AUTOMATIC TRANSCRIPTION OF TETRA-TRANSCODED BROADCAST NEWS

Linares Georges (1), Nocera Pascal (1), Ravera Bertrand(2), J.F Bonastre (1)
{georges.linares,pascal.nocera, jean-francois.bonastre}@univ-avignon.fr
Bertrand.RAVERA@fr.thalesgroup.com
(1) Laboratoire Informatique d'Avignon, Université d'Avignon
(2) Thalès Commnucations, Laboratoire Multimédia Processing

ABSTRACT

The use of speech based distant services from mobile devices requires sufficient network perfomance and disponibility. Low rate speech coding reduces significantly these ressource requirements, but it could have also a negative impact on speech quality. In this paper, we study the effect of Tetra [1] transcoding on a speech recognition system.

Experiments are conducted on an continuous speech large vocabulary task, using the LIA automatic speech recognition system (SPEERAL). We use the French broadcast news corpus provided for the Ester evaluation campaign [2].

We first perform recognition using our baseline wide band system; results are then compared to one obtained using acoustic models trained on transcoded datas.

Our results show that processing of transcoded speech requires the adaptation of ASR system in order to reach a good level of performance.

1. INTRODUCTION

High quality speech coding is crutial in the field of mobile communication networks, specially for low-cost mobile services, telephone voice services, speech compression for transmission and storage, etc.

Over the past three decades, the technology of voice coding has been intensively developped, improving speech quality and reducing the coding and transmission costs [3].

Tetra (Terestrian Truncked Radio) [1] is an ETSI standard adopted by many organisations for the development of their network infrastructure, mainly for voice communications. This algorithm provides an high quality low bit rate speech coding. Nevertheless, speech quality is subjective and an automatic speech processing system could be disrupted by signal distortions caused by the transcoding method.

In this paper, we study the impact of the Tetra transcoding algorithm to an automatic transcription system. Experiments are conducted using the LIA broadcast news system wich as been involved into the Ester French evaluation campaign [2]. The performance of recognizer on transcoded data are evaluated and compared to the original system (processing non-transcoded datas). A scheme for adaptating the speech recognition system to transcoded data is also proposed.

The first part of this paper describes the LIA broadcast news system and the Ester evaluation framework.

The second part presents Tetra coding algorithms and the method used for building a Tetra-like speech database from original Ester corpus.

The third part describes our experimental work. Results obtained using several acoutic model training schemes are compared and discussed.

At last, conclusion and perspectives are proposed in the fourth part.

2. THE LIA BROADCAST NEWS SYSTEM

2.1 The LIA speech processing toolkit

The broadcast news system used in this work is based on a toolkit developed at the LIA. This toolkit provides software for transcription system design and implementation. It is composed of three main packages addressing a large part of speech-to-text related tasks. The first one contains software components for segmentation and speaker recognition. This package is based on the free Alize toolkit [4] for speaker recognition and segmentation. The second one is composed of tools for HMM based acoustic modeling and decoding. At last, the core of transcription toolkit consists of an A* based speech recognition engine [5]. Moreover, a few additional tools are provided (automatic text-to-phone phone system, lattice analysis, etc.). All these tools are freely distributed for research activities [4].

2.2 Ester corpus

Ester is an evaluation campaign for transcription systems organised jointly by the Francophone Speech Communication Association (AFCP) the French ministry of Defense (DGA) and the Evaluation and Langage Ressources Distribution Agency (ELDA). The core of the campaign is constitued by transcription and segmentation tasks. Database are composed of radiophonic broadcast news shows coming from different sources.

The train corpus contains about 90 hours of manually anotated speech. For system tuning, organizators provide

a development corpus composed by 8 hours of BN shows (also manually transcribed)..

The train corpus covers the 1998 to 2003 period. Morover, the test corpus contains 10 hours of signal, recorded during the end of 2004.

2.4 General architecture of the LIA Broadcast news system

In addition to intrinsic difficulty of speech recognition, broadcast news transcription adds specific problems related to the signal stream continuity as well as the diversity of the acoustic conditions. Recognizers require manageable speech segments and high level acoustic information about the nature of segments (speaker identity, recording conditions, etc.). The LIA Broadcast news system is composed of two main parts addressing respectively segmentation and recognition tasks.

In our system, 3 successive segmentation passes are performed. Speech segments are initially isolated from audio flow; then, a narrow/wide band segmentation identifies telephone segments. The final pass achieves a speaker-based segmentation under a maximal segment size constraint. This last point is motivated by some technical reasons; nevertheless, such an over-segmentation could introduces some linguistic model breaking or intra-word cuts. So, contiguous segments are overlapped. Concurrent transcriptions of overlapped segments are merged after last decoding stage.

The second stage consists in transcribing homogeneous segments extracted by the segmentation stage. Two recognition passes are performed. The first one generates an intermediate transcriptio which is used. for an unsupervised adaptation of acoustic models, using a classical Maximum Likelihood Linear Regression method (MLLR). Transformations are trained using the speaker segmentation results. A single transformation is trained for each speaker. This blind adaptation process allows the estimation of both speaker and acoustic-condition dependant models.

The second pass performs the final decoding based on these adapted models.

2.3 Segmentation

The goal of the audio partitioner is to divide each broadcast show into homogeneous segments according to predefined classes. This segmentation aims at discarding non speech signal (like music, silence, …), and labelling wide and narrow band signal.

The system relies on a hierarchical segmentation performed in three successive steps :

- during the first step, a speech / non speech segmentation is performed using "MixS" and "NS" models. The first model represents all the speech conditions while the second represents the non speech conditions. Basically, the segmentation process relies on a frame-by-frame best model search. A set of morphological rules are then applied to aggregate frames and to label the segments.
- during the second step, a segmentation based on 2 classes – wide-band speech ("S" model) and telephone speech ("T" model) is performed only on the speech segments detected by the previous segmentation step. All models involved during this step are gender-independent. The segmentation process is a Viterbi decoding ausing an ergodic HMM, composed, here, of two states (“S” and “T” models).
- the last segmentation stage performs a speaker segmentation. It is based on an incremental identification mechanism of new speakers [7].

2.3 Transcription

2.3.1. Acoustic modeling

We use a classical PLP parameterization ; feature vectors are composed of 12 coefficients, plus energy, and first and second derivatives of these 13 coefficients. At last, we perform a cepstral normalization (mean removal and variance reduction) in a 500ms sliding window.

The context-dependent models are trained on the 90 hours of Ester transcribed data. State tying is performed by a decision tree algorithm, using acoustic context related questions. Each acoustic models set contains about 230k gaussians for 3600 emitting states.

Ester corpus provides a small amount of narrow-bandwidth data; so, narrow-bandwidth models were first trained using filtered large bandwidth data (using a low-pass filter); finally, telephone models were mapped to real narrow-bandwidth data extracted from the Ester train corpus.

2.3.2. Lexion and language models

The linguistic resources are extracted from two corpora:

- newspaper “Le Monde” from 1987 to 2003 (330 Million words) ,
- ESTER (960K words).

The lexicon contains 65K words. It is composed of all the Ester corpus words, as well as the most frequent words of the corpus “le Monde” 1987-2003. The phone labeling comes from 2 sources : the phonetic lexicon ILPHO ([]) and the LIA_PHON text-to-phone system ([]) for the unknown words of ILPHO. The forms generated by LIA_PHON were partially checked and corrected manually (especially for proper names).

The language model was learned on the corpus “le Monde” and Ester training set, with the SRI-LM toolkit

[]. It is obtained by a linear combination of three models; the first one was learned on the data of "le Monde" 1987-2002, the second on "le Monde" 2002-2003, and the last on the Ester corpus. All these models are trigram models with the modified Kneser-Ney backoff method (open lexicon).

Lastly, these models are mixed into an unique model with interpolation coefficients determined by the Ester development corpus entropy (relative weights of 0.41, 0.24 and 0.35 for respectively "le Monde" 1987-2002, "le Monde" 2002-2003, ESTER). At last, language models include 16.7 Million of bigrams and about 20M of trigrams.

2.3.3. Search algorithm

The LVCSR engine is an asynchronous stack decoder derived from the A* search algorithm. The choice of this decoding strategy was motivated by scalability offered by this kind of search algorithm. Graph exploration is driven by an estimate function composed of two terms : (1) current hypothesis scoring and (2) a probe function estimating the minimal cost of ending path. Plugging additional (or alternative) information sources to the search may be easily achieved by adding specific cost terms to the estimate function.

As most of state-of-the-art recognition machines, our system uses HMM for acoustic modeling and n-gram linguistic models. Cross-word acoustic contexts are taken into account, and the search algorithm is able to dynamically combine a set of language models.

A special effort was made on search efficiency, both by pruning the search graph size and by reducing the computational cost of explored path scoring. The probe function plays a significant part for search performance. We use a combination of a pruned Viterbi-back acoustic decoding [5] and a fast language model look-ahead [7].

At last, likelihood computation is performed using a gaussian selection method, reducing significantly the computational cost due to acoustic scoring.

This engine allows real time transcription, while the baseline system runs at 10x the real time (less than 5xRT per pass).

3. TETRA TRANSCODING

The applied transcoding aims at modifying the quality of the audio records to obtain a telephony network related quality. To do that, we use two different stages. The former performs a telephony-band weighting by applying a band pass filtering which models an analog telephone based network. This model defines an average send and receive frequency response. An example of such a spectral weighting is given by the standard G712 [2] where a non linear phase IIR filter is used. The G.712 based telephony-band weighting is used as a first stage in Tetra transcoding. The latter is the Tetra speech coding ETSI standard [7] which is typically dedicated to radiocommunications system for professional users. The standardised speech coder is based upon an ACELP (Algebraic Code Excited Linear Prediction) coder at 4567 b/s.

4. EXPERIMENTS

We first perform a baseline experiment using our standard system for decoding the original signals from Ester development corpus. Theses signals are mainly large band ones, less than 10% of the database is phone speech. We obtain a WER of about 23.1% during the first pass, and about 21.2% after the blind acoustic adaptation.

In order to evaluate the impact of Tetra transcoding on recognizer performance, we test this baseline system (using narrow-bandwidth models) for decoding Tetra transcoded data. The test corpus is composed from the 8 hours of the Ester developement corpus transcoded as described in the previous section. Results show that the WER increases to 46.2%. It seems clear that the recognizer robustness is not sufficient regarding to distortions caused by the voice coding system. MLLR adaptation allows a significant improvement of system performance (-4.5% of absolute WER). This gain outperforms significantly gains obtained by applying MLLR on the baseline system (-1.9%).

The next step consists in training models on the full transcoded database. The model structure remains unchanged (topology, complexity, tying, etc...). We obtain a 36.5% WER during the first pass, and 34.7% after unsupervised adaptation. The gain involved by specific models is closed to the gain obtained during the baseline test. These experiments shows two major points :

- specific training is required for recognition efficiency. The relative WER improvement is about 40% since the baseline system obtains weak performance (46% absolute),
- in spite of good intelligibility of transcoded signals, it seems that usefull information has been removed by the coding algorithm. Classical modeling based on Markov models and EM algorithm are probably not able to achieved a full compensation of this information lost.

Table 1 : *Word Error Rates (WER) obtained by respectivly baseline system on original baseline database, baseline system on Tetra transcoded database, Tetra system on transcoded database.*

	Pass 1	Pass 2
Baseline system Original database	23,40%	21,90%

	Pass 1	Pass 2
Baseline wide band system Tetra transcoded database	46,50%	40,20%
Tetra system Tetra transcoded database	36,50%	34,70%

5. CONCLUSION AND FUTURE WORK

Our experiments show that classical speech recognition systems fail to process Tetra transcoded data without significant a significant loss of performance in term of word error rate. Using the narrow-bandwidth models trained on the original Ester database, the word error rate increases to more than 20% relative. The unsupervised adaptation allows a significant improvement (from 46.7% to 40.2%, about 16% relative) but the final word error rate remains at a relativly low level. Nevertheless, the interest of transcription is preserved and good performance could be expected on more constrained tasks (small vocabulary, closed linguistic field, indexing task, etc.).

Combining the model re-estimation wich have been achieved in these experiments and low level signal processing methods could improve the recognition system performance by restoring speech signal quality and increasing signal-to-model matching. Such methods have been evaluated for speech enhancement or for signal normalisation. We plan now to work in this way.

6. REFERENCES

[1] ITU-T, “Recommendations G.712, performance characteristics of PCM channels”, ITU, Geneva, 1992.

[2] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.GF. Bonastre, G. Gravier, « The Ester Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News », Interspeech 2005, Lisboa , Sept. 2005 (to be published)

[3] W.T.K Wong, R.M. Mack, B.M.G Cheetham and X.Q. Sun « Low rate speech coding for telecommunications », BT Technology Journal, Vol 14 No 1 January 1996

[4] http://www.lia.univ-avignon.fr

[5] P. Nocera, G. Linares, D. Massonié, L. Lefort, « Phoneme lattice based A* search algorithm for speech recognition », 2002 TSD2002, Brno

[6] Meignier, S. and Moraru, D. and Fredouille, C. and Besacier, L. and Bonastre, J.-F, “Benefits of prior acoustic segmentation for automatic speaker segmentation”, ICASSP-04, Mai 2004,Montreal, Canada

[7] ETSI, “EN 300 395-4 V1.3.0: Trans European Trunked Radio (TETRA) voice coding”, 1999.