A SILENCE/NOISE/MUSIC/SPEECH SPLITTING ALGORITHM

Claude Montacié, Marie-José Caraty

LIP6 - Université Pierre et Marie Curie
4, place Jussieu - 75252 Paris cedex 05 - France

ABSTRACT

In this paper, we present techniques to warp audio data of a video movie on its movie script. In order to improve this script warping, a new algorithm has been developed to split audio data into silence, noise, music and speech segments without training step. This segments splitting uses multiple techniques such as voiced/unvoiced segmentation, pitch detection, pitch tracking, speaker and speech recognition techniques.

The 102.47 minutes of the film movie « Contes de Printemps » produced by E. Rohmer have been indexed with these techniques with an average shifting lower than one second between the time-code script and audio data.

1. INTRODUCTION

The interest in video databases is increasing fast. National or corporate archives store millions of hours of video. Presently such archives are indexed manually and can only be accessed through the mediation of human experts, whose availability is limited. One major aim of the video indexing is to obtain the script movie directly from audio data. The first difficulty is to split audio data in silence/noise/music/speech segments.

The great variability of noises (e.g., rumbling, explosion, creaking) and music (e.g., classic, pop) used on the audio-video databases (e.g., broadcast news, movie film) makes difficult an a priori training. This new algorithm of splitting has no training step, and adapt the silence/noise/music/speech models to audio data. In a previous paper [1], we have yet developed a silence/noise/music/speech detection algorithm based on a single Auto-Regressive Vector (ARV) model, the results on the film movie « Un indien dans la ville » produced by H. Palud are poor (i.e., 20% detection rate). Now, multiple techniques are used, such as voiced/unvoiced segmentation, pitch detection, pitch tracking, speaker and speech recognition techniques. Each of these techniques has been adapted to these different kinds of sound. The new algorithm we designed is based on a gradual splitting using a decomposition of the silence/noise/music/speech detection algorithm on three binary decisions. An adapted modeling is used for each decision. At first, a silence/non-silence detection based on an histogram of the short-time energy is developed. Then, each non-silence segment is labeled as noise/non-noise segment using a fusion of energy, voiced/unvoiced segmentation and acoustic-phonetic decoding. At last, speaker recognition techniques based on ARV models are used for a speech/music splitting. At each step, our script warping technique is used as paradigm of evaluation. Of course, vocal dictation [2] could also be used, but this technique is too time-consuming.

2. SCRIPT WARPING

The warping technique [3] is based on the Hidden Markov Model (HMM) technique. A network warping is made from the phonetic transcription of the words of the script. The phonetic dictionary [4] allows word phonetic variants as elisions and liaisons (cf. fig. 1). 37 phonetic models including the silence (sil), represented by 3-states Bakis models, are used. 16 gaussian mixtures represent the HMM state distribution. These mixtures are first trained [5, 6] on a reference database [7], then the audio signal is segmented by the Viterbi decoding algorithm. For each phonetic segment, a segment likelihood coefficient is computed from the segment duration and the segment probability. The mixtures are trained again on the well warped phonetic segments (i.e., for which the segment likelihood coefficient is lower than a given threshold). This procedure is iterated until no segmentation difference is observed. Word segmentation is built from the phonetic segmentation using dynamic programming algorithm.

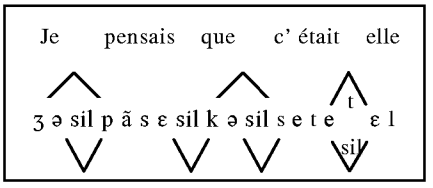

Figure 1 : Example of warping network

2.1. Database and Experiments

The 6,167 seconds of audio signal of the entire film movie "Contes de Printemps" are chosen as database. There are 10 speakers and 14,706 occurrences of words (i.e., 2,132 different words). There are 47 interior and exterior scenes. Many kinds of noise (e.g., slamming of doors, sounds of footstep, motor car noise) and various classical music pieces (e.g., Beethoven, Schumann) are in the audio signal. Only one decoding/training session has been necessary for the script warping. The

segmentation computational cost is about half a day (Pentium II). Three kinds of time-code error have been defined : Phone Shifting segmentation, Word Shifting segmentation (i.e., excluding monophone words), and Sentence Shifting segmentation. But it's taking a long time listening the result of the warping script (i.e., segmentation of 14,706 words). Therefore, 103 words have been selected (i.e., one word per minute). For each step of the gradual silence/noise/music/speech splitting, the time-code of each of the 103 words is presented in the Table 6. In audio data, the reference time-code of the words are found in Table 6-1. The average shifting, between the time-code script and audio data, and the time-code errors are computed from this table. 101 words are only used : two words of the script have no occurrence in audio data.

Phone Shifting	Word Shifting	Sentence Shifting
8	2	22

Table 1 : Warping results without splitting algorithm.

The segmentation results of script warping are not very good. There is many Sentence Shifting segmentation and the average shifting is about 21 seconds (cf. Table 6-2). The worst shifting is about 6 minutes, and is brought by a long classical music segment at the beginning of the movie. But for many segments there are not any shift between speech and script.

3. SILENCE / NON-SILENCE SPLITTING

The first step of the gradual silence/noise/music/speech splitting is the silence/non-silence segmentation. This splitting algorithm is based on the histogram of the short-time energy computed on signal segments of 15 sec duration. This duration is chosen to be higher than the length of a sentence and lower than the duration of a scene. Indeed, a scene change may induce a background noise (e.g., interior/exterior).

For a segment, the average μ and the standard deviation σ of these short-time (i.e., 10 msec) energies are computed. If the histogram can be represented by a single gaussian, 95% of these energies are located between μ-2σ and μ+2σ. In this case, the signal segment is homogeneous and is made up of only silence or non-silence. In the contrary case, the average μ_1 and the standard deviation σ_1 of the silence, the average μ_2 and the standard deviation σ_2 of the non-silence are computed by the k-means algorithm. The threshold silence/non-silence is estimated by $(\mu_1 + (\mu_2 - \mu_1).\sigma_1)/(\sigma_1 + \sigma_2)$. A 4-states automaton (i.e., silence, up, non-silence, down) uses this threshold for an ultimate segmentation.

The silence/non-silence splitting gives 5,141 segments including 2,567 silence segments (i.e., 2,743 sec) and 2,574 non-silence segments (i.e., 2,284 sec). None error is found at listening verification.

4. NOISE / NON-NOISE SPLITTING

Two successive methods are used to split the non-silence segments into noise/non-noise segments. The first one looks for the voiced/non-voiced parts in the non-silence segments, the second one uses an acoustic-phonetic decoding of the voiced segments.

4.1. Voiced / Non-Voiced Splitting

Practically, in all languages a word has to include at least a voiced sound. This voiced sound is characterized by a pseudo-periodicity (i.e., the pitch) and has generally a high energy. The algorithm we use looks for the presence of the periodicity, by the Average Magnitude Difference Function (AMDF) algorithm, in 10% of the higher short-time energy segments. An algorithm of pitch tracking is used to confirm these anchor points.

The voiced/non-voiced splitting algorithm gives 2,574 segments including 514 non-voiced (i.e., noise) segments (i.e., 224 sec) and 2,060 voiced (i.e., non-noise) segments (i.e., 3,219 sec). One error is found at listening verification. At listening, this error is probably due to a low intensity voice overlapping a high noise.

4.2. Training of Noisy Phonetic Models

The non-voiced segments are useful to compute a HMM model of the noise. When this model is used in a HMM recognizer in competition with phonetic models trained on clean database, the segmentation is wrong. Indeed, the segments of noisy speech are mistaken for noise. To compute noisy phonetic models, we choose to add noise to the signal of the French reference database [7] and to re-estimate the phonetic models. This solution is costly enough (30 CPU hours) but efficient. In the future, less costly noise processing techniques can be used ; for example the combination of phonetic HMM models with the noise HMM model [8].

A first script warping uses these noisy phonetic models. The warping network is also modified using a parallel connection between the silence model and the noise model. The average shifting falls down 19 sec (cf. Table 6-3).

Phone Shifting	Word Shifting	Sentence Shifting
18	4	16

Table 2 : Warping results with noisy phonetic models and noise model.

The second script warping uses the previous warping network, moreover an average frame of silence computed over the entire movie is substituted for each frame of silence or noise. The average shifting falls down 13 sec (cf. Table 6-4).

Phone Shifting	Word Shifting	Sentence Shifting
19	4	14

Table 3 : Warping results with the voiced/non-voiced splitting algorithm.

We remark the noise model is insufficient (e.g., HMM), the removal of noisy segments and silence segments is necessary.

4.3. Acoustic-Phonetic Decoding of Voiced Segments

The second way of searching for the noise segments is based on the acoustic-phonetic decoding of the voiced segments. The recognizer uses the noise model and the noisy phonetic models. When the phonetic decoding of a segment does not include any phoneme, the segment is considered as a noise segment.

The phonetic decoding of the voiced segments gives 143 additional noise segments (i.e., 83 sec). None error is found at listening verification. The average shifting falls down 8 sec (cf. Table 6-5).

Phone Shifting	Word Shifting	Sentence Shifting
20	4	12

Table 4 : Warping results with phonetic decoding of voiced segments.

5. MUSIC / SPEECH SPLITTING

The algorithm of music/speech splitting uses two types of information : the percentage of voiced parts in a segment (i.e., voice rate) and the speaker-recognition distance of a segment in relation to a set of speech segment references of the movie. This voice rate is defined as the ratio of the duration of the voiced parts and the duration of the non-noise parts in a segment.

The set of speech segments is made to choose the segments with a voice rate between 40% and 60% with less than 10% of noisy parts. 119 speech segments (i.e., 6 minutes) are selected in this way in the database. None error is found at listening verification.

5.1. Music Detection technique

At first, the algorithm of music detection searches a set of music segments in the non-noise segments using the voice rate and the speaker recognition technique. Then the other music segments are selected using this set. Our speaker (and music) recognition technique is based on the Auto-Regressive Vector model. ARV model is successfully used in speaker recognition [9].

In the first step, the non-noise segments are labeled as music in the two following cases -the segments having a voice rate higher than 90%, and -the segments with a voice rate higher than 80% having a distance to the speech segments higher than a threshold. The first step labels 15 music segments (i .e., 149 sec). 2 errors are found at listening verification. The wrong segments are two short speech segments characterized by screams and a high pitch.

The second step searches for the remaining non-noise segments near from the music segments yet labeled. This step is iterated until no new music segment is found. This method find 5 additional music segments (i.e., 130 sec). Two errors are found at listening verification. The wrong segments are short segments characterized by phone ringing.

Phone Shifting	Word Shifting	Sentence Shifting
22	2	4

Table 5 : Warping results with music segments detection.

The average shifting falls down 1 sec (cf. Table 6-6). The residual error is due to an incomplete music segments detection.

6. CONCLUSIONS

A new algorithm has been developed to split audio data into silence, noise, music and speech segments. This algorithm doesn't use training data. With this splitting technique, the average shifting is reduced about 95% and new information useful for indexing is found : noise and music segmentation. A future extension of this work is the identification of music segments in few clusters (e.g., classical, jazz, pop) and the detection of scene changes using the background noise.

7. REFERENCES

1. Montacié, C., and Caraty, M.-J., « Sound Channel Video Indexing », *Eurospeech*: 2359-2362, 1997.
2. Woodland, P.C., Hain, T., Johnson, S.E., Niesler, T.R., Tuerk, A., and Young, S.J., « Experiments in Broadcast News Transcriptions », *ICASSP*: 909-912, 1998.
3. Weatley, B., Doddington, G., Hemphill, C., Godfrey, J., Holliman, E., Mac Daniel, J., and Fisher, D., « Robust Automatic Time Alignement of Ortographic Transcription with Unconstrained Speech », *ICASSP*: 533-536, 1992.
4. Caraty, M.-J., and Montacié, C., « Dynamic Lexicon for a Very Large Vocabulary Vocal Dictation », *Eurospeech*: 2359-2362, 1997.
5. Young, S.J., « HTK Version 1.4 : Reference Manual and User Manual », *Cambridge University Engineering Department - Speech Group*, 1992.
6. Barras, C., Caraty, M.-J., and Montacié, C., « Temporal Control and Training Selection for HMM-based System », *Eurospeech*: 2359-2362, 1995.
7. Gauvain, J.-L., Lamel, L.-F., and Eskénazi, M., « Design Considerations and Text Selection for BREF, a large French read-speech Corpus », *ICSLP*: 2359-2362, 1990.
8. Matrouf, D., and Gauvain, J.-L., « Model Compensation for Noises in Training and Test Data », *ICASSP*: 831-834, 1997.
9. Montacié, C., Deléglise, P., Bimbot, F., and Caraty, M.-J., « Cinematic Techniques for Speech Processing : Temporal Decomposition and Multivariate Linear Prediction », *ICASSP*: 153-156, 1992.

1	2	3	4	5	6	Word	1	2	3	4	5	6	Word
404.3	34.3	53.3	53.7	53.7	381.8	bonjour	3163.9	3163.9	3163.6	3163.6	3163.6	3163.6	enseignement
409.6	61.7	118.0	132.6	378.5	409.6	pensais	3235.8	3235.7	3235.7	3235.7	3235.7	3235.7	intellectuelle
413.4	134.2	128.2	183.9	404.4	413.4	pensais	3287.9	3287.8	3287.9	3287.9	3287.9	3287.9	psychanalyse
415.7	233.9	174.0	378.5	409.9	415.7	simplement	3313.8	3313.8	3313.8	3313.8	3313.8	3313.8	transcendantale
424.4	281.9	319.1	413.5	424.4	424.4	simplement	3407.8	3407.7	3407.7	3407.7	3407.7	3407.7	naturellement
428.4	326.9	347.2	420.2	435.7	435.7	comprends	3473.6	3473.6	3473.6	3473.6	3473.6	3473.6	mathématiques
446.7	410.5	407.9	446.4	446.5	446.5	demander	3520.9	3520.8	3520.8	3520.8	3520.8	3520.8	recommencera
468.9	453.9	453.9	468.4	468.4	468.4	meilleures	3564.6	3564.5	3564.6	3564.6	3564.6	3564.6	impression
492.2	492.2	492.0	492.0	492.0	492.0	effectivement	3633.0	3633.0	3630.9	3632.8	3632.8	3632.8	quotidiennes
550.0	549.9	549.9	550.0	550.0	550.0	Montmorency	3705.8	3705.8	3706.1	3706.1	3706.1	3706.1	intelligente
679.5	649.8	679.8	668.0	680.0	680.0	précipitamment	3776.8	3777.0	3777.2	3777.2	3777.2	3777.2	enthousiasmante
708.0	698.2	708.3	708.3	708.3	708.3	spécialement	3780.2	3780.2	3780.2	3780.2	3780.2	3780.2	enthousiasmes
747.0	747.0	747.0	747.0	747.0	747.0	complètement	3858.4	3858.4	3858.4	3856.5	3856.5	3856.5	éventuellement
828.0	828.0	826.0	826.0	828.0	828.0	appartements	3949.3	3949.3	3949.3	3949.3	3949.3	3949.3	institutrice
846.5	846.5	846.5	846.5	846.5	846.5	précisément	3962.6	3962.6	3962.5	3962.5	3962.5	3962.5	certainement
936.7	936.7	936.7	936.7	936.7	936.7	conservatoire	4021.2	4021.3	4022.8	4021.3	4021.3	4021.3	tranquillement
1014.8	1014.7	1014.9	1014.9	1014.9	1014.9	nécessairement	4135.2	4135.2	4135.2	4135.2	4135.2	4135.2	heureusement
1052.2	1052.0	1052.2	1052.2	1052.2	1052.2	administratif	4190.4	4190.4	4190.4	4190.4	4190.4	4190.4	simplement
absent	1104.8	1128.0	1093.6	1107.4	1107.4	vitrage	4257.6	4257.6	4257.6	4257.6	4257.6	4257.6	malheureusement
1178.5	1178.5	1178.5	1178.5	1178.5	1178.5	complètement	4291.3	4291.2	4291.2	4291.2	4291.2	4291.2	contrecoeur
1226.1	1226.1	1226.1	1226.1	1226.1	1226.1	marteau-piqueur	4346.5	4346.5	4346.5	4346.5	4346.5	4346.5	pratiquement
1309.3	1309.3	1309.3	1309.3	1309.3	1309.3	actuellement	absent	4419.0	4420.4	4436.7	4436.7	4436.7	comprends
1372.8	1372.7	1372.7	1372.7	1372.7	1372.7	administration	4480.5	4480.4	4480.4	4480.4	4480.4	4480.4	raccompagner
1386.6	1386.6	1386.6	1386.6	1386.6	1386.6	incompatibles	4527.5	4527.9	4527.4	4527.4	4527.4	4527.4	invraisemblable
1466.3	1466.3	1466.3	1466.3	1466.3	1466.3	journalistique	4608.0	4608.0	4608.1	4608.1	4608.1	4608.1	fougueusement
1510.6	1510.6	1510.7	1510.7	1510.7	1510.7	épouvantable	4675.3	4675.3	4675.3	4675.3	4675.3	4675.3	caractérielles
1553.3	1568.9	1562.3	1553.2	1553.2	1553.2	comprendra	4713.1	4713.1	4713.1	4713.1	4713.1	4713.1	quelquefois
1575.6	1684.0	1705.0	1627.9	1625.9	1575.6	indulgente	4748.7	4748.7	4748.6	4748.6	4748.6	4748.6	caractérielle
1577.0	1699.6	1709.8	1649.4	1649.4	1622.1	commencer	4808.7	4808.7	4808.7	4808.7	4808.7	4808.7	scientifique
1789.0	1789.0	1788.9	1788.9	1788.9	1788.9	appartement	4864.0	4875.6	4875.6	4862.8	4862.8	4864.7	peut-être
1803.2	1803.2	1803.2	1803.2	1803.2	1803.2	rapporter	4967.0	4967.0	4967.0	4967.1	4967.1	4967.0	connaissez
1867.5	1867.5	1865.3	1867.6	1867.6	1867.6	heureusement	5036.6	5036.5	5036.5	5036.5	5036.5	5036.5	précipitamment
1959.4	1959.4	1959.4	1959.4	1959.4	1959.4	effectivement	5067.8	5067.7	5067.7	5067.7	5067.7	5067.7	arrière-fond
1997.7	1997.7	1997.8	1997.8	1997.8	1997.8	malheureusement	5112.3	5112.2	5112.2	5112.2	5112.2	5112.2	transcendantale
2098.2	2098.2	2098.6	2098.6	2098.6	2098.6	maintenant	5162.6	5162.5	5162.5	5162.5	5162.5	5162.5	définissent
2156.8	2156.8	2156.8	2156.8	2156.8	2156.8	conservatoire	5235.0	5238.8	5235.0	5235.0	5235.0	5235.0	uniquement
2198.2	2198.2	2198.1	2198.1	2198.1	2198.1	extraordinaire	5285.7	5291.8	5256.8	5280.1	5280.1	5280.1	Kriesleriana
2264.4	2264.4	2264.4	2264.4	2264.4	2264.4	actuellement	5382.2	5374.7	5382.2	5382.2	5382.2	5382.2	aujourd'hui
2317.0	2317.0	2317.3	2317.0	2317.0	2317.0	malheureusement	5460.2	5460.2	5460.4	5460.4	5460.4	5460.4	probablement
2345.8	2345.6	2345.6	2345.6	2345.6	2345.6	malheureusement	5492.8	5492.8	5492.8	5492.8	5492.8	5492.8	aujourd'hui
2435.5	2435.4	2435.4	2435.4	2435.4	2435.4	anniversaire	5560.8	5560.9	5560.9	5560.9	5560.9	5560.9	téléphoner
2472.4	2472.5	2472.5	2472.5	2472.5	2472.5	anniversaire	5596.1	5596.1	5596.1	5596.1	5596.1	5596.1	enregistrement
2562.2	2562.1	2562.2	2562.2	2562.2	2562.2	farfouillant	5671.5	5671.5	5671.5	5671.5	5671.5	5671.5	viendraient
2631.6	2631.5	2631.4	2631.5	2631.5	2631.5	recommencera	5755.0	5755.0	5755.0	5755.1	5755.1	5755.0	comprendras
2640.1	2640.1	2640.1	2640.1	2640.1	2640.1	restaurant	5767.9	5767.9	5767.9	5767.9	5767.9	5767.9	machiavélique
2709.4	2709.4	2709.5	2709.5	2709.5	2709.5	félicitations	5809.6	5825.2	5844.3	5809.5	5809.5	5809.5	tortueux
2805.7	2805.7	2805.7	2805.7	2805.7	2805.7	confidences	5904.2	5904.2	5904.2	5904.2	5904.2	5904.2	quelquefois
2859.8	2859.8	2859.8	2859.8	2859.8	2859.8	incendiaire	5956.9	5979.6	5956.8	5976.1	5976.1	5956.8	chaussures
2909.7	2909.7	2909.7	2909.7	2909.7	2909.7	actuellement	5980.6	6041.8	5980.5	6035.4	6035.4	5980.5	contrairement
2973.7	2973.7	2973.7	2973.7	2973.7	2973.7	appartement	6020.7	6105.2	6035.4	6093.0	6093.0	6020.5	complètement
3063.2	3035.4	3063.2	3063.2	3063.2	3063.2	rencontrées	6048.1	6178.9	6155.4	6172.5	6172.5	6048.1	sûrement
3074.8	3074.8	3074.8	3074.8	3074.8	3074.8	précisément							

Table 6 : Time-code of each 103 word for each step of the gradual splitting algorithm. -1) Time-code reference -2) Without splitting algorithm -3) Use of noisy phonetic models and noise model -4) Removal of non-voiced segments -5) Removal of noisy segments -6) Removal of music segments