Genre Categorization and Modeling for Broadcast Speech Transcription

Qingqing Zhang, Lori Lamel, Jean-Luc Gauvain

Spoken Language Processing Group, LIMSI-CNRS
BP 133, 91403 Orsay cedex, France
{qing,lamel,gauvain}@limsi.fr

10.21437/Interspeech.2011-188

Abstract

Broadcast News (BN) speech recognition transcription has attracted research due to the challenges of the task since the mid 1990's. More recently, research has been moving towards more spontaneous broadcast data, commonly called Broadcast Conversation (BC) speech. Considering the large style difference between BN and BC genres, specific modeling of genres should intuitively result in improved system performance. In this paper BN- and BC-style speech recognition has been explored by designing genre-specific systems. In order to separate the training data, an automatic genre categorization with two novel features is proposed. Experiments showed that automatic categorization of genre labels of the training data compared favorably to the original manually specified genre labels provided with corpora. When test data sets were classified into BN or BC genres and tested by the corresponding genre-specific speech recognition systems, modest but consistent error reductions were achieved compared to the baseline genre-independent systems.

Index Terms: Broadcast news, broadcast conversation, automatic genre categorization, Mandarin speech recognition

1. Introduction

Broadcast News (BN) speech recognition has been an active research topic in ASR for more than 15 years. More recently, research on this task has been moving towards Broadcast Conversation (BC) speech that is produced spontaneously and in conversational settings. In contrast to the well-planned speech that is characteristic of the BN domain, BC speech is more interactive and spontaneous, corresponding to free speech in news-style TV and radio programs such as talk shows, interviews, call-in programs, live reports, and round-tables [1]. As opposed to BN, BC speech recognition is more challenging due to the changes of speaking style, speech rate, hesitation, filled pauses etc. It is well acknowledged that state-of-the-art speech recognition systems perform considerably worse on BC data than on BN data, with word error rates about three times higher [2]. Considering the differences between BN and BC data, intuitively it is better to separate BN and BC data, and build corresponding models to improve the recognition accuracy for these two genres.

To find how much performance gain can be obtained by explicit modeling of these two genres, genre-specific speech transcription systems were investigated in this paper. Since BC and BN genres are different not only acoustically but also grammatically, besides acoustic models, genre-specific language models were also explored in this work. To build genre-specific systems, all the data needs to be classified into BN and BC. The Linguistic Data Consortium (LDC) provided each show or each snippet[1] with a single genre label. A close examination shows that both genres can occur within a single show or snippet. In order to improve the accuracy for BN/BC genre classification, a more precise genre categorization is clearly needed. Some previous work attempting to automatically categorize audio data into BN and BC genres explored occurrence counts of lexical N-grams and selected words that had the highest mutual information for class variability [1]. This method relies on the precise transcriptions of speech. For the test set, if no reasonable hypothesis could be obtained, the performance of this method will decrease.

In this paper, we investigate an efficient automatic genre categorization method for a large vocabulary Mandarin ASR system. Motivated by human perception, two distinctive features to categorize BN/BC genres are proposed: one is the Normalized Duration Variance related to speech style (prepared or spontaneous), and the other is the Silence Ratio related to speech continuity. Both features are extracted from audio data directly. Our automatic categorization classifies input audio into BN and BC genres based on these two features. This method does not rely on high accuracy speech transcriptions, thus it can be used to classify the training data for which there are transcriptions and as well as the test data using a less accurate automatic hypothesis. The automatic genre categorization was compared with the LDC provided categorization. Using the labels provided by automatic genre categorization, genre-specific acoustic models and language models were built and compared with the genre-independent models on separated BN and BC subsets and the full data set. Experiments were carried out using a single-pass decoding system and also a multi-pass one with unsupervised adaptation.

2. Genre-independent and Genre-specific modeling

Approximately 1640 hours of Mandarin BN and BC speech data from the DARPA GALE program are used for training. All of the data are assigned to BN or BC using the genre label for each show released by LDC. For test sets, LDC provided manual snippets, each snippet being labeled as BN or BC. The LIMSI P5 Gale Speech-To-Text system (similar to that described in P4 system[2]) is used as a starting point to explore acoustic and language modeling. The language model of this system is trained with large amounts of Mandarin data thus providing the system with robust LM estimates. The language model training data consists of 48 different text sources. The total amount of data available for training is 3.2 billion word tokens (af-

[1] Annotators listened to the audio files and selected short snippets of approximately 1-2 minutes' duration. Each snippet was chosen to contain a single topic.

ter segmentation)[3]. A 56K vocabulary is used for language model and recognition. This system uses multiple discriminatively trained acoustic models with both cepstral and probabilistic features, which have been growing in popularity.

Probabilistic features produced by Multi Layer Perceptron (MLP) in STT transcription systems, have been shown improve system performance when concatenated with cepstral features [4-8]. In our experiments, acoustic models were trained with probabilistic features obtained from MLP. To extract MLP features, the wLP-TRAP raw features [9] [10] are used as the input to the 4-layer bottle-neck MLP [8], and the MLP features are taken from the "bottle-neck" hidden layer and de-correlated by a PCA transformation.

To extract the Genre-Independent (GI) MLP features, the MLP and the PCA transformation are estimated from the combined BN and BC training data. For the Genre-Specific (GS) MLP training, according to the genre labels released by LDC, there are about 805 hours of BN data and 835 hours of BC data. We use the corresponding data to train MLPs and PCA transformations for BN and BC respectively. The cross-validation (CV) frame accuracies for GI MLP and GS MLP are shown in Table 1. The GS MLP outperforms the GI MLP on classification of cross-validation data. Consistent improvements are observed for the BN data (1.5%) and for the BC data (1.7%).

Table 1: % Cross-validation frame accuracies for BN and BC data based on Genre-independent (GI) and Genre-specific (GS) MLP training

CV frame accuracy%	BN	BC
GI MLP	59.3	52.0
GS MLP	61.0	53.5

For comparison, all of GI and GS MLP HMMs were trained with 81-dimensional MLP+PLP+F0 feature vectors, which were formed by the 39-dimensional MLP feature vector and the 42-dimensional PLP+F0 feature vector. All the acoustic models are sets of 3-state left-to-right HMM with Gaussian mixture, with 32 Gaussians per state. Gender-dependent triphones are adapted from gender independent models with MAP. Besides MLP features, we also train GI and GS models based on standard 42-dimensional PLP+F0 features. For the PLP models, a maximum-likelihood linear transform (MLLT) is also used.

The GALE Phase 4 development subset dev09s is used to assess the models. The dev09s has 3.0 hours of speech data containing 182 snippets. LDC provided the genre label for each snippet, with 92 snippets in BN and 90 snippets in BC. For a first set of comparisons word recognition is performed in a single pass decoding. This is carried out via one pass cross-word trigram decoding with gender-specific sets of position-dependent triphones and a trigram language model. The trigram lattices are rescored with a 4-gram language model. The words with the highest posterior are hypothesized as the final recognition results [2].

Table 2 shows the Character Error Rates (CERs) on the whole set of dev09s data and the BN and BC subsets with GI/GS MLP+PLP+F0 models and PLP+F0 models(All the acoustic models were trained based on BN or BC data using the genre label for each show, and all the test sets were classified with the genre label for each snippet). Results from both PLP+F0 system and MLP+PLP+F0 system show that GS acoustic models achieve 4% relative reduction on BC compared to GI acoustic models. But on BN subset, GI acoustic models outperform. On the entire dev09s set, modest but consistent reductions in recognition error rate are achieved by GS acoustic models.

Table 2: % The CER of Genre-independent (GI) systems and Genre-specific (GS) systems based on LDC provided genre labels in the first pass decoding

CER%	dev09s(bc+bn)	dev09s_bc	dev09s_bn
GI PLP+F0	13.8	19.2	8.0
GS PLP+F0	13.6	18.5	8.5
GI MLP+PLP+F0	12.0	17.3	6.4
GS MLP+PLP+F0	11.9	16.7	6.8

In order to have a complete set of contrastive results the GS models were also used to decode the data of the other genre. Results are thus obtained for the full dev09s data sets with both sets of GS models (BN and BC), and performance can be compared under matched and mismatched genre conditions. As shown in Table 3, improved results can be achieved on BC subset when BC models are used, but on BN subset, the matched BN models do not outperform the mismatched BC models, which leads to the suspicion that in the test set some of the data manually classified as BN by LDC are conversational-like, likely because the classification is made at a relatively coarse level.

Table 3: % The CER of Genre-specific (GS) models when decoding matched/mismatched genres of subsets on dev09s

CER(Del,Ins)%	dev09s(bc+bn)	dev09s_bc	dev09s_bn
BC PLP+F0	13.4(3.1,1.3)	18.5(4.5,2.0)	8.1(1.7,0.6)
BN PLP+F0	14.6(3.4,1.3)	20.3(4.9,1.9)	8.5(1.8,0.6)
BC MLP+PLP+F0	11.8(2.5,1.3)	16.7(3.6,2.1)	6.7(1.4,0.5)
BN MLP+PLP+F0	12.9(2.9,1.3)	18.8(4.3,2.1)	6.8(1.4,0.5)

The results also show that regardless of types of features (PLP+F0 or MLP+PLP+F0), when BN models are used to decode BC data, the Deletion Error rate (Del) increases and the Insertion Error rate (Ins) decreases. This could be caused by the more casual style and faster, less carefully articulated speech in the BC portion. One possible explanation is when the BN model is used to decode BC data, the BN model cannot catch up the fast speed, so words tend to be deleted. The reverse trend can be seen when the BC model is used to decode the BN data, the Insertion Error rate tends to increase and the Deletion Error rate tends to decrease, but to a lesser extent.

3. Automatic genre categorization

As discussed previously, not only some BN/BC data were mislabeled and included in the wrong subset in LDC data, but also it only has genre labels at whole show or snippet level. Considering the fact that even within a given snippet both BN and BC can occur, a more detailed and accurate labeling on smaller unit is indeed needed for improving the modeling precisions of BN and BC. However given the huge amount of training data, it is impossible to get all the data checked manually. An automatic categorization is needed to re-classify all the data.

In this paper, we investigate an efficient automatic genre categorization method with smaller units: speaker clustered segments. The speaker clustered segments are obtained from the data partitioning [2]. Data partitioning, based on an audio stream mixture model, serves to divide the continuous stream of acoustic data into homogeneous segments, and associates cluster, gender and labels with each non-overlapping segment. The output of the partitioning process is a set of speaker clustered speech segments. With this automatic segmentation, one segment is contiguous for one speaker in one acoustic condition

(The speech from the same speaker may occur in different parts of the broadcast, and with different background noise conditions), and there is no mixture of BN and BC segments within one segment theoretically. These speaker clustered segments are appropriate units for genre categorization.

For automatic genre categorization, motivated by characteristic differences between BN/BC genres, two distinctive features extracted from the audio data directly are proposed: one is the Normalized Duration Variance related to speech style, and the other is the Silence Ratio related to speech continuity. Our automatic categorization classifies input audio into BN and BC genres based on these two features. The detailed algorithm is described as follow.

3.1. Categorization algorithm

3.1.1. Normalized Speed Variance

Compared to BN speech which is primarily prepared in studio conditions, BC speech is more spontaneous. It consists of talk shows, debates, and interactive programs, etc. It is unusual for BC speakers to keep a stable speed when talking. This leads a larger variation in syllable/phone durations for BC speech. On the other hand, the anchors in BN are well trained. When reporting news, they keep their speed reasonably stable (less than 300 syllables per minute nowadays). Thus it is reasonable to assume if the speed of a sentence varies largely, this sentence tends to be BC instead of BN.

The proposed Normalized Duration Variance (NDV) aims to differentiate the two speaking styles (prepared or spontaneous) based on the speech speed and the speech variation. The mean of syllable durations obtained from forced-alignment at the syllable level in a segment is used to represent the speech speed and the variance of syllable durations is used to represent the speech variation, and the NDV is defined as follows:

$$\mathrm{NDV} = \frac{\sqrt{V(X)}}{\mathrm{E(X)}} \tag{1}$$

E(X) and V(X) refer to the mean and the variance of syllable duration respectively. When the variance of syllable duration is large and the speech speed is fast, the NDV will increase, and this segment is more likely to be BC. Otherwise, it should be BN.

3.1.2. Silence Ratio

Besides differences in speech speed, BN and BC genres also have different characteristics in non-speech parts. BN speech is more fluent and there are not as many interruptions, so the occurrences of pauses/silences should be much less than that of BC speech. We exploited this phenomenon with the proposed Silence Ratio feature. The Silence Ratio (SR) standing for the average silence occurrence per syllable is defined as below:

$$SR = \frac{\#Silence}{\#Syllable} \tag{2}$$

#Silence denotes the number of silences and #Syllable denotes the number of syllables. The larger the SR is, the more likely this segment is BC.

In our experiments, only silences longer than 200 ms[2] were taken into account (any segment beginning silence or ending silence was excluded). Silences shorter than 200 ms are considered as the necessary pauses between words when people are talking no matter in BN or BC styles.

[2]200 ms is shortest average duration for each character in BN style according to anchors speeds nowadays.

3.1.3. Genre categorization

For categorization, NDV and SR, characterizing the differences between BN and BC data, are used as distinctive features in our Automatic Genre Categorization (AGC) which is defined as follows:

$$\mathrm{AGC} = \alpha \times \mathrm{NDV} + (1-\alpha) \times \mathrm{SR} \tag{3}$$

Here α denotes the interpolation weight. In our experiment, equal weights are given for NDV and SR.

BN/BC genre decision is made based on a preset threshold. When one segment has an AGC larger than the threshold, it will be classified as BC. Otherwise, it will be classified as BN.

Since AGC obtained from different languages can be different, the threshold for genre decision is optimized by a development data set. Dev09s is used as the development set. We manually checked this dev set and labeled the genre information for each speaker clustered segment. The optimal threshold is obtained when the minimum classification error rate is achieved compared to the manually checked classification.

3.2. Categorization comparison

To test the effectiveness of AGC and compare it with the LDC provided labels, we keep using the GS acoustic models described in Section 2, but just change the categorizations for the test set. There are 320 speaker clustered segments in dev09s. Compared to genre labels provided by LDC (153 segments in BN and 167 segments in BC), there are 40 segment genre labels changed by AGC (126 segments in BN and 194 segments in BC). Table 4 gives the results of MLP+PLP+F0 features based on different categorizations. When the automatic genre categorization is used to classify BN and BC genres, it outperforms the LDC provided labels. Compared with the manually checked classification, our automatic genre categorization achieved the same level of accuracy.

Table 4: % The CER of the BN/BC subsets based on different categorizations

CER %	LDC label	AGC	Manual check
dev09s	11.9	11.8	11.8

Results show that the proposed AGC method, albeit its simplicity, can capture most characteristics of underlying differences between BN and BC, and works as well as the manual classification. Therefore, all the training data were re-classified by the proposed AGC. There are 95728 speaker clustered segments in the training set. Compared to genre labels provided by LDC (44755 segments in BN and 50973 segments in BC), 18375 segment genre labels are changed by AGC (39381 segments in BN and 56347 segments in BC). After automatic genre categorization, there are about 783 hours of BN data and 857 hours of BC data.

4. Genre-specific system based on AGC

With the new genre labels provided by AGC, new genre-specific models are trained. For the acoustic part, the corresponding genre data are used to retrain the GS MLP and PLP models. Since BC and BN genres are different not only acoustically

but also grammatically, besides acoustic models, GS language models are also explored based on the new labels. For the GS language modeling, the AGC is used to separate all the training sources into BN and BC subsets. The BN subset is used to train a small BN LM and the BC subset is used to train a small BC LM. Considering the baseline GI LM was trained by all of text sources, it could be more stable, the GS BN/BC LM is thus obtained by the baseline GI LM interpolated with the small BN/BC genre LM. Dev0910, which includes the development and evaluation sets of GALE Phase 4 and all the development sets of Phase 5, is separated into dev0910_bc and dev0910_bn subsets by AGC, and is used to optimize the LM interpolation weights.

The perplexities of dev0910_bc and dev0910_bn with GS and GI 4-gram LMs are shown in Table 5. Compared to the GI baseline LM, the perplexity of the BC subset with GS LMs is reduced by about 5% relatively, but the perplexity of the BN subset changes little.

Table 5: perplexities of GI and GS LMs.

Perplexity	dev0910_bn	dev0910_bc
GI LM	165	256
GS LM	161	243

Performances between these rebuilt GS systems and GI systems are compared in Table 6 based on MLP+PLP+F0 features. To investigate GS systems thoroughly, we compared three conditions. The first one is to use GI AMs and GS LMs. The second one is to use GS AMs and GI LMs. The last one is to use both GS AMs and GS LMs. Compared to the baseline GI system, only changing to use GS LMs hardly improves the performance, but when GS AMs are used, some improvements can be obtained (0.3% absolute). When both GS AMs and GS LMs are applied, results on full dev09s set show that 4.6% relative error reduction is achieved compared to the GI system.

Table 6: % The CER of GI systems and GS systems based on AGC labels in the single-pass decoding.

CER%	dev09s(bc+bn)	dev09s_bc	dev09s_bn
GI AM, GI LM	12.0	18.0	4.7
GI AM, GS LM	11.9	17.9	4.7
GS AM, GI LM	11.6	17.6	4.6
GS AM, GS LM	11.4	17.2	4.5

5. Multi-pass adaptation

The above results have shown that for each genre improvements were achieved by using genre-specific systems. These results are based on the one-pass decoding. In this section, a complete multi-pass adaptation framework is used to evaluate the systems based on the previous results. In LIMSI STT multi-pass adaptation systems, the second- and third-pass decoding performs unsupervised acoustic model adaptation for each segment cluster using the CMLLR and MLLR prior to the next decoding pass. To provide complementary work, PLP+F0 models are used in the 2nd pass, while MLP+PLP+F0 models are used in the 1st and 3rd passes. All the models used here are based on Maximum Mutual Information Estimation (MMIE) training.

Table 7 gives the CERs on eval08, dev09s and dev10c in the first pass (1p) and the third pass (3p) decoding. The eval08 is the subset of the GALE Phase 3.5 evaluation set, containing 4.5 hours of speech data. The dev10c is the subset of the GALE Phase 5 development set, containing 5.7 hours of speech data. After the multi-pass adaptation, compared to the baseline genre-independent systems, modest but consistent error reductions are obtained on all the test sets by using the genre-specific systems.

Table 7: % The CERs of GI and GS multi-pass adaptation systems on eval08, dev09s and dev10c.

CER%	eval08	dev09s	dev10c
GI (1p)	9.7	11.3	16.3
GS (1p)	9.4	11.0	16.2
GI (3p)	8.6	10.1	14.9
GS (3p)	8.4	9.9	14.8

6. Conclusions

The paper has explored genre-specific speech recognition systems for BN and BC, and compared them with the genre-independent system. One key issue in building such a system is how to reliably classify the genre. To avoid costly manual annotation and improve the accuracy for BN/BC genre classification, we proposed an automatic genre categorization at a speaker cluster level based on our two novel features. Experiments showed that the proposed methods performed favorably compared to the original genre labels provided by LDC. When genre-specific systems were built based on this categorization, compared to the genre-independent system, modest but consistent error reductions were achieved.

7. Acknowledgements

This work was in part supported under the GALE program of the Defense Advanced Research Projects Agency, and by OSEO under the Quaero program.

8. References

[1] W. Wang, A. Mandal, X. Lei, A. Stolcke, J. Zheng, "Multifactor Adaptation for Mandarin Broadcast News and Conversation Speech Recognition," Interspeech'09, 2103-2106, Brighton, UK, September 2009.

[2] L. Lamel, J-L. Gauvain, V.B. Le, I. Oparin and S. Meng, "Improved Models for Mandarin Speech-To-Text Transcription," ICASSP'11, 4660-4663, Prague, Czech Republic, May 2011.

[3] I. Oparin, L. Lamel and J.-L. Gauvain, "Improving Mandarin Chinese STT System with Random Forests Language Models," ISCSLP'10, 242-245, Tainan, Taiwan, 2010.

[4] Q. Zhu and A. Stolcke and B.Y. Chen and N. Morgan, "Using MLP features in SRIs conversational speech recognition system," Interspeech'05, 2141-2144, Lisbon, September 2005.

[5] S. Chu, et al. "The IBM 2009 Mandarin Broadcast Transcription System," ICASSP'10, 4374-4377, Dallas, US, March 2010.

[6] X. Liu, M.J.F. Gales, P.C. Woodland, "Language Model Cross Adaptation For LVCSR System Combination," Interspeech'10, 342-345, Makuhari, Japan, September 2010.

[7] T. Ng et al. "Progress in the BBN 2007 Mandarin Speech to Text System," ICASSP'08, 1537-1540, Las Vegas, NV, March 2008

[8] P. Fousek, L. Lamel and J.L. Gauvain, "Transcribing Broadcast Data Using MLP Features," Interspeech'08, 1433-1436, Brisbane, Australia, 2008

[9] P. Fousek, "Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics," PhD, thesis, CzechTechUniv, Prague, 2007.

[10] F. Valente, M. Magimai Doss, C. Plahl, S. Ravuri, and W. Wang, "A comparative large scale study of MLP features for Mandarin ASR," Interspeech'10, 2630-2363, Makuhari, Japan, September 2010