550, rue Sherbrooke ouest, bureau 100
Montréal (Québec), Canada H3A 1B9
Téléphone : (514) 840-1234
Télécopieur : (514) 840-1244
http ://www.crim.ca/

Projet E-Inclusion

Acoustic segmentation

Rapport technique
CRIM-06/05-02

Michel Comeau
Groupe Reconnaissance de la parole

Mai 2006

ISBN-13: 978-2-89522-071-8
ISBN-10: 2-89522-071-9

Table of Contents

List of Tables

1 Introduction

In this work we apply automatic speech recognition technology to the task of audio classification. The audio is segmented into a set of predefined sound classes. Each of these classes is modeled by a hidden Markov model (HMM) that is trained on a limited of amount of manually segmented transcriptions.

Our developped procedure is applied to media recordings chosen from our laboratory database. In the future, it will serve as a basis for the classification and indexing of audio streams associated with the AV (Audio/Visual) media content provisions of the E-Inclusion project.

2 Data

The data arises from the following french language media sources :

TVA : News broadcasts of "Le réseau TVA".
GJ : "Le grand journal" ; broadcast news segment of "Télévision Quatre Saisons"
MM : "Météomédia" ; Weather channel.
CA : "Canal Argent" ; Channel devoted to financial information.

The data represents some 8 hours and 17 minutes of audio that have been manually segmented into the following classes :

Pure speech : Speech with no background noise.
Speech+noise : Speech with underlying background noise
Speech+music : Simultenous speech and music.
Pure Music : Music.
Publicity : Commercials.
Other : This category includes such sounds as noise, coughing, undiscernible sound and silence.

The sound class distribution amongst the media sources is shown in table 1. Commercials may be segmented according to any of the other classes. Because of this, we can anticipate a corresponding model endowed with poor discriminating capabilities. To remedy for this confusion, we have deleted all publicity from the data. [1].This reduces the audio duration to 6 hours 9 minutes[2] and leads to the class distribution shown in table 2. The data has been divided into train and test sets as shown in tables 3 and 4. As can be seen, the scarcity of available data[3] has constrained us to define a single media transcript for testing (a one half-hour news broadcast of TVA).

[1] This implies editing out the corresponding segments from the audio and textual transcripts, and correcting the timestamps in the latter.

[2] Publicity accounts for 2 hours 8 minutes.

[3] In the sense of manually segmented data at hand.

Table 1: *Sound class distribution (%) of media corpa.*

	TVA	GJ	MM	CA
Pure speech	24	42	34	28
Speech+noise	38	20	7	34
Speech+music	5	5	6	4
Pure music	3	2	22	3
Other	4	7	3	4
Publicity	25	24	27	27

Table 2: *Sound class distribution (%) of media corpa stripped of publicity.*

	TVA	GJ	MM	CA
Pure speech	33	55	47	38
Speech+noise	51	27	10	47
Speech+music	7	6	8	5
Pure music	4	2	30	4
Other	6	9	5	6

Moreover, training times of less than 1 hour of data (speech+music for example) may translate into undertrained models. It should be noted however that the models trained over such a limited amount of data represent a viable starting point for an augmented training performed iteratively in a unsupervised way, although that refinement was not pursued at the present time.

3 Procedure

Each acoustic class is represented by a left-right 3 state HMM, with a single emitting state (state 2) associated with an output distribution corresponding to a Gaussian Mixture Model (GMM) with individual mean vector and diagonal covariance parameters. Input parameters are 25-dimensional vectors of 12 static MFCCs, absolute energy being excluded, corresponding first-order derivatives plus the first-order derivative of the energy. Each GMM is separately trained on the set **E** as follows : *i)* initial parameter estimation is provided by iterative Viterbi segmentation using the HTK tool HINit and *ii)* the initialized HMM is re-estimated according to the Baum-Welch procedure using the HTK tool HRest. After initial production of single-gaussian HMMs, the number of gaussian components is doubled, the parameters (means, variances, transition probabilities and component weights) are reestimated and so on until the production of mixture models of 64 gaussians.

The language model is a bigram, trained on set **E** using a closed vocabulary made up of the 5 acoustic classes (table 2). The computed perplexities of **E** and **T** are 2.1 and 2.2, respectively.

Table 3: *Train (*$\mathbf{E}$*) and Test (*$\mathbf{T}$*) durations (h :min :s) of media corpa*

Media	$\mathbf{E}$	$\mathbf{T}$	$\mathbf{E} \cup \mathbf{T}$
TVA	0 :44 :30	0 :22 :45	1 :07 :15
GJ	1 :54 :38		1 :54 :38
MM	1 :27 :24		1 :27 :24
CA	1 :40 :17		1 :40 :17
Total	5 :46 :50	0 :22 :45	6 :09 :35

Table 4: *Train (*$\mathbf{E}$*) and Test (*$\mathbf{T}$*) durations (h :min :s) of sound classes*

Sound Class	$\mathbf{E}$	$\mathbf{T}$	$\mathbf{E} \cup \mathbf{T}$
Pure speech	2 :37 :49	0 :06 :51	2 :44 :40
Speech+noise	1 :48 :05	0 :12 :08	2 :00 :13
Speech+music	0 :23 :21	0 :00 :48	0 :24 :09
Pure music	0 :34 :58	0 :01 :01	0 :35 :59
Other	0 :22 :37	0 :01 :58	0 :24 :34
Total	5 :46 :50	0 :22 :45	6 :09 :35

The word-level network, in *Standard Lattice Format* (SLF), is built from the language model using the HTK tool HBuild.

A single pass of the Viterbi algorithm is applied to the complete data set $\mathbf{E} \cup \mathbf{T}$ to yield the corresponding hypothesized segmented transcripts, using the HTK tool HVite.

4 Evaluation metrics

i) **Recognition Accuracy**

Let N^{ref} represent the number of labels in the reference transcript $\mathcal{R}$ and N^{reco} the number of correctly recognized labels in the hypothesized transcript $\mathcal{H}$. The latter may contain a number of $N^{ref} - N^{reco}$ errors with respect to $\mathcal{R}$ that translate into N_s label substitutions and N_d label deletions ($N^{ref} = N_d + N_s + N^{reco}$). The correctness gives the proportion of correctly hypothesized labels in $\mathcal{H}$:

$$C = \frac{N^{reco}}{N^{ref}} \tag{1}$$

The hypothesized transcript may also be associated with N_i incorrect label insertions.The accuracy is then computed as :

$$A = \frac{N^{reco} - N_i}{N^{ref}} \tag{2}$$

Introducing now

$$D = N_d/N^{ref} \tag{3a}$$

$$I = N_i/N^{ref} \tag{3b}$$

$$S = N_s/N^{ref} \tag{3c}$$

as the proportion of deletions, insertions and substitutions, it is[4] :

$$A = 1.0 - (D + I + S) \tag{4}$$

ii) **Segmentation Accuracy**

Our evaluation metric for segmentation measures the coincidence of segmented labels between the reference and hypothesized time-aligned transcripts. The reference transcript will be segmented into N classes. Let t_{ji}^{ref} represent the duration of segment j associated with acoustic class i in the reference transcript ; $j = 1, M$ where M is the corresponding total number of segments of type i. The duration of class i in the reference transcript is :

$$t_i^{ref} = \sum_j^M t_{ji}^{ref}$$

such that the duration of the reference transcript is expressed as :

$$t^{ref} = \sum_i^N t_i^{ref}$$

where N is the number of classes in the reference transcript. Let τ_{ki} represent the coincident segment $k \subset i$ duration of the time-aligned reference and hypothesized transcripts. The duration of coincident segments of class i is then :

$$\tau_i = \sum_k^L \tau_{ki} \tag{5}$$

where L is the number of coincident segments. The segmentation accuracy of labelled class i is computed as :

$$\mathcal{A}_i = \frac{\tau_i}{t_i^{ref}} \tag{6}$$

The overall segmentation accuracy is then given by :

$$\mathcal{A} = \frac{\sum_i \tau_i}{t^{ref}} \tag{7}$$

[4] In terms of 3, the correctness is given by $C = A + I$.

5 Results

5.1 All classes

We have performed a series of recognition experiments on the complete data set $\mathbf{E} \cup \mathbf{T}$ with respect to the parameters $\mathcal{P} = (\rho,\ \pi,\ \beta)$, namely the language model weight, insertion penalty and beam. [5]. The corresponding optimized parameters are found to be $\mathcal{P} = (25, -15, 2000)$. The hypothesized complete set yielded by each of these experiments was then split into $\mathbf{E}$ and $\mathbf{T}$ and the corresponding performances were measured. The best performance observed for $\mathbf{E}$ is associated with $\mathcal{P} = (30, -15, 2000)$; in the case of $\mathbf{T}$ it is $\mathcal{P} = (25, -10, 2000)$.

Table 5 reports the recognition results of the sets $\mathbf{E} \cup \mathbf{T}$, $\mathbf{E}$ and $\mathbf{T}$ for the best-performance

Table 5: *Recognition and segmentation accuracies (percent) of best-recognition sets.*

	ρ	π	β	D	I	S	A	$\mathcal{A}$
$\mathbf{E} \cup \mathbf{T}$	25	-15	2000	13.1	12.7	9.0	65.3	76.8
$\mathbf{E}$	30	-15	2000	16.6	9.2	8.3	65.9	76.8
$\mathbf{T}$	25	-10	2000	13.6	3.6	13.9	68.8	78.9

parameters of each of these (equation 4). The accuracie of the acoustic training set $\mathbf{E}$ is conspicuously low (below 70%) ; this may be a direct result of insufficient training data. This situation may also be reflected in the best result of $\mathbf{T}$ (A = 68.8%) with corresponding values of $\mathcal{P}$ that yield A = 54.7% for $\mathbf{E}$.

Table 5 also reports segmentation accuracies of the data sets, as mesured by equation 7. Contrary to the previous metric, the values of $\mathcal{A}$ show much less variation with respect to the reported $\mathcal{P}$ parameters.

The segmentation class accuracies of the best-recognition sets (equation 6) are shown in table 6. Segmentation accuracies for pure music are high for the training data ; the corresponding value for set $\mathbf{T}$ is much lower but it is difficult to conclude on its significance in view of the small sampling involved (1 minute ; *cf.* table 4). The same caution applies to the speech+music class.

[5] Common practice is to optimize with respect to the test set only. In our case the latter contains a single of the 4 media types included in the training set and the corresponding optimized $\mathcal{P}$ may be biaised towards that particular media type (**TVA**).

Table 6: *Acoustic class segmentation accuracies (percent) of each best-recognition set.*

	$\mathbf{E \cup T}$	$\mathbf{E}$	$\mathbf{T}$
Pure speech	79.8	79.6	92.3
Speech+noise	72.8	72.8	76.9
Speech+music	79.1	78.8	91.7
Pure music	90.9	92.2	35.6
Other	52.7	50.7	62.5
Total	76.8	76.8	78.9

5.2 Speech and music classes

Our evaluations have been performed using HMM class-specific models for pure speech, speech+noise, speech+music, pure music and Other (Table 4). We now inquire on the performances related to the 2 augmented classes : speech and music. These can be inferred by applying the following groupings on the existing reference and hypothesized transcripts[6].

$$\begin{aligned} \mathbf{Speech} &= \mathbf{Pure\ speech} \ \cup \ \mathbf{speech + noise} \\ \mathbf{Music} &= \mathbf{Speech + music} \ \cup \ \mathbf{pure\ music} \end{aligned} \tag{8}$$

The augmented class durations for the 3 sets considered are indicated in table 7. The existing

Table 7: *Train (*$\mathbf{E}$*) and Test (*$\mathbf{T}$*) durations (h :min :s) of speech and music classes.*

$\mathbf{Sound\ Class}$	$\mathbf{E}$	$\mathbf{T}$	$\mathbf{E \cup T}$
Speech	4 :25 :55	0 :18 :58	4 :44 :53
Music	0 :58 :18	0 :01 :49	1 :00 :08
Total	5 :24 :13	0 :20 :48	5 :45 :01

transcripts associated with the experiments performed in section 5.1 were subjected to the transformation 8. The best-recognition results of each set (associated with parameters $\mathcal{P}$) and correponding segmentation accuracies are indicated in table 8.

[6]Specifically, we replace labels *i)* $\mathbf{Pure\ speech}$ and $\mathbf{speech + noise}$ by $\mathbf{Pure\ speech}$ and *ii)* $\mathbf{Speech + music}$ and $\mathbf{pure\ music}$ by $\mathbf{Music}$. For the purpose of computing recognition accuracies, we delete the remaining label $\mathbf{Other}$. On the other hand, segmentation accuracies rely on the boundary timestamps of the time-aligned transcriptions, so all segments (regardless of their groupings) should be included (*i.e*, label $\mathbf{Other}$ is retained in that case).

Table 8: *Recognition and segmentation accuracies (percent) of modified best-recognition sets.*

	ρ	π	β	D	I	S	A	$\mathcal{A}_{\mathbf{Music}}$	$\mathcal{A}_{\mathbf{Speech}}$	$\mathcal{A}$
$\mathbf{E} \cup \mathbf{T}$	30	-35	2000	12.4	3.2	0.0	84.3	89.4	95.7	94.6
E	25	-35	2000	9.3	4.1	0.0	86.6	89.8	95.8	94.7
T	50	-35	2000	0.0	0.0	0.0	100.0	71.2	93.6	91.6

6 Conclusion

Using standard technics of automatic speech recognition, we have developped a procedure for the segmentation of an audio stream into a predefined set of 5 sound classes. With a limited amount of manually segmented training data at our disposal, we achieve moderate recognition and segmentation accuracies.

The improved results obtained by reduction of the hypothesized transcripts into 2 well defined classes - speech and music - lend support to the feasiblity of our approach. We can anticipate performance enhacements for the multi-class case using acoustic models trained on a greater amount of data, and a more specialized language model than the bigram used.