1. Introduction

Based on the GALE project, we build a highly accurate automatic speech recognizer for continuous Mandarin speech, handling broadcast news (BN) and broadcast conversations (BC). This paper summarizes the further development and improvement of our work presented in [1]. The final system is competitive to current Mandarin speech recognizers, [2, 3, 4].

We start by introducing two different toneme sets and pronunciation dictionaries. The LC-STAR toneme set (denoted as RWTH-83) and the corresponding pronunciation lexicon are taken from the preceding system, [1]. In addition, we use a new toneme set and pronunciation lexicon derived from the lexicon of the University of Washington (UW). Section 3 describes the acoustic models based on a MFCC and a PLP front-end in combination with two neural network posterior feature sets [5, 6]. While the computationally most expensive part in discriminative training is the generation of word lattices for a huge amount of data, we present a fast and simple way for porting word lattices between systems using different kinds of neural network toneme posterior features.

While different kinds of systems are developed a two-stage decoding framework for combining these systems is applied. Moreover we show detailed recognition results of the development cycle of the systems. Finally, two methods to integrate tonal features are compared.

Index Terms: Mandarin speech recognition, LVCSR, system combination, feature extraction, VTLN

2. Pronunciation Dictionary and Language Model

The RWTH Mandarin LVCSR system follows a common approach for Mandarin LVCSR systems: word-based toneme pronunciation models [2, 4, 3].

(a) Pinyin-

(b) Special syllables

The pronunciations of the above mentioned syllables are listed in Table 1(a). Finally, in RWTH-71 we use v-glide (instead of y-glide, as most Mandarin systems do) for the four syllables in Table 1(b).

The two language models (LMs) used in this work are kindly provided by UW and SRI. Both LMs share the same 60K vocabulary. The first 4-gram LM (LM.v1), used in all recognizers, is the same pruned LM as the one used in the GALE 2007 summer evaluation. The second 4-gram LM (LM.v2) is an improved version of LM.v1, using more data and almost no pruning. LM.v2 is used in lattice rescoring in the final system.

3. Acoustic Modeling

Similar to the systems presented in [7] and [1], the subsystems differ only in their acoustic front-ends, and the toneme set, and
the pronunciation dictionary, resp. The toneme set and the pronunciation dictionary are described in Section 2. The final system in the GALE 2007 re-evaluation consists of two subsystems labeled s1 and s2. While s1 is trained using RWTH-83, s2 is based on RWTH-71. The acoustic training is performed independently for each of the subsystems.

3.1. Acoustic Features

The acoustic front-ends of the subsystems consist of MFCCs or PLPs as base features. In addition, a voicedness feature [8] is augmented to the PLP feature extraction of s2 while s1 consists of MFCCs only. The features are normalized by segment-wise mean and variance normalization and are fed into a sliding window of length nine. All feature vectors within the sliding window are concatenated and projected to a 45 dimensional feature space using a linear discriminative analysis (LDA).

In addition, a tonal feature and its first and second derivatives, represented by the first and second order regression coefficients, are augmented to the LDA-transformed features. Tonal information is crucial for Mandarin ASR systems, because tonal patterns play an important role in distinguishing tones and words in the Mandarin language. The tonal feature used is described in [9].

For the experiments in Section 5.2, the setup is slightly different. Instead of augmenting the tonal features to the LDA-transformed features to, a common LDA for both feature streams is used, following [10]. In this case no derivatives of the tonal features are used.

Finally these features of s1 and s2 are concatenated with toneme posterior features produced by neural networks trained on a subset of 1200 hours of training data. S1 consists of hierarchical MRASTA (HMRASTA) features, produced by a hierarchical neural network [11] with multiple time resolution features (MRASTA) [5] as input. We use a hierarchy of three nets to produce the HMRASTA-features following [12]. The first and second net uses the higher and lower frequency parts of the MRASTA features, combined with plp features in the last net. At the end, the toneme posterior features are transformed by a logarithm and reduced by a principal component analysis (PCA) to 51 dimensions. Overall, concatenation of all features leads to a feature dimension of 99 for s1.

In contrast to s1, s2 uses neural network features based on Tandem [13] and hidden activation temporal patterns phoneme posterior (HATs) described in [2, 6]. Finally the Tandem and HATs features are combined using the Dempster-Shafer[14] algorithm, transformed by a logarithm and reduced by a PCA to 32 dimensions. Overall, s2 consists of 80 feature components.

3.2. Acoustic Training

The acoustic models for all systems are based on triphones with cross-word context, modeled by a 3-state left-to-right hidden Markov model (HMM). A decision tree based state tying is applied, resulting in a total of 4500 generalized triphone states. The acoustic models consist of 1M Gaussian distributions and a diagonal global covariance matrix. Both maximum likelihood (ML) and discriminative training are applied.

The filterbanks underlying the MFCC and PLP feature extraction undergo a vocal tract length normalization (VTLN). The warping factor classifier is trained beforehand on the complete training corpus.

In order to compensate for speaker variations we have used constrained maximum likelihood linear regression speaker adaptive training (SAT/CMLLR). While s1 uses the standard approach, for s2 a modified version of the SAT/CMLLR training is applied. The speaker adaptive training is combined with the LDA-transformation step, resulting in speaker-specific dimension reducing feature transformation matrices as introduced in [15]. In addition, during recognition, MLLR is applied to the means of the acoustic models.

Discriminative training, i.e. minimum phone error (MPE) [16] is applied to refine the ML trained acoustic models. For the MPE training of the two different systems we generate word-conditioned word lattices using the SAT/CMLLR model of s1 in combination with a bigram language model. System dependent alignments are produced for the accumulation and are kept fixed during the training iterations. The optimal number of training iterations is determined by recognition on the development corpus.

In order to save computation time, we convert the generated word graphs of s1 with the following procedure to the new toneme set of s2. First we map the pronunciation lemmas from s1 to all possible pronunciations provided by the pronunciation dictionary of s2. In order to cope with the different pronunciations in the word graph we introduce a weight factor for each possible pronunciation. During the procedure the word boundary times are kept fixed. Finally, the transformed word lattices are used for discriminative training for s2. As shown in the Section 6, the new word lattices work very well.

4. Corpora

1534h of broadcast news (BN) and broadcast conversations (BC) of speech data collected by LDC are used for training the final system. The corpus includes data from the Hub4 and TDT4 corpora and from the first three years of the GALE project (releases P1R1-4, P2R1-2, P3R1). Training transcripts are pre-processed by UW-SRI as described in [17].

However for the system development cycle, a 230h subset of the training corpus has been created. The subset contains the HUB4 corpus (30h), 100h of BN and 100h of BC from the four releases of the first year of the GALE project. Table 2 gives detailed statistics for the corpora used.

5. System Development

5.1. Development of the Toneme Set

In this section we present the results concerning the improvements introduced by RWTH-71 in contrast to the old toneme set RWTH-83.