Detection of Acoustic Patterns in Broadcast News using Neural Networks

H. Meinedo, J. Neto

Spoken Language Systems Lab, INESC-ID Lisboa, R.Alves Redol, 9, 1000-029 Lisboa, Portugal
Hugo.Meinedo@l2f.inesc-id.pt
Instituto Superior Técnico, Lisboa, Portugal

ABSTRACT: This paper describes the use of neural networks to detect jingles that mark the start and end of broadcast news. Contrarily to many other applications, in this particular one, we are not interested in the generalization capabilities of the neural network, as the goal is to detect a single acoustic pattern and not generalize to similar ones.

Accurate jingle detection is crucial to the performance of our automatic broadcast news transcription system, in order to delimit the parts of the audio signal that belong to the news show and are worth feeding to a large vocabulary speech recognition system.

In order to train the neural networks we collected several samples of patterns to be either detected or rejected. That was done for the 2 public TV channels in Portugal (RTP), and for a private one (TVI). In a limited test set, our approach obtained a very high accuracy in frame classification (jingle/non-jingle), allowing the retrieval of 100% of the signal worth transcribing in the news shows. Our jingle detector took 0.028 times real time in a Pentium IV, operating at 2.66 GHz.

1. INTRODUTION

The development of the work described in this paper was initiated in the scope of the ALERT European project [1] (2000-2002). The goal of this project was to develop a media watch system that continuously monitors a TV channel, searching inside its news programs for stories that match the profile of a given user [2]. The system may be tuned to automatically detect the start and end of a particular news program. Once the start is detected, the system automatically records, transcribes, indexes, summarizes and stores the program.

Accurate detection of the programs start and end jingles is crucial to the performance of our automatic broadcast news transcription system, in order to delimit the parts of the audio signal that belong to the news show and are worth feeding to a large vocabulary speech recognition system.

Our approach uses artificial neural networks to detect the jingles that mark the start and end of broadcast news show. Neural networks are widely known pattern classifiers that possess good generalization capabilities when correctly trained. But unlike many other pattern classification applications we are not interested in the generalization capabilities of the neural network, as the goal is to detect a pre-determined acoustic pattern and not generalize to similar ones.

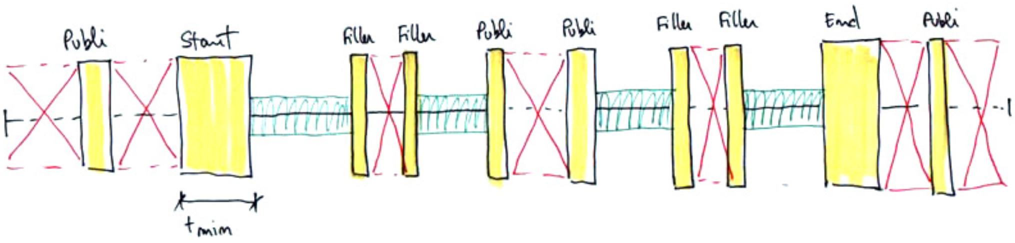
Section 2 briefly describes the different jingles that are used by the Broadcast News programs that we intent to process automatically. The major part of this paper is devoted to the detailed description of the processing blocks (Section 3). In Section 4, we shall briefly describe the training steps for the neural network classifiers and the amount of training material collected which allowed the training and evaluation of this block. The overall evaluation is presented in Section 5. Finally on Section 6 we present some concluding remarks and our most recent research trends.

2. BROADCAST NEWS PROGRAMS

In the Broadcast News shows that we are processing there are four types of jingles: start of show which indicates the beginning of the program, end of show marking the end of the program, the publicity jingle either marking the beginning/ending of a commercial break or appearing between commercials and filler jingles.
The filler jingles appear when the newscaster is emphasizing some news stories that will be expanded later in the show or is summarizing the news stories that were covered during the show. In either situation these filler sequences don't convey relevant information and can induce errors in the story classification module because the topic is changing rapidly.

Currently we are monitoring three different BN shows. The "Telejornal" which is the main 8 o'clock news show from RTP1 (Radio Televisao Portuguesa channel 1) the Portuguese public broadcast news company; The "Jornal 2" which is 9 o'clock news show from RTP2 station and the "Jornal Nacional" which is again the main news show (8 o'clock) but from the private station TVI.
"Telejornal" news show is currently characterized by two jingles, one for delimiting the main body of the broadcast news (start, end and fillers) and another to mark publicity. "Jornal 2" news show has a jingle for the start of the show and a different one for the end. It has no commercial breaks or filler blocks. The private channel news show ("Jornal Nacional") has four different jingles: start, end, publicity and filler. Each of these jingles has minimum time duration. In Fig. 1 we represent the time sequence of the "Jornal Nacional" program and one hypothetic jingle sequence.

Figure 1. *"Jornal Nacional" news show jingle sequence.*

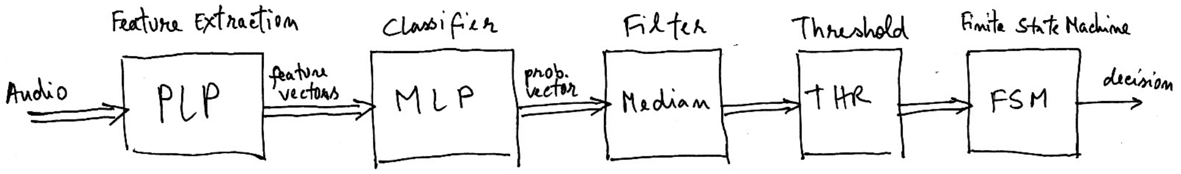


In Fig. 1 the jingles (start, end, filler and publicity) are marked in yellow, the useful content is in green and the non useful parts (that will be discarded) are in red.

3. MODULE ARCHITECTURE

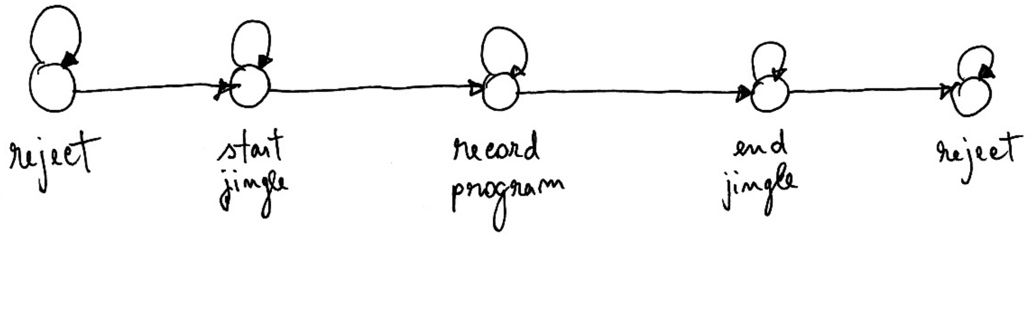
The block diagram of our jingle detection module, represented in Fig. 2, includes 5 main components. The first one extracts PLP (Perceptual Linear Prediction) [3] features from the incoming audio signal. It uses a sliding window of 20 ms, which is updated every 10 ms and extracts 26 parameters per frame (12th order plus energy plus first order derivatives). The second block is a neural network classifier of the type MLP (Multi-Layer Perceptron) that classifies these acoustic feature vectors and is trained to estimate at the output the probability of the given time frame being a certain jingle. The MLP architecture includes 9 input context frames, 25 hidden units and 2 complementary output units. The output of this binary classifier is then smoothed by a median filter with a 21 frame window and compared to a pre-determined threshold value. After some adjustments in the training set, this threshold was set to 90%, that is, it only considers that the jingle occurred if the frame probability is higher than 0.9.

Figure 2. *Jingle Detector Block Diagram.*



The last block is a finite-state machine that defines the possible jingle/non-jingle sequences. This block receives as input the signals from the several jingle classifiers and takes a decision regarding the recording of the audio signal (useful audio). Fig. 3 represents the finite state diagram for one of the news shows that is processed ("Jornal2").

Figure 3. *Jornal2 Finite State Transitions Diagram.*



There is one state for rejecting the audio, one state while the audio belongs to the start jingle, one for recording the useful audio, one for the end jingle and finally one state for rejecting the audio after the news show ends. The finite state machine also takes into consideration the minimum time duration of the jingles, that is, if a jingle occurred but it lasted less than its minimum duration, the finite state machine does not change state. There is a recording buffer for compensating this decision delay. The size of the buffer has to be larger than the longest minimum duration of all the jingles. Presently it is set at one second.

4. MLP TRAINING

The jingle MLP classifiers were trained using the back-propagation and gradient descent algorithm [4] in stochastic mode with non-adaptive training step. The stop criterion was the mean square error in the cross validation set. We also used output error weighting factors in order to balance class *a priori* distributions.

Table 1 – *Amount of training material for each jingle.*

Jingle	**training**	**validation**	**total time frames**
Jornal2 start	20	5	16047
Jornal2 end	5	2	3694
Telejornal main	18	5	23726
RTP1 publi	8	2	2983
Jornal Nacional start	7	2	6599
Jornal Nacional end	8	2	12313
Jornal Nacional filler	26	8	8972
TVI publi	10	3	3617
Reject patterns	26	3	79726

As we can see from Table 1, an adequate total number of training frames was chosen in order to have a high pattern to weight ratio. The cross validation set was chosen in order to have about 20% of the total number of training patterns. The frame classification rate obtained after training was well above 95% for the cross validation sets and approaches 100% for the training sets.

5. TEST RESULTS

The evaluation of the different jingle detectors was done with three test sets, one for each different type of news show that we are processing. Each of these test sets is composed by one news show which typically lasts almost an hour. We considered as evaluation criterions the frame classification rate (FCR) and the percentage of useful content retrieved (% of CR). Table 2 shows our results including the time taken to process the news shows, expressed in terms of real time percentage (% RT).

Table 2 – *Global results.*

News show	FCR	% of CR	% RT
Telejornal	79.3	94.8	2.0
Jornal 2	89.2	100.0	2.0
JornalNacional	84.6	100.0	2.8

These results were obtained in a Pentium IV computer operating at 2.66 GHz. Although the frame classification rate is less than 100%, that is, the algorithm is not capable of correctly tagging all the time frames that belong to the jingles, the percentage of content retrieve approaches 100%. In the "Telejornal" news show, the algorithm failed to classify one of the filler segments, which was marked as containing useful audio. Sometimes these filler jingles are shorter than the minimum duration or played with volume fade-in which makes their correct detection very difficult.

6. CONCLUSIONS AND FUTURE WORK

The jingle detection module plays a very important part in our daily automatic transcription system for Broadcast News shows. It permits the correct identification of relevant audio by detecting program start/end jingles. Furthermore it detects non useful commercial breaks and filler segments. The evaluation results are encouraging although not perfect especially with the correct detection of some filler segments. The algorithm is very fast, taking only a fraction of real time and having an acceptable maximum delay of one second.

Furthermore, this technique is being used successfully by us in other applications like voicemail jingle identification with similar good results.

From time to time the broadcast companies change the jingles. Future work could be done regarding the automatic detection of these newly observed jingle patterns, in order to prevent system failure. This would also decrease the time consuming task of manually collecting new samples in order to retrain the neural network classifiers.

ACKNOWLEDGEMENT

This work was partially funded by FCT project POSI/PLP/47175/2002. Hugo Meinedo was also supported by a FCT scholarship SFR/BD/6125/ 2001. INESC ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

REFERENCES

[1] ALERT project web page http://alert.uni-duisburg.de/

[2] J. Neto, H. Meinedo, R. Amaral and I. Trancoso, *A system for Selective Dissemination of Multimedia Information Resulting from the ALERT project*, In Proc. ISCA ITRW on Multilingual Spoken Document Retrieval, Hong Kong, China, April 2003.

[3] H. Hermansky, N. Morgan, A. Baya and P. Kohn, *RASTA-PLP Speech Analysis Technique*, In Proc. ICASSP 92, San Francisco USA, April 1992.

[4] L. Almeida, *Multilayer Perceptrons*, Handbook of Neural Computation, Editors Fiesler E. and Beale R., IOP Publishing Ltd and Oxford University Press, 1996.