TOWARDS AUTOMATIC CORPUS PREPARATION FOR A GERMAN BROADCAST NEWS TRANSCRIPTION SYSTEM

Wolfgang Macherey and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology, 52056 Aachen, Germany
{w.macherey, ney}@informatik.rwth-aachen.de

ABSTRACT

When setting up a speech recognition system for a new domain, a lot of manual effort is spent on corpus preparation, i.e., data acquisition, cutting and segmentation of the audio material, generation of pronunciation lexica, as well as the definition of suitable training and test sets. In this paper we describe several methods that help to automate and thus to speed up this procedure. For this purpose, we assume that only a preliminary, partially incorrect textual transcription is available. The effectivity of the proposed methods is demonstrated with the development of a transcription system for the recognition of German broadcast news.

1. INTRODUCTION

Fast and low-cost system development plays an important role if an existing speech recognizer is adapted to a new language or domain. Usually a lot of manual work is necessary to collect and segment the audio data, to transcribe the resulting segments, to create new pronunciation lexica, to define suitable training and test sets, and to estimate proper acoustic models and language models. However, the amount of required personnel resources could drastically be reduced if the process of collecting and preprocessing was automated. In the recent past, several research groups investigated the capabilities of unsupervised training procedures to rapidly prototype a speech recognition system for new languages and domains [1, 2]. These approaches have in common that at least a small portion of manually transcribed data as well as an initial alignment between the feature vectors and the Hidden Markov Model (HMM) states must be provided in order to bootstrap the training.

This paper focuses on a different scenario. We assume that low-quality transcriptions are provided for parts of large unsegmented audio files, however, without any time alignment with the audio data. Moreover, we proceed on the assumption that no domain specific speech recognizer is available. Now, given both the audio and the text sources, the objective is to automatically compile a speech corpus that enables both a training of the acoustic models and a reliable evaluation of the achieved recognition performance using as little manual effort as possible. To this purpose, we employ a set of new tools that help to automate the process of collecting and preprocessing the training and the test data. The effectivity of the proposed methods is demonstrated with the domain change of the RWTH large vocabulary continuous speech recognition system from the *Verbmobil* time scheduling task [3] to the automatic transcription of German broadcast news.

2. DATA ACQUISITION

Nowadays, more and more radio and TV stations provide parts of their program on the Internet. Particularly for broadcast news and TV reports one can often find not only the recordings of the broadcasts but also the respective scripts. However, since the audio and the text data usually comprise complete shows, several preprocessing steps have to be carried out in order to transform the raw data into a format that is more appropriate for a subsequent training or recognition. Moreover, depending on the format, the audio data is provided as, it might be necessary to detect and eliminate transfer artifacts such as zero-byte sequences and multiple repetitions of frames which both may occur in real audio streams because of high server loads.

For the development of a German broadcast news transcription system, three different shows of the two broadcasting corporations "Westdeutscher Rundfunk" (WDR) and "Südwestrundfunk" (SWR) were collected. Among these are two radio broadcast news shows that are transmitted daily every hour as well as the TV magazine *Report* which is broadcast twice a month. The audio material of the first two programs is provided as a real audio stream on the Internet and was collected over a period of 10 months. The latter was provided as MPEG-1 encoded recording including nearly all shows that were broadcast between January 1997 and January 2001. In Table 1 some information on data collection is summarized.

2.1. Elimination of Transfer Artifacts in Audio Streams

Since parts of the audio data are provided as real audio stream, the respective recordings often contain transmission artifacts like zero-byte sequences and multiple repetitions of frames (cf. Figure 1). However, unlike locating zero-byte sequences the detection and elimination of frame repetitions is more complicated since neither the frame length nor the number of repetitions is known in advance. Nevertheless, the problem can be solved very efficiently in the following way: in a first step each sample is linked with its successor and after that, the whole sample sequence in sorted in ascending order. Because of the preceding linking and sorting stage it is now easy to determine all sample subsequences with identical prefix and to check whether the length of a prefix that occurs several times, exceeds a certain value. If so, the prefix is part of a frame repetition and all occurrences but one must be deleted. Afterwards, the original relative order of the residual samples is restored. After performing these preprocessing steps, both the real audio stream and the MPEG encoded audio data were re-sampled at 16kHz, resulting in the final audio format.

Table 1. *Corpus statistics for the German broadcast news transcription system.*

	WDR	SWR	
show	radio news	radio news	*Report*
period	Mar 2000 – Dec 2000	Jul 2000 – Dec 2000	Jan 1997 – Jan 2001
# broadcasts	ca. 2800	ca. 1600	48
bandwidth	11.025kHz	11.025kHz	22.05kHz
audio format	real audio	real audio	MPEG-1
speech	read	read	spontan.
ave. duration	5-7min	5-7min	42min

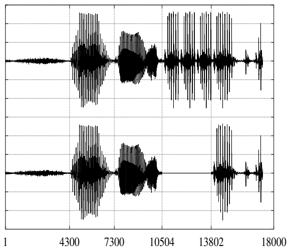

Fig. 1. *Audio stream before (above) and after (below) elimination of frame repetitions.*

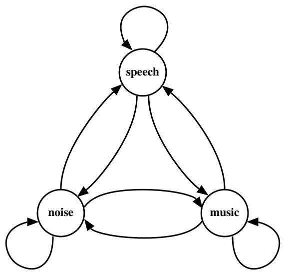

Fig. 2. *Ergodic Markov topology for speech-noise-music segmentation.*

3. GENERATING PRONUNCIATION LEXICA

The recorded broadcast news data comprise a vocabulary of approximately 41,000 words. The respective phonetic transcriptions were generated using an automated tool which integrates several machine readable pronunciation dictionaries, among them the *Verbmobil* lexicon, Lex-DB, Celex, the pronunciation lexicon of the *Süddeutsche Zeitung*, and a statistical grapheme-to-phoneme translator. Compound words that are not covered by the lexicon entries were composed from their respective constituents. The resulting pronunciation lexicon contained 61,700 entries, thus including approximately 22,000 pronunciation variants (35.5%).

4. AUDIO STREAM SEGMENTATION

When applying automatic speech recognition to broadcast news data, a preliminary segmentation step is required. The goal of this segmentation stage is on the one hand to partition the whole audio stream into speech and non-speech segments and on the other hand to cut the speech segments into reasonably short segments while discarding the non-speech portion. Moreover, for the automatic preparation of a training corpus a sentence-wise cutting is useful as well as a technique that allows a reliable verification of the textual transcriptions. As will be shown in the next sections, both requirements can be considered as specific decoding tasks.

4.1. A Discriminatively Trained Segmenter

Discriminative training criteria such as Maximum Mutual Information (MMI) and Minimum Classification Error (MCE), often outperform maximum likelihood (ML) based training methods, especially if the model complexity is low [4]. So far, discriminative training techniques were used to optimize speech recognition systems at the model level [5, 6]. Nevertheless, due to their general formulation they can also be used for a segmentation task. In this work, the MMI criterion was applied to train a segmenter using a generalized probabilistic gradient descent method for parameter optimization.

For this, the training data shall be given by $r = 1, \ldots, R$ training patterns, each consisting of a sequence X_r of acoustic observation vectors $x_{r1}, x_{r2}, \ldots, x_{rT_r}$ and the corresponding sequence k_r of class labels $\kappa_{r1}, \kappa_{r2}, \ldots, \kappa_{rN_r}$ where κ denotes a segmentation class. The *a posteriori* probability for a segmentation k_r given the acoustic observation vectors X_r shall be denoted by $p_\lambda(k_r \mid X_r)$ where λ denotes the set of all distribution parameters. Similarly, $p_\lambda(X_r \mid k_r)$ represents the emission probability for the acoustic observation sequence X_r. The priors $p(k_r)$ are supposed to be given. Then the MMI criterion which is defined as the sum over the logarithms of the *a posteriori* probabilities of each training pattern, is given by:

$$F(\lambda) = \sum_{r=1}^{R} \log \frac{p(k_r) \cdot p_\lambda(X_r \mid k_r)}{\sum_{\widetilde{k}} p(\widetilde{k}) \cdot p_\lambda(X_r \mid \widetilde{k})} \tag{1}$$

To segment an audio stream we use the Markov topology as depicted in Figure 2. Since each state of this Markov network is associated with one of the three segmentation classes *noise*, *speech*, and *music*, respectively, we can identify the class labels with the corresponding HMM states, i.e.

$$\kappa \equiv s_\kappa \qquad \text{with} \;\; \kappa \in \{\textit{noise}, \textit{speech}, \textit{music}\} \tag{2}$$

In the following a segmentation of the acoustic observations X_r will be represented as the according sequence of hidden Markov states $s_{\kappa_1}, \ldots, s_{\kappa_{T_r}}$. The emission probability of an observation x given the HMM state s is modeled as a Gaussian mixture density with the distribution parameters $\lambda_s = \{c_{sl}, \lambda_{sl} \mid l = 1, ..., L\}$ where c_{sl} denotes the mixture weight of the component density l and $\lambda_{sl} = \{\mu_{sl}, \sigma_{sl}\}$ denotes a density specific mean μ_{sl} and a diagonal covariance matrix σ_{sl}, respectively. The objective function in Eq. (1) can be maximized using a gradient descent method [5]:

$$\overline{\lambda} = \lambda + \epsilon \cdot \frac{\partial F(\lambda)}{\partial \lambda} \tag{3}$$

The derivation of the MMI criterion with respect to λ_{sl} yields

$$\frac{\partial F(\lambda)}{\partial \lambda_{sl}} = \Gamma_{sl}\left(\frac{\partial \log p(x \mid \lambda_{sl})}{\partial \lambda_{sl}}\right) \tag{4}$$

where Γ_{sl} denotes the discriminative average for the mixture component l of state s which is defined by:

$$\Gamma_{sl}\big(g(X)\big) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \Big(\gamma_{rt}(s \mid k_r) \, - \, \gamma_{rt}(s)\Big) \cdot g(x_{rt})$$

Using the Viterbi approximation the forward-backward (FB) probabilities γ_{rt} are given by

$$\gamma_{rt}(s \mid k_r) := p_\lambda(s_t = s \mid X_r, k_r) \underset{\text{Viterbi}}{=} \delta(s_{k_r(t)}, s) \tag{5}$$

$$\gamma_{rt}(s) := \sum_{\widetilde{k}} p_\lambda(\widetilde{k} \mid X_r) \cdot \gamma_{rt}(s \mid \widetilde{k}) \tag{6}$$

Here, δ denotes the Kronecker delta and $s_{k_r(t)}$ is the HMM state of the best alignment path at time frame t. The generalized FB probabilities $\gamma_{rt}(s)$ in Eq. (6) can be estimated using a forward-

Table 2. *Comparison of two segmentation approaches on the test set of the Report corpus.*

approach	speech lost [%]	purity [%]	avg. segment length	words cut [%]
ML	1.58	97.57	21.20s	0.17
MMI	0.89	98.84	26.30s	0.12

backward algorithm on word graphs [7]. Using state specific variances σ_s we obtain the following reestimation equations:

$$\widehat{\mu}_{sl} = \frac{\Gamma_{sl}(x) + D_s c_{sl}\mu_{sl}}{\Gamma_{sl}(1) + D_s c_{sl}}$$

$$\widehat{\sigma}_s^2 = \frac{\Gamma_s(x^2) + D_s\sigma_s^2 + \sum_l \left[c_{sl}\mu_{sl}^2 - \left(\Gamma_{sl}(1) + D_s c_{sl}\right)\widehat{\mu}_{sl}^2\right]}{\Gamma_s(1) + D_s}$$

$$\widehat{c}_{sl} = \frac{\dfrac{\Gamma_{sl}^{\text{cor}}(1)}{\Gamma_s^{\text{cor}}(1)} - \dfrac{\Gamma_{sl}^{\text{gen}}(1)}{\Gamma_s^{\text{gen}}(1)} + C_s}{\sum_{\tilde{l}} c_{s\tilde{l}}\left[\dfrac{\Gamma_{s\tilde{l}}^{\text{cor}}(1)}{\Gamma_s^{\text{cor}}(1)} - \dfrac{\Gamma_{s\tilde{l}}^{\text{gen}}(1)}{\Gamma_s^{\text{gen}}(1)}\right] + C_s} c_{sl}$$

Here, D_s is a state specific iteration constant which controls the convergence of the training process. The choice of D_s has to ensure that all denominators in the respective reestimation equations are positive which can be achieved by setting

$$D_s = h \cdot \max\left\{D_s^{\min},\ 1/\beta - \Gamma_s(1)\right\}$$

where $D_s^{\min}$ is a constant which guarantees that all variances are positive definite. $\beta > 0$ is chosen to prevent overflows caused by low-valued denominators and $h > 1$ is a scaling factor that controls the step size of the gradient descent method. The iteration constants C_s enable a fast convergence of the mixture weights. For this, the discriminative averages Γ_{sl} are decomposed into the correct model Γ'_{sl} and the competitive model Γ''_{sl}:

$$\Gamma'_{sl}\big(g(X)\big) := \sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_{rt}(s \mid k_r) \cdot g(x_{rt}) \tag{7}$$

$$\Gamma''_{sl}\big(g(X)\big) := \sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_{rt}(s) \cdot g(x_{rt}) \tag{8}$$

$$C_s = \max_l \left\{\frac{\Gamma''_{sl}(1)}{\Gamma''_s(1)} - \frac{\Gamma'_{sl}(1)}{\Gamma'_s(1)},\ 0\right\} + \varepsilon, \quad 0 < \varepsilon \tag{9}$$

Comparative experiments using the discriminatively trained segmentation algorithm and a ML based segmenter were performed on the test set of the *Report* data. Both segmenters use 16 cepstral coefficients with the 16 first derivatives and the second derivative of the energy. Table 2 shows the segmentation results. Compared to the ML based training the discriminative approach improved the segmentation purity from 97.57% to 98.84%. The amount of lost speech could be reduced from 1.58% to 0.89%.

4.2. Networks of Hidden Markov Models

Even though some broadcasting companies provide transcriptions of some shows on the Internet, the quality is often insufficient for training purposes. Particularly for broadcast news there are often major changes between the available transcriptions and the respective recordings. Missing transcriptions on the one hand and transcribed sentences that were not spoken on the other hand are counted among the most frequently occurring errors.

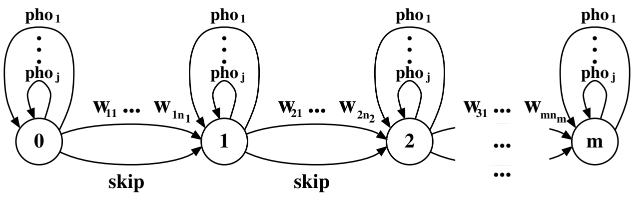

Fig. 3. *Network of HMMs for transcription verification and sentence-wise segmentation.*

To mitigate the effect of transcription errors in the training data, the segmentation procedure was extended to automatically verify the given transcriptions and to discard incorrect parts. The method is based on an algorithm that aligns a sequence of acoustic observations to a network of HMMs. This network is a final state transducer which allows a more flexible alignment of the acoustic observations. Besides the word automata for the original transcription, the HMM network also contains monophone transitions which serve as filler models and optionally pronunciation variants and noise models (cf. Figure 3). A sequence of word labels $w_{i1}, \dots, w_{in_i}$ as depicted in Figure 3 represents the respective Markov automaton for the ith sentence of a given transcription. Sentences that were not spoken are modeled by skip transitions. Untranscribed utterances are modeled by monophone arcs. During the segmentation process all acoustic observations that are aligned to filler models (i.e. monophone arcs or noise models) are considered as incorrectly transcribed data and will be left out in a subsequent training of the acoustic models. The determination of the best path through the network yields both a more reliable transcription and an alignment path which can be used to bootstrap the training. In order to prevent the alignment of the acoustic observations to monophone arcs only, the respective HMM state hypotheses are penalized by adding an offset to the acoustic score. Based on the resulting alignment path, the audio data was cut sentence-wise. Finally, all training sentences were discarded whose acoustic score (normalized to the number of time frames) exceeded a certain threshold. Thus, the HMM network has the advantage to focus the search on a given, perhaps flawed transcription while preserving the flexibility to align untranscribed data. Using the HMM networks, an automatic transcription verification in combination with a sentence-wise cutting was even feasible for shows with runtimes of more than 45 minutes.

5. ACOUSTIC AND LANGUAGE MODEL ESTIMATION

Given the new (verified) transcriptions and the alignments as generated by the HMM networks, a ML training for the estimation of monophone models was performed. After that, a linear discriminant analysis (LDA) matrix was estimated. The resulting alignment paths were used to train a phonetic classification and regression tree (CART) followed by a reestimation of the LDA.

For training the language model we used a newspaper text collection that was enriched with reports of a news ticker. The resulting corpus comprises 61.3 million running words.

6. EXPERIMENTAL RESULTS

Experimental results were performed with the RWTH *Verbmobil* recognizer [3]. The recognizer uses a time-synchronous beam search algorithm based on the concept of word-dependent tree co-

Table 3. *Recognition results on the WDR/SWR broadcast news corpus and the Report corpus.*

	WDR				SWR	*Report*	
	train	test1	test2	test3	test	train	test
period	Mar 2000	Apr 2000	May 2000	Jun 2000	Aug 2000	1997-2000	Mar 2, 1998 Jan 29, 2001
duration [h]	13.3	18.8	24.1	21.8	1.1	11.2	1.6
# sentences	6,947	10,130	13,138	11,793	731	8,928	1,054
# running words	107,579	150,447	194,163	173,115	7932	100,682	12,115
perplexity (trigram LM)	39.17	38.30	38.80	178.35	150.89	23.01	110.78
silence portion [%]	18	19.3	18.7	19.2	22.2	17	16.8
OOV rate [%]	0.0	0.0	0.0	0.0	5.17	0.0	5.2
word error rate [%]	**6.57**	**6.50**	**7.67**	**9.59**	**16.50**	**11.1**	**35.1**
real time factor	5.3	8.5	9.5	4.0	5.4	7.4	8.9

pies and integrates the trigram language-model constraints in a single pass. The pruning strategy includes a look-ahead technique of the language model probabilities [8]. For all experiments, gender independent models were used. No speaker-adaptive or normalization methods were applied. Similar to the segmentation algorithm 16 cepstral coefficients with the 16 first derivatives and the second derivative of the energy were used.

Recognition results were produced in the following way: in a first experiment, HMM networks as described in section 4.2 were used to verify and align the transcriptions of the WDR/SWR radio broadcast news. Since the radio data contained only read speech, a previous segmentation stage was not necessary. The initial alignment and a subsequent training were performed using 15 hours of speech of the WDR data that was broadcast during the second half of March 2000. Monophone references trained with the *Verbmobil* system [3] served as initial models for the network-based alignment. The recordings of the SWR radio broadcasts were used as test set. To prevent speaker overlaps the SWR data was excluded from the training set. Table 3 shows the recognition results on the WDR training and the SWR test set. The WDR test sets 1-3 show the evolution of the word error rate over three months. For all recognition results the text sources used to estimate the language model comprised news paper texts enriched with task specific data collected until March 2000 inclusive.

In a second experiment, HMM networks and monophone models trained with the WDR corpus were used to verify the *Report* transcriptions and to produce initial alignments. First of all, a segmentation step was applied in order to separate the speech signal from the non-speech portion (cf. section 4.1 for segmentation results). Thereby, approximately two-thirds of the audio data was automatically discarded due to errors in the given transcriptions. Table 3 shows the word error rates obtained on the *Report* training corpus and test corpus, respectively. The durations reported refer to the segmented speech signal. Real time factors were determined on a Pentium-III processor (450MHz).

7. CONCLUSIONS

In this paper we presented several methods that enable almost a fully automated speech corpus acquisition and preparation. Among these approaches, a discriminatively trained segmentation algorithm and a new training procedure based on networks of Hidden Markov Models (HMM) was presented. The proposed methods were tested on the RWTH large vocabulary continuous speech recognizer which had originally been developed for the *Verbmobil* time scheduling task [3] and was now adapted to the recognition of German broadcast news. Starting from low-quality transcriptions, the HMM networks enabled a transcription verification as well as a sentence-wise cutting of the training data. Experiments were performed on the WDR/SWR radio broadcast news shows and the SWR *Report* corpus.

Acknowledgments: This work was partially funded by the European Commission under the Human Language Technologies Project CORETEX (IST-1999-11876). Additionally, we would also like to thank SWR for kindly providing the *Report* data.

8. REFERENCES

[1] L. Lamel, J. L. Gauvain, and G. Adda, "Lightly supervised acoustic model training," in *Automatic Speech Recognition Workshop*, Paris, France, Sep. 2000, pp. 150–154.

[2] F. Wessel and H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," in *Automatic Speech Recognition and Understanding Workshop*, Trento, Italy, Dec. 2001.

[3] S. Kanthak, A. Sixtus, S. Molau, R. Schlüter, and H. Ney, "Fast search for large vocabulary speech recognition," in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed., pp. 63–78. Springer, 2000.

[4] R. Schlüter, W. Macherey, S. Kanthak, H. Ney, and L. Welling, "Comparison of optimization methods for discriminative training criteria," in *ESCA 5th European Conference On Speech Communication And Technology*, Rhodes, Greece, 1997, vol. 1, pp. 15–18.

[5] R. Schlüter and W. Macherey, "Comparison of discriminative training criteria," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1998, vol. 1, pp. 493–496.

[6] P. Beyerlein, "Discriminative model combination," in *Automatic Speech Recognition and Understanding Workshop*, Santa Barbara, Dec. 1997, pp. 238–245.

[7] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 9, pp. 288–298, March 2001.

[8] S. Ortmanns, A. Eiden, and H. Ney, "Improved lexical tree search for large vocabulary speech recognition," in *Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, USA, May 1998, vol. 2, pp. 817–820.