DSP-MCOM 2005
Košice, Slovak republic
13-14 September 2005

Speech Detection in the Broadcast News Processing

Matúš Pleva[1], Jozef JUHÁR[2], Anton ČIŽMÁR[3]
Department of Electronics and Multimedia Communications,
Technical University of Košice, Park Komenského 13, 041 20 Košice,
Slovak Republic.
[1] Matus.Pleva@tuke.sk
[2] Jozef.Juhar@tuke.sk
[3] Anton.Cizmar@tuke.sk

Abstract **- This paper deals with speech detection problem which is a part of broadcast news stream preprocessing before continuous speech recognition. The whole broadcast is recorded and then offline processed with blocks of speaker turn detection, speech detection, gender detection, speaker clustering and finally with speech recognizer. The results are so called metadata files, which contains the information about the broadcast news show recorded. These files could be used for information searching in broadcast news shows archive.**

The paper also describes the algorithm used in the experiments and compares the obtained results with the results of other institutions cooperated in *COST-278 broadcast news special interest group*[1]. The whole procedure consists of Perl scripts running on BN COST-278 database.

Index terms **– broadcast news, speech detection, segmentation**

1 Introduction

In the first section of the paper the broadcast news COST 278 database [1] will be briefly introduced. The structure, collaborating institutions and the purpose of this database will be mentioned, too.

Then the process of data preparation will be briefly described, especially the conversions between lots of file and data formats which are commonly used for special purposes and tasks.

In the next section the algorithm of the speech detector training [2] will be described, and then the algorithms of others institutions cooperating in COST 278 broadcast news special interest group [3] will be briefly mention.

The last section, the testing algorithm and the results, will be depicted and compared with all the cooperating institutions.

2 Preparation of the data

The BN COST 278 database was collected by 10 institutions and each of them recorded a national data set consisting of approximately 3 hours of material. The material was manually transcribed to text form using the Transcriber[2] tool. Also the focus conditions, speaker turns, silence inside a speaker turn, jingle segments and foreign language utterances was described in the transcription files (.trs).

The complete database contains 30 hours of data originating from TV stations (public or commercial), and covering 9 European languages, namely Dutch, Portuguese, Galician, Czech, Slovenian, Slovak, Croatian, Hungarian and Greek [3].

Due to the limited size of the national data sets we can not use them for recognition system training, but they are very suitable for the evaluation of acoustic model adaptation methods and audio indexing systems**.**

2.1 Format conversion

Firstly, there was a need to extract the transcribed data from XML format transcriber (.trs) files and then convert them to HTK[3] master label files (.mlf). Namely to extract the speech / non-speech information about the segments and then to produce the list of training and testing segments (trainlist.lst, testlist.lst).

During the tests we tried to use more then two models (not only speech / non-speech), but we obtained no significantly better results, because not very often used models (for example speech+music) do not contain so much training materials.

2.2 Feature extraction

The feature extraction methods used in HTK tutorial [4] was implemented. So that the HCopy tool to extract 12 mel-cepstral coefficients and energy coefficient was used, and during the training period the HRest tool was used to extract next 12 delta and acceleration coefficients.

2.3 HMM Model configuration

To compare the results more than one HMM model types and structures were used. For example there were used from 3 to 7 states HMM models, ergodic and left-to-right topology, but the results were not significantly

[1] http://chardonnay.elis.ugent.be/cost278/cost278.html

[2] http://www.etca.fr/CTA/gip/Projets/Transcriber/

[3] Hidden Markov Model Toolkit - Speech Recognition toolkit - http://htk.eng.cam.ac.uk/

better. Finally a 7 state ergodic HMM models were used and during the training period the number of PDF[4] on state were duplicated to 64 PDF on one state.

3 Speech detection training

For speech detection training the HTK Tools was used like it was mentioned before. Especially the HInit, HRest and HERest tools was usually used.

3.1 Initializing the models

The two models (speech and non-speech) was initialized using HInit tool and all the training material which consists of one full national set (all material from one language). Then the models are used for testing on other full national sets. This type of tests is used to call as C5 tests (training on one national set – testing on all the others national sets). For comparing the results with other institutions the training procedure was done only on BE, GA, PT and SK national sets.

3.2 Reestimation

The models were then reestimated by the HRest HTK Tool and then the number of PDF on state were doubled by HHEd HTK Tool. Then the new HMM models were reestimated two times with HERest Tool which used the Baum-Welch reestimation algorithm.

This training procedure was inspired by examples from HTK Book tutorial [4], too.

3 Speech detection testing

For speech detection testing the HTK Tool HVite was used realizing the Viterbi algorithm. The whole testing material was transformed with HCopy to mel-frequency coefficients feature files. Then the Viterbi algorithm produced the results and the results were compared with manually transcribed metafiles.

The results from all recordings were averaged for every national set tested. The results were then converted to tabular, which are very useful form for comparing and graphical representation generating.

Finally, it is possible to compare the results with other participating institutions and their algorithms.

3 Comparing the results

The TUK results (Technical University Kosice – TUK) were compared with institutions which made exactly the same tests but using another algorithm for speech detection on University of Ljubljana (ULJ) and Technical University of Liberec (TUL), which are cooperated the COST 278 Broadcast News special interest group, too. These results depicted on Fig. 1 will also be published on InterSpeech Conference this year [3].

[4] Power Density Function

The algorithms results were compared on the basis of their accuracy or the opposite of error rate (which means 100% - accuracy) [3].

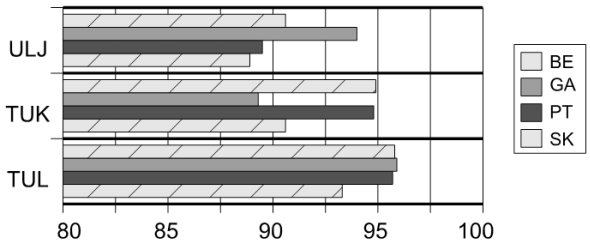

Fig.1 Average results of testing the speech / non-speech detection on all other national sets with BE, GA, PT and SK national set trained models.

4 Conclusions

The speech detection is the first step in the chain of speech preprocessing before speech recognition in the topic of broadcast news processing. The goal is to make the whole broadcast news automatic archive engine with speaker turn detection, gender and speaker detection, and continuous speech recognition cooperating with N-gram modeling.

This research was performed with support of the grant project N° 1/1057/04 of the Grant Agency for Science (VEGA), Slovak Republic and COST 278 (http://cost278.org/) Spoken Language Interaction in Telecommunication.

References

[1] A. Vandecatseye, J. P. Martens, J. Neto, H. Meinedo, C. G. Mateo, J. Dieguez, F. Mihelic, J. Zibert, J. Nouza, P, David, M. Pleva, A. Čižmár, H. Papageorgiou, Ch. Alexandris: "*The COST278 pan-European Broadcast News Database",* Proc. LREC 2004 – 4th International Conference On Language Resources And Evaluation, 873–876, 2004

[2] A. Zgank, Z. Kacic, A. Moreno, M. Caballero, F. Diehl, K. Vicsi, G. Szaszak, J. Juhár, S. Lihan: "*The COST 278 initiative – crosslingual speech recognition with large telephone database"*, Proc. LREC' 2004 – 4th International Conference On Language Resources And Evaluation, 2107–2110, 2004

[3] J. Zibert, F. Mihelic, J. P. Martens, J. Neto, H. Meinedo, C. G. Mateo, L. Docio, J. Zdansky, P, David, M. Pleva, A. Čižmár, A. Zgank , Z. Kacic, C. Teleki, K. Vicsi: "*The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation - Overview, Methodology, Systems, Results*", Proc. Interspeech 2005 (Eurospeech) – 9th European Conference on Speech Communication and Technology (to be published)

[4] S. Young, J. Odell, D. Ollason, V. Valtchev and P. Woodland, "*The THK Book"*, Cambridge University 1995