TRANSCRIPTION OF BROADCAST NEWS - SYSTEM ROBUSTNESS ISSUES AND ADAPTATION TECHNIQUES

R. Bakis, S. Chen, P. S. Gopalakrishnan, R. A. Gopinath, S. Maes, and L. Polymenakos

Human Language Technologies, Computer Science Dept.,
IBM T. J. Watson Research Center, Yorktown Heights, NY.,
email:rameshg@watson.ibm.com, phone: (914)-945-2794

ABSTRACT

This paper describes some of the main problems and issues specific to the transcription of broadcast news and describes some of the methods for solving them that have been incorporated into the IBM Large Vocabulary Continuous Speech Recognition System

1. INTRODUCTION

Significant advances in speech recognition technology have been achieved recently, as seen on tests conducted with read speech corpora such as the Wall Street Journal corpus [1]. The focus of research has shifted recently to transcription of "found" speech like radio/TV broadcast news. Transcription of broadcast news presents technical challenges arising from several sources of signal variability. A typical broadcast news segment contains speech and non-speech data from several sources, such as the signature tune of the show, interviews with people on location - possibly under very noisy conditions - and interviews over the telephone, commercials, etc. Broadly speaking, the data in such broadcasts can be characterized using three criteria: the quality of the microphone or channel, the characteristics of the speaker, and the condition of the background. The signal might be acquired using a high quality microphone, a low bandwidth microphone, or could be telephone quality. The speaker may be an experienced announcer or correspondent or an inexperienced speaker.The speech from the former appears similar to read speech, whereas the latter produces largely spontaneous speech. The background may contain music, noise, or other interfering speech. In some cases, there is no speech present - the signal might consist of a musical interlude or an extended period of noise such as street noises added to evoke an environment.

Decoding this data with a system trained on a clean training corpus such as the Wall Street Journal gives very high error rates [5]. It is necessary to develop new techniques to deal with such data. Preliminary ideas along these lines were explored in the IBM system used in the ARPA sponsored, November 1995 Hub4 radio broadcast news transcription task. Error rates dropped from 55% to 27% on some test data [5, 6, 7]. This paper describes continuing work on the various problems encountered and the solutions attempted for transcription of broadcast news.

The basic philosophy is to first try and identify the segments of input data that belong to one of several classes and use separate modeling techniques appropriate for each class. For instance, segments detected as pure music are discarded and not decoded, segments identified as telephone quality speech are decoded by a system trained on telephone bandwidth speech, and so on. In the following sections, we describe techniques to handle issues in each class.

A brief description of our base recognition system follows (see [2, 4, 3] for details). The system uses acoustic models for sub-phonetic units with context-dependent tying. The instances of context dependent sub-phone classes are identified by growing a decision tree from the available training data [2] and specifying the terminal nodes of the tree as the relevant instances of these classes. The acoustic feature vectors that characterize the training data at the leaves are modeled by a mixture of Gaussian pdf's, with diagonal covariance matrices. The HMM used to model each leaf is a simple 1-state model, with a self-loop and a forward transition.

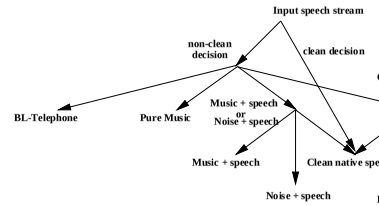
The training data used for the models in this paper comes from three sources: WSJ-SI284 [5], MP-10 [5], and BN-87 (the official 1996 Hub4 evaluation training data distribution consisting of 30 hours of broadcast shows from radio and TV). The test data is from one of the following sources: Dev95H4, Eval95H4, Dev96H4 and Eval94H1. For example, Dev94H4 stands for development test data distributed in 1994 for the Hub4 task.

Section 2. describes the segmentation and classification scheme, Section ?? the models for the various conditions, and Section ?? the adaptation used in our experiments.

For the Dev95H4 test data conditions were separated one at a time, i.e. first pure music segments, then telephone segments, and then music-corrupted segments were identified and separated. Finally one is left with clean speech. This organization enables use of different feature spaces for each binary classification problem. For instance, the feature space used to model the pure music segments was the 60-dimensional feature space that was also used for decoding. The feature space used to model telephone-speech was 24-dimensional cepstra augmented with their first and second differences to make up a 72-dimensional feature vector. Table 2. shows the performance of this segmentation algorithm on Dev95H4

On Dev96H4, a different strategy was used. Only linear *HMM* models were used. The separation between classes is also no more a binary condition. All the gaussians are trained on the corresponding con-

tagged as clean. Using *HUB4'96* cle
often classification of clean speech
even speech plus music. Figure 2.
mentation algorithm.



It has been observed on Dev
Dev96H4 that this strategy sometir
segments as pure music or music pl
fore, after segmentation, music, mus
telephone segments are further clas
NBL using the system previously
sic.*BL*, telephone.*BL* and music+sp
are tagged as telephone.*BL*, music.
considered as pure music and spe
considered as speech+music. On
ment data, this strategy takes care
misclassifications of long telephone s

3. CONDITION-SPECIFI

3.1. Clean Speech

For clean speech better models using
normalization (VTL) [11] and speake
(SAT) [12] are explored. These me
reduction in relative error rate of up t
by preliminary experiments.

3.2. Detection Of Music

There is no universal definition of r
from rhythmic speaking to long sus
Techniques for detecting music mu
devised to the kind of "music" we a

In present-day broadcasting in
world, background music is predom
electronic instruments producing no
tempered 12-tone scale, tuned to a
cisely phase-locked to quartz crysta
music is easily detected by the prese
tones. An algorithm based on this c
extremely sensitive to even low lev
music in speech.

training data is used to train music-corrupted models that are then MAP adapted using music-corrupted broadcast news training data. During testing a music-cancellation scheme is applied on the test data which is then decoded with music-corrupted models. On the 1995 Hub4 development test data error rates dropped from the 56% baseline to about 27% using this algoritm.

3.4. Noise Corrupted Speech

For noise corrupted speech, and speech on degraded microphones PLP-based feature space gives a sufficient degree of robustness [13].

3.5. Telephone Bandwidth Speech

The main problem is that the speech has low-bandwidth. Two approaches are attempted. Firstly, we use data from the Switchboard telephone corpus (which contains restricted domain conversational speech). Secondly, we bandlimit the WSJ SI-284 data to 200-3500 Hz and use this data. On The Telephone Portion Of Hub4 Development Test Data The Error Rates Dropped From 55% (With Switchboard Training) To 40%. Based On Preliminary Experiments A Plp-Based Feature Space Seems To Be More Robust Than Mel-Cepstral Feature Speace.

3.6. Rapid Evaluation Of Acoustic-Feature Information Capacity

Traditionally, a new set of acoustic features, such as cepstra from a filterbank with a different set of center frequencies, must be tested by training a complete system on the new features and then running it on test data—a very time-consuming project.

There exists, however, a simple algorithm for estimating the mutual information between a set of acoustic features and any given set of phonetic labels. This mutual information sets an upper bound on the performance of those acoustic features, but takes much less computation than the complete training and testing of a recognition system. Such rapid acoustic feature testing is a prerequisite for the development of powerful new parametric features.

3.7. Parametric Acoustic Features

Modern recognition systems have very large numbers of adjustable parameters, typically of the order of 10^5 or more. Very few of these parameters, however, are in the signal-processing component of the system. Current signal processors are largely algorithmic, such as Fourier transforms and cepstra. They do have a few parameters—filter center frequencies, for example, but are essentially not specialized for the particular task of speech recognition. The reason for this lack of specialization is, undoubtedly, the above alluded-to difficulty in determining the performance of any given set of acoustic features. With the availability of the mutual-information estimator, however, it becomes feasible to adapt much larger numbers of parameters in the signal processor, and thus generate truly speech-specific signal processors.

An example of such a speech-specific processor is a new formant-tracker. In contrast to traditional algorithmic trackers based, e.g. on a linear predictor model, the new tracker at first glance appears heuristic, with its many "arbitrary" parameters such as cepstral liftering bandwidths, etc. These parameters, however, are not adjusted heuristically, but by strict numerical optimization of the mutual information objective, leading to a set of parameters accurately adapted for its specific task.

4. CONCLUSIONS

Transcription of radio broadcasts poses several challenges. Many of these are problems whose solution will significantly advance the state-of-the-art in speech recognition. Recognition systems have to be developed that can cope with a variety of signal environments, speaking styles and accents, and multiple background noise sources. We have made an initial attempt at developing a system for transcription of broadcast news shows. The results obtained in the initial test are encouraging. Clearly much more work needs to be done in order to obtain an acceptable level of accuracy.

REFERENCES

[1] Proceedings of ARPA Speech and Natural Language Workshop, 1995, Morgan Kaufman Publishers.

[2] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, M. A. Picheny, "Robust Methods for using Context-Dependent features and models in a continuous speech recognizer", Proc. Intl. Conf. Acoust., Speech and Sig. Proc., 1994.

[3] P. S. Gopalakrishnan, L. R. Bahl, R. Mercer, "A tree search strategy for large vocabulary continuous speech recognition", Proceedings of the ICASSP, pp , 1995.

[4] L. R. Bahl et al., "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA Wall Street Journal task", Proceedings of the ICASSP, pp 41-44, 1995.

[5] P. S. Gopalakrishnan, R. Gopinath, S. Maes, M. Padmanabhan, L. Polymenakos, H. Printz, M. Franz, "Transcription of Radio Broadcast News with the IBM Large Vocabulary Speech Recognition System," Proc. of ARPA SLT Workshop, Feb 1996.

[6] P. S. Gopalakrishnan, R. Gopinath, S. Maes, M. Padmanabhan, L. Polymenakos, "Acoustic Models Used in the IBM System for the ARPA Hub 4 Task," Proc. of ARPA SLT Workshop, Feb 1996.

[7] L. Polymenakos, M. Padmanabhan, D. Nahamoo, P.S. Gopalakrishnan, "Suppressing background music from music corrupted data of the ARPA Hub 4 task," Proc. of ARPA SLT Workshop, Feb 1996.

[8] C. J. Legetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous density HMM's", Computer Speech and Language, vol. 9, no. 2, pp 171-186.

[9] J. L. Gauvain and C. H. Lee, "Maximum-a-Posteriori estimation for multivariate Gaussian observations of Markov chains", IEEE Trans. Speech and Audio Processing, vol. 2, no. 2, pp 291-298, Apr 1994.

[10] G. Zavagliogkos, R. Schwartz and J. Makhoul, "Batch, Incremental and Instantaneous Adaptation techniques for Speech Recognition", Proceedings of the ICASSP, pp 676-679, 1995.

[11] S. Wegmann, D. McAllister, J. Orloff and B. Peskin, "Speaker Normalization on Conversational Speech", Proc. of ICASSP 96, pp. 339-343, 1996.

[12] G. Zavagliakos, J. McDonough, "Speaker Adapted Training", presentation at LVCSR Workshop, Baltimore, 1996.

[13] P. C. Woodland, M. J. F. Gales and D. Pye, "Improving Environmental Robustness in Large Vocabulary Speech Recognitionh", Proc. of ICASSP 96, pp. 65-69, 1996.

Table I

Class	Corr	Miss%	Err%
Music	163.53	9.2	5.3
Telephone	766.62	0.13	4.2
Music & speech	308.66	2.8	39.6
Correct speaker	1185.96	17.3	13.6