is used.

In next subsection, we show that speech/music classification can be succesfully achieved with only one threshold instead of two.

5.3. Results

Three case studies are presented and illustrated (Figure 1).
1. *Speech*/*Music* **discrimination:** (Fig. 1 (a))
The optimal threshold is located at the curves intersection with approximately a value of 0.001 for the threshold and a discrimination score of 90%. However, if one wants to introduce the notion of loss associated to decisions by favoring the recognition of one class relatively to another, the optimal threshold should be moved to the right or to the left direction A better discrimination of speech is observed when the threshold is lower. On the opposite, when the threshold is increased, the scenario is completely reversed (good discrimination for music but not for speech).
2. *Speech*/*Music with Songs* **discrimination:** (Fig. 1 (b))
The optimal threshold is moved considerably to the right, makes the average discriminant score fall to 80%. This means that the

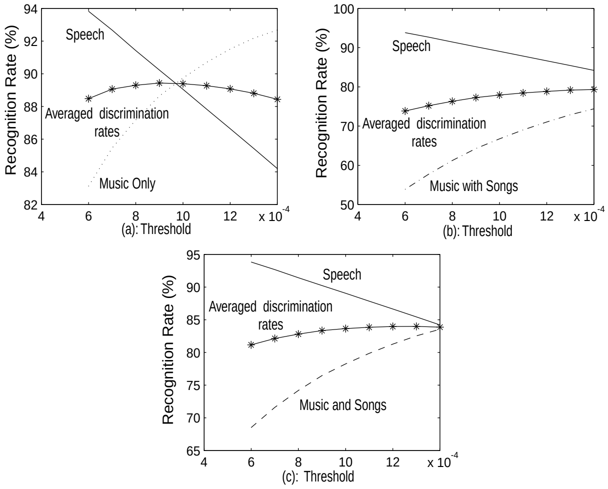

Fig. 1. System performance with threshold strategy: **(a)**: classification of *Speech*/*Music Only*. **(b)**: classification of *Speech*/*Songs*. **(c)**: classification of *Speech*/*Music with Songs*. Continuous curves in (a),(b) and (c) are the recognition rate for *Speech*. Dot line curves are the recognition rates (a) for *Music Only*, (b) for *Songs* and (c) for both *Music* **and** *Songs*.-*- is the total averaged discrimination rate.

overlap with speech is more significant for the class *Music with Song* than for *Music Only*.

3. **Speech/merged** *Music* **class with the** *Song* **class discrimination:** (Fig. 1 (c))

The optimal threshold has a value of 0.0014 and a score of 84%.

For the reason that 3 classes (*Speech*, *Music*, *Music with Songs*) are considered, the classification should be performed with at least two boundaries (i.e. thresholds). But we observe that the *Music with Songs* class can be merged with the *Music Only* class to create one class. By doing so, the discrimination can still be performed with satisfactory results even if this architecture is not associated with the optimal solution of the problem. An optimal solution would require a greater processing time.

5.4. Discussion

In comparison to most common works that usually use test durations on the order of 1000 ms, we obtain a comparable score with a shorter test duration (100 ms). This represents a factor of reduction of 9/10. To increase the discrimination rate, we can estimate the standard deviations of the proposed distances over durations longer than 100 ms.

On the other hand, for *Music with Songs* and when no specific model is used for *Music* or *Songs*, the performance decreases according to the coupling between the music and the singer signal. We retain that the singer signal which does not dominate the music signal (energy) is recognized as music. On the other hand, if the singer signal dominates and affects the regularity of music signal, it is recognized as speech. Combining an energy tracker and a silence detector could be a good strategy to retrieve the singer segments. But, the processing requires more computations and a greater time analysis.

6. CONCLUSION

We investigated two systems to discriminate *Speech*, *Music* and *Music with Songs* in the context of telephony with handsets variability. The first system uses three Gaussian Mixture Models (GMM) for speech, music and songs respectively. The reference system uses two GMMs, the first for *Speech* and the second for *Music Only* along with *Music and Songs*. Each GMM comprises 8 Gaussians trained on very short sessions and tested on very large sessions.

We have shown that, when channel normalization is used to improve speech recognition or identification rates in the context of handsets variability, then the speech/music discrimination score reduces by 10% for *Music* and by 20% for *Music with Songs*. Indeed, the normalization of feature vectors removed the regularity structure of musical signal.

Over a 100 ms of test duration, the scores obtained for the reference system with normalization and the proposed system respectively are 89.96% and 94.28% in the case of speech, 83.50% and 97.65% in the case of music and 70.67% and 91.91%in the case of song music. The score for test durations of 100 ms and 200 ms are almost similar, so we think that if we reduce the duration to 50 ms the score should not change too much.

The second system is based on an empirical transformation of the ΔMFCC that enhances the dynamical evolution of tonality. Even if it is optimized for integration in real time applications, it yields an acceptable discrimination rate of 84% with a 100 ms test duration. The results are comparable to other scores reported in the literature that use on the average 1 second. Therefore a reduction of 9/10 is observed.

Finally, in the context of handsets variability, when the discrimination system is used in conjunction with Speech or Speaker Recognition, the normalization of the MFCC should be only performed by the recognition systems and not by the speech/music discriminator

7. REFERENCES

[1] Ajmera J., McCowan I. , and Bourlard H., "Robust HMM-Based Speech/Music Segmentation," in *ICASSP'02*, 2002.

[2] Carey M. J., Parris E. S., and Lloyd-Thomas H., "A comparison of features for speech, music discrimination," in *ICASSP'99*, 1999.

[3] El-Maleh K., Klein M., Petrucci G., and Kabal P., "Speech/music discrimination for multimedia applications," in *ICASSP'00*, 2000.

[4] Tancerel L., Ragot S., and Lefebvre R., "Speech/music discrimination for universal audio coding," in *20th Biennal Symposium on Communications*, 2000, pp. 28–31.

[5] Samouelian A., Robert-Ribes J., and Plumpe M., "Speech, silence, music ans noise classification of tv broadcast material," in *ICSLP'98*, 1998.

[6] Saunders John, "Real-time discrimination of broadcast speech/music," in *ICASSP'96*, 1996, pp. 993–996.

[7] Scheirer E. and Stanley M., "Construction and evaluation of a robust multifeature speech/music discriminator," in *ICASSP'97*, 1997, vol. II, pp. 1331–1334.