Automatic Classification and Transcription of Telephone Speech in Radio Broadcast Data

Alberto Abad, Hugo Meinedo, and João Neto

L^2F - Spoken Language Systems Lab
INESC-ID / IST, Lisboa, Portugal
{Alberto.Abad,Hugo.Meinedo,Joao.Neto}@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt/

Abstract. Automatic transcription of telephone speech involves additional challenges compared to wideband data processing, mainly due to channel limitations and to particular characteristics of conversational telephone speech. While in TV speech recognition applications, such as automatic transcription of broadcast news, the presence of telephone data is nearly insignificant (less than 1 %), in most radio broadcast stations the presence of telephone speech grows significantly. Thus, transcription of telephone speech data deserves special attention in radio broadcast applications. In this work, we describe our initial efforts to tackle this particular problem. First, a telephone channel classifier is proposed to automatically detect telephone segments. Then, some strategies for increasing robustness of the automatic transcription system are investigated.

Keywords: Speech recognition, radio broadcast transcription, telephone speech processing, channel classification.

1 Introduction

Continuous advances in speech and language technology, and more concretely, in automatic speech recognition (ASR) have made possible the development of successful very large vocabulary continuous speech recognition systems in certain constrained conditions. Particularly, high quality speech – free of noise and reverberation – and planned non-spontaneous speaking style are usually required.

Due to the generally favorable speech data characteristics, automatic transcription of TV broadcast news has been one of the application fields that has received major attention by the research community. As a consequence, several research groups worldwide have developed their own high performance broadcast transcription system for different languages [1,2,3]. In the particular case of the European Portuguese language, the AUDIMUS.media system described in [4] is up to our knowledge the most successful one.

In the context of our actual research projects, we are currently investigating application of broadcast news transcription technology to the problem of automatic transcription of commercial radio broadcast stations.

A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, pp. 172–181, 2008.

Although the TV and radio broadcast transcription problems share many similarities, there are some major differences that make the radio broadcast problem more challenging. Mainly, there is a considerable increase in the amount of telephone data that is present in radio broadcast programs compared to TV shows, where most of the speech data is wideband data recorded in a free of noise environment.

On the one hand, the problems of speech recognition in telephone applications are very well-known. In addition to the inner limitations of narrow band speech, in most cases a considerable presence of environmental noise appears due the use of mobile telephones in adverse environments. Consequently, the performance of speech recognition systems well-matched to the clean wideband problem fail dramatically in these conditions.

On the other hand, increase of the amount of telephone speech in radio programs is usually related with the presence of live press conferences, interviews to personalities, audience calls and participation of journalists out of the studio. In general, a common characteristic of these telephone contributions is that they are highly spontaneous. This fact results in similar difficulties to the problems of conversational telephone speech recognition, which is known to be significantly more challenging than the transcription of broadcast news [5,6]. Actually, this problem has been extensively tackled in the context of the Switchboard [7] benchmark tasks for the English language.

In this work, automatic detection and transcription of excerpts of telephone speech in radio broadcast data is investigated and some directions for future improvements are drawn. For this purpose, a small corpora of one complete day of broadcast of a Portuguese commercial station was collected and telephone segments were manually transcribed.

With respect to the telephone/non-telephone speech detection, a channel classifier based on linear discriminant analysis (LDA) of logarithmic filter bank energies is proposed.

Regarding the adaptation of the TV broadcast news system for tackling the problem of conversational telephone speech, initial efforts have been focused on the acoustical missmatch problem. New phonetic classifiers for connectionist speech recognition system [8] have been trained using both downsampled TV broadcast news data and real telephone (fixed and mobile) speech data. A considerable improved performance was achieved compared to alternative use of phonetic networks trained only with TV broadcast news data (11.8 % relative word error reduction) or only with telephone data (28.5 % relative word error reduction).

The rest of this paper is organized as follows. The two baseline systems for TV broadcast news transcription and telephone speech recognition are described in next section. Corpora considered in the work is reported in Section 3. Sections 4 and 5 are respectively devoted to the description of the proposed telephone channel classifier and to the developed radio telephone transcription system. Some future work and challenges are also drawn at the end of Section 5 before the concluding remarks.

2 Baseline Transcription Systems

2.1 TV Broadcast News Transcription System

In this work, the broadcast news transcription (BNT) system for the European Portuguese language described in [4] is adapted to the particular needs of radio telephone broadcast speech. A block diagram of the BNT system is shown in Figure 2.1.

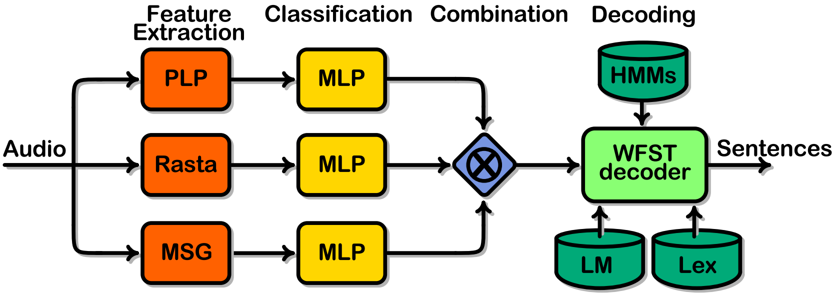

Fig. 1. Block diagram of the broadcast news transcription system after [4]

The system is based on the hybrid ANN/HMM paradigm for speech recognition [8]. This kind of recognisers are generally composed by a phoneme classification network, particularly a Multi-Layer Perceptron (MLP), that estimates the posterior probabilities of the different phonemes for a given input speech frame (and its context). These posterior probabilities are associated to the single state of context independent phoneme hidden Markov models (HMM). An appealing characteristic of the hybrid systems is that they are very flexible in terms of merging multiple input streams.

Concretely, the BNT system combines three network outputs trained with Perceptual Linear Prediction (PLP) features (13 static + first derivative), log-RelAtive SpecTrAl (log-RASTA) features (13 static + first derivative) and Modulation SpectroGram (MSG) features (28 static). In addition to the feature representation, MLP networks are characterized by the size of their hidden layers (2 hidden layers of 2000 units) and the size of the output layer (39 phonemes including silence pattern). The phonetic networks of the recognizer have been trained and adapted along years of speech recognition research with 57 hours of manually annotated data (46 train + 11 development) and more than 300 hours of automatically transcribed broadcast news data.

The decoder of the BNT system is based on weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [9]. In this approach, the decoder search space is a large WFST that maps observation distributions to words. The language model (LM) in the one described in [10] with an active lexica size of 100K word. It is build based on a daily and unsupervised adaptation approach which dynamically adapts the active vocabulary and LM to the topic of the current news. Thus, a remarkable reduction of the out-of-vocabulary (OOV) words and of the word error rate (WER) is achieved.

In an evaluation set composed of 6 broadcast news programs recorded in 2007, the best WER performance achieved up to now with the BNT system is 20.6% for all conditions and 10.1% for F0 condition (read speech in studio). Current state of the art BN ASR systems for the English language have WER performances of less than 13% with 10x Real-Time [1] and less than 16% in real-time [6]. It is worth to notice that Portuguese BNT system results are in real-time performance.

2.2 The Telephone Speech Recognizer

A telephone speech recognizer (TSR) similar to the one described in [11] – known as AUDIMUS.telephone – has also been considered in this work for baseline comparison. The recognizer is particularly developed for both fixed and mobile telephone dedicated applications, such as automatic informational retrieval systems based on voice-command operated dialogs.

The architecture of the TSR system is the same multi-stream ANN/HMM paradigm shown in Figure 2.1. Main differences rely on the MLP networks and the corpora used for training them.

The phone classification networks were trained following the refrec 0.96 training procedure for SpeechDat [12] (without *garbage* model) using fixed telephone data (∼115 hours) and mobile telephone data (∼6 hours). Actually, the total amount of effective data was considerably reduced due to the unbalanced representation of some phone patterns and to the excessive amount of silence (approximately 36 % in fixed telephone and 50 % in mobile telephone data). Finally, networks with 7 window context and a unique hidden layer of 1500 units were trained. The different size with respect to previous BNT system is basically due to the different amount of available data.

In Table 1, WER results on SpeechDat II test sets proposed in [12] are shown together with the performace of some reference systems in other languages.

Table 1. WER results of the telephone speech recognizer (TSR) for Portuguese language compared to other language references after [12]. SpeechDat test categories are: isolated digits (I), yes/no (Q), application words (A), connected digits (BC), city names (O) and phonetically rich words (W).

Language	I	Q	A	BC	O	W
TSR system	0.4	0.1	1.8	8.2	5.6	6.8
Danish	0.0	0.3	1.9	2.4	13.8	46.2
English	3.5	0.0	0.8	4.4	6.0	30.8
German	0.0	0.0	1.7	2.8	5.3	7.1
Norwegian	3.5	0.0	2.8	5.3	14.9	22.1
Slovenian	5.2	1.2	3.5	4.7	7.3	15.9
Swiss German	0.2	1.0	0.6	2.5	9.2	25.0

3 Corpora Description

TV broadcast news data and fixed and mobile telephone speech data were used on the training and development of the telephone radio broadcast speech recognition system. Additionally, real radio data was collected for both development of a telephone classifier and for evaluation of this transcription system.

3.1 TV Broadcast News Corpus (TVBN)

The TV Broadcast News Corpus in an excerpt of the data collected from April 2000 to January 2001 to support the research and developments associated with automatic transcription of Portuguese BN. A total of 123 programs have been considered with an approximate duration of 57 hours. This corpus was divided in two sets: training (46 hours) and development (11 hours). Audio data is stored at 16 kHz sampling and 16 bits PCM encoding.

3.2 Fixed Telephone Corpus (FT)

A sub-set of the well-known Portuguese SpeechDat corpora have been used for training and development purposes. Concretely, the training data set consists of 24 hours of phonetically rich sentences and 12 hours of spontaneous speech from the SpeechDat II database. The development data set consists of 7 hours of phonetically rich sentences and 2 hours of spontaneous speech from the SpeechDat I database.

3.3 Mobile Telephone Corpus (MT)

Mobile telephone data of about 800 sessions recorded from the mobile GSM network in Portugal following the model of SpeechDat (yes/no categories, digit strings, application words...) was also considered. Non-spontaneous speaking style is dominant. Additionally, although being mobile data, there is not a significant amount of background noise. Data was also split into training ($\sim$ 11 hours) and development sets ($\sim$ 2 hours).

3.4 Radio Broadcast Corpus (RB)

One entire day, that is 24 hours, of a Portuguese commercial radio station was collected at 16 kHz sampling frequency. The data was used for developing the channel classifier described in next Section 4 in order to automatically detect the segments of telephone speech. The classified segments resulted in 116.6 minutes of telephone speech data. These telephone segments were orthographically transcribed to define the test data set of the speech recognition system for telephone radio broadcast data. Hereinafter, this test sub-set of only radio telephone speech will be referred to as RTB. For this corpus, the OOV word rate with the 100K words vocabulary is 0.33 %. Most of the OOV words are typical forms of conversational speech, such as clitics and some verb conjugations.

4 Detection of Telephone Segments in Radio Broadcast

Telephone channel is characterized by narrow band transmission in the frequency range from 300 Hz to 3400 Hz. Thus, a simple way for detecting telephone speech consists on computing energies in the different frequency bands and classify it.

Concretely, 15 logarithmic filter bank energies of 20 msec frames at 16 kHz sampling frequency are extracted with a time shift of 10 msec. The feature vectors are complemented with their first derivative. Then, each speech frame is classified with a binary LDA classifier into non-telephone or telephone classes.

In a first stage, the LDA classifier was initially trained with less than 4 minutes of telephone data and around 5 minutes of randomly selected non-telephone data (also including music and jingles) that were manually extracted from the RB corpora. A small portion of the training data was used for validation purposes. The rate of correct classified frames in the validation data set was of 99.8 %.

This initial classifier was then used to detect telephone segments in the whole RB corpus (24 hours). According to automatic frame classification, the segments with more than 1 second of duration and with a rate of telephone class labels above a fixed threshold were marked as telephone segments. Then, these automatically detected telephone segments were manually validated. In general, only few errors could be observed due to short telephone segments missed because of simple detection rule and mainly false positive detection corresponding to segments of music and jingles.

This new telephone segmentation (partially automatic) resulted in 116.6 minutes of telephone data. This sub-set constitutes the RTB test set. Notice that these almost two hours represents around 8 % of one complete day of radio broadcast, which is quite significant if it is taken into account that in the rest of the data there is a significant amount of non-speech acoustic events.

In a second stage, semi-automatically detected telephone segments together with around 4 hours of non-telephone data randomly selected and extracted from the same RB corpus were used to train a new LDA classifier. Again a short set of data was used for validation. In this case, the rate of correct classified frames in the validation data set was of 96.5 %. The drop in the classification rate is due to the higher variability in the data used for training. However, a generalized improved performance of the resulting classifier could be observed, particularly, when it is combined with a robust speech-non-speech detector that permits rejecting music and other non-speech acoustic events.

5 Automatic Transcription of Radio Telephone Speech

The object of detecting telephone speech segments in radio broadcast data is to apply a different processing to that applied to regular wideband data.

In the context of speech recognition area there is a countless number of robust techniques aimed to improve the performance of telephone speech recognition in different domains such as speech enhancement, robust feature extraction or acoustic model adaptation [13]. Additionally, attending to the fact that most of telephone data corresponds to spontaneous and even conversational speech, adaptation of the lexica and of the language model might also provide some additional benefits.

However, in this work we have focused only in the construction of robust acoustic phonetic classifiers adapted to the characteristics of conversational telephone data. The language model used in all the following experiments is the one described in Section 2.1 and the large lexicon of 100K words.

5.1 Baseline Systems Performace

Some initial experiments were carried out to confirm the need of developing new phonetic classification networks matched to the characteristics of telephone speech radio broadcast shows.

The well-trained system for transcription of Portuguese BN described in Section 2.1 and the set of well-trained phonetic classifiers for automatic recognition of telephone speech described in Section 2.2 were assessed with the RTB test set. In the case of the telephone dedicated system, the RTB test set was downsampled to 8 kHz.

Table 2 shows the performance of the two baseline systems. The TSR system, which is entirely trained with telephone speech data, is not well-matched to the problem of continuous speech recognition and obtains a poor WER of 72.0 % for all conditions. On the other hand, the BNT recognizer achieves a considerable error reduction with respect to the TSR system (WER of 58.4 % for all conditions). Despite the BNT system is trained with wideband data and suffers from channel missmatch problem, it is more appropriate for this concrete task. However, its performace in both planned and spontaneous speech is still far of the reference results provided in Section 2.1 obtained on TV broadcast news data. In general, it can be clearly stated that both systems fail to provide a reasonable performace independently of the speaking style. These observations are in well-accordance with [5], where a state of the art BN transcription system had a WER of around a 50 % in Switchboard data.

5.2 Robust Network Training

The TVBN corpus, the FT corpus and the MT corpus are used to develop a system in more accordance to the needs of telephone speech in radio broadcast shows. The combination of the three corpora results in a training set of approximately 93 hours (46 TVBN + 36 FT + 11 MT) and a development set of 22 hours (11 TVBN + 9 FT + 2 MT). The development data is used to define the stopping criteria in the process of training the MLPs as it is usually done in this kind of approaches.

Manually generated transcriptions were used to obtain frame-to-phone alignments needed for training the phonetic classifiers. In the case of TVBN data, the BNT system was used; while the TSR system was used to align telephone data, both fixed and mobile.

As in the case of the previous systems, a multistream system is built with PLP, log-RASTA and MSG feature parametrizations. Network characteristics are similar to those of the BNT system, but the size of the two hidden layers was fixed to 1500 due to the reduced amount of available data.

Phonetic networks are trained with 8 kHz sampling rate data, thus speech from the TVBN corpus was previously downsampled. In order to simulate more accurately telephone channel characteristics, we experimented to apply an additional pass-band filtering stage in the frequency range of telephone speech. However, not remarkable differences were found depending on wether telephone-like filtered or not filtered data was used for training the networks. Thus, the results shown in this work were obtained with downsampled data without filtering.

In next Table 2 the WER performance of the new proposed system referred to as the radio telephone transcriber (RTT) is compared to the two previous baseline systems. Two different test conditions are considered: planned and spontaneous speech speaking style. The WER average of the two conditions is also provided.

Table 2. WER results of the telephone speech recognizer (TSR), the broadcast news transcriber (BNT) and the radio telephone transcriber (RTT) in planned and spontaneous test conditions

	TSR	BNT	RTT
planned	66.2	51.7	43.7
spontan	85.2	69.3	64.3
average	72.0	58.4	51.5

In both planned and spontaneous speech, the new RTT system achieves a considerable improvement with respect to the best baseline system performance. The most noticeable improvement is obtained in planned speech condition. In this case, the impact of missmatched language model is less important and a 15.5% relative error reduction is obtained with respect to the BNT system thanks to better acoustic modeling. However, the relative improvement in spontaneous condition is quite lower due to main influence of inappropriate language modelling. For all conditions, the RTT system achieves a relative word error rate reduction of 28.5% with respect to the TSR system and 11.8% with respect to the BNT system.

5.3 Future Work and Challenges

According to the reported results of our ongoing research activities, it is clear that many challenges still need to be faced in order to achieve reasonable performance of transcription of telephone speech in radio broadcast applications.

On the one hand, there is a need for real conversational telephone speech data in Portuguese language to develop robust acoustic modelling. In this work, it has been shown that the use of other sources of data can help to alleviate this problem, but it is not a definite solution. With regards to other robustness issues, we are currently investigating the use of alternative feature parametrizations that might better match the telephone speech recognition problem, such as the standard advanced front-end of ETSI [14]. Additionally, the use of speaker normalization techniques like vocal tract length normalization (VTLN) [15] is being investigated.

On the other hand, adaptation of current broadcast news language model to better match the spontaneous speaking style of conversational speech appears as a necessary step for future improvements. Thus, the problem of corpora resources is present again, since there exists a limited amount of language model training data of the desired characteristics.

Finally, it must be noticed that both TV broadcast news transcription and telephone based information retrieval systems are usually limited by the need of real time functionality. However, there is not need for real time limitations in most applications of speech recognition to radio broadcast data. The typical application is to generate information (transcriptions) of already stored data. In this case, more computational demanding decoding strategies can be applied. For instance, multiple stage decoding steps based on adaptation and re-decoding of automatically transcribed data.

6 Conclusions

Everyday huge amounts of multimedia data are generated by broadcast media world-wide, which would be desirable to have automatically segmented and transcribed. Actually, there exist real systems capable of providing accurate transcriptions in some contexts, such as in TV broadcast news applications. In this work, we have started to investigate the possible re-usability of a broadcast news transcription system for the European Portuguese language to the similar radio broadcast transcription problem. The main challenges that one can find are the significant increase of both telephone speech and spontaneous speaking style. Thus, we have initially focused on the automatic detection of telephone speech and the improvement of phonetic acoustic modelling for particular conversational telephone speech. A relative WER reduction of 11.8% was achieved with respect to the broadcast news system, besides an high classification rate of the telephone channel detector proposed. Finally, some thoughts for future development have been provided.

Acknowledgements

This work was funded by PRIME National Project TECNOVOZ number 03/165.

References

1. Nguyen, L., Xiang, B., Afify, M., Abdou, S., Matsoukas, S., Schwartz, R., Makhoul, J.: The BBN RT04 English Broadcast News Transcription System. In: Proceedings of Interspeech 2005, Lisbon, Portugal (2005)

2. Gales, M.J.F., Kim, D.Y., Woodland, P.C., Chan, H.Y., Mrva, D., Sinha, R., Tranter, S.E.: Progress in the CU-HTK Broadcast News Transcription System. IEEE Transactions on Audio, Speech, and Language Processing 14(5), 1513–1525 (2006)
3. Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., Gravier, G.: The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In: Proceedings of Interspeech 2005, Lisbon, Portugal (2005)
4. Meinedo, H., Caseiro, D., Neto, J., Trancoso, I.: AUDIMUS.media: A Broadcast News speech recognition system for the European Portuguese language. In: Proceedings of PROPOR- 2003, Portugal (2003)
5. Gauvain, J.-L., Lamel, L., Schwenk, H., Adda, G., Chen, L., Lefèvre, F.: Conversational telephone speech recognition. In: Proceedings of ICASSP-2003, pp. 212–215 (April 2003)
6. Matsoukas, S., Prasad, R., Laxminarayan, S., Xiang, B., Nguyen, L., Schwartz, R.: The 2004 BBN 1xRT Recognition Systems for English Broadcast News and Conversational Telephone Speech. In: Proceedings of Interspeech 2005, Lisbon, Portugal (2005)
7. Godfrey, J.J., Holliman, E.C., McDaniel, J.: Switchboard: Telephone speech corpus for research and development. In: Proceedings of ICASSP-1992, pp. 517–520 (March 1992)
8. Morgan, N., Bourlard, H.: An introduction to hybrid HMM/Connectionist continuous speech recognition. IEEE Signal Processing Magazine, 25–42 (1995)
9. Mohri, M., Pereira, F., Riley, M.: Weighted finite-state transducers in speech recognition. In: ISCA ITRW Automatic Speech Recognition, Paris, pp. 97–106 (2000)
10. Martins, C., Teixeira, A., Neto, J.: Dynamic language modeling for a daily broadcast news transcription system. In: Proceedings of ASRU-2007, Kyoto, pp. 165–170 (2007)
11. Hagen, A., Neto, J.: HMM/MLP Hybrid Speech Recognizer for the Portuguese Telephone SpeechDat Corpus. In: Proceedings of PROPOR-2003, Portugal (2003)
12. Lindberg, B., Johansen, F., Warakagoda, N., Lehtinen, G., Kacic, Z., Zgank, A., Elenius, K., Salvi, G.: A noise robust multilingual reference recogniser based on SpeechDat(II). In: Proceedings of ICSLP 2000, Beijing, pp. III, 370–373 (2000)
13. Junqua, J.-C., Haton, J.P.: Robustness in Automatic Speech Recognition: Fundamentals and Applications. Kluwer Academic Publishers, Dordrecht (1996)
14. ETSI standard doc.: Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced feature extraction algorithm. ETSI ES 202 050 Ver. 1.1.5 (2002)
15. Kamm, T., Andreou, G., Cohen, J.: Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. In: Proceedings of the 15th Annual Speech Research Symposium, Baltimore, USA (1995)