RECOGNITION OF CONVERSATIONAL TELEPHONE SPEECH USING THE JANUS SPEECH ENGINE

Torsten Zeppenfeld *Michael Finke* *Klaus Ries* *Martin Westphal* *Alex Waibel*

Interactive Systems Laboratories
Carnegie Mellon University, USA
University of Karlsruhe, Germany

ABSTRACT

Recognition of conversational speech is one of the most challenging speech recognition tasks to-date. While recognition error rates of 10% or lower can now be reached on speech dictation tasks over vocabularies in excess of 60,000 words, recognition of conversational speech has persistently resisted most attempts at improvements by way of the proven techniques to date. Difficulties arise from shorter words, telephone channel degradation, and highly disfluent and coarticulated speech. In this paper, we describe the application, adaptation, and performance evaluation of our JANUS speech recognition engine to the Switchboard conversational speech recognition task. Through a number of algorithmic improvements, we have been able to reduce error rates from more than 50% word error to 38%, measured on the official 1996 NIST evaluation test set. Improvements include vocal tract length normalization, polyphonic modeling, label boosting, speaker adaptation with and without confidence measures, and speaking mode dependent pronunciation modeling.

1. INTRODUCTION

The recognition of conversational speech over telephone lines such as the Switchboard LVCSR corpus represents one of the most challenging speech recognition tasks to date. The Switchboard corpus and conversational speech in general have persistently resisted attempts to improve results to the level of read speech. After several years of intense research by a number of large research teams, error rates on conversational telephone speech (the Switchboard corpus) have been improved considerably from an initial 70+% word error, but still remain stubbornly high. Official test results in 1995 still averaged 52% word error across participating sites. Difficulties arise from phenomena found mainly in this type of speech, such as the usage of shorter words and the significant presence of highly disfluent and coarticulated speech. Acoustic degradations from the telephone channel, such as cross-talk, clicks, channel noise and spikes, also have a negative effect on performance.

In the following, we describe our work on developing a speech recognition system for the Switchboard Large Vocabulary Conversational Speech Recognition (LVCSR) task. In addition to giving an overview of our system, we will highlight several noteworthy enhancements. These include vocal tract length normalization, polyphonic modeling, label boosting, speaker adaptation with and without confidence measures, and speaking mode dependent pronunciation modeling.

2. SYSTEM OVERVIEW

2.1. Preprocessing

During the pre-processing stage, we perform several operations which make it easier for the recognizer to do its job. First, an adaptive crosstalk filter is used to eliminate much of the channel crosstalk present in the 4-wire setup of the Switchboard recordings.

We want to remove crosstalk in signal $a(t)$ caused by the speech signal $s_b(t)$ of speaker B with

$$a(t) = s_a(t) + x(t) * s_b(t)$$

where $s_a(t)$ is the speech signal of speaker A. The linear model of the crosstalk path $x(t)$ can be estimated by using a FIR filter $h(t)$ and the LMS adaptation algorithm. We allow adaptation only when the power of both channels indicate that significant crosstalk is present (see figure 1).

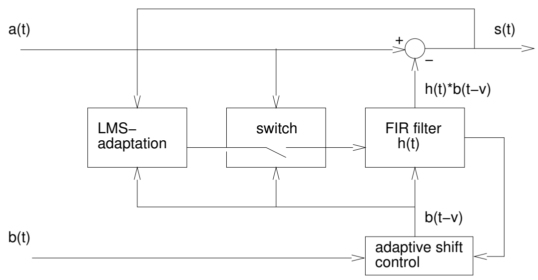

Figure 1. Adaptive Crosstalk filter

Due to recording conditions of the Switchboard data the two channels are not always synchronous and the time shift happens to drift up to a second over a conversation. For this reason we applied an adaptive shift control that moves the input window of the filter according to the correlation of the two channels which is implicitly encoded in the coefficients of the filter. Given an FIR filter with the impulse response $\{h[t] \mid t = -L \ldots L\}$ we define the mass center of the filter as $m = \sum_{i=-L}^{L} i \cdot |h_i|$. We then control the shift v by changing it by $\Delta v = \alpha(\frac{m}{L})$. The constant α determines the adaptation rate of the shift control.

After the crosstalk is filtered out, the speech signal is passed through a silence detector in order to segment the conversation into reasonably sized chunks and filter out long sections of silence. Using the channel signal power as its input feature, a continuous density silence/speech classifier is trained, and proceeds to label the input signal as speech or silence. Recent tests have shown that we loose roughly 2% absolute from this automatic segmentation procedure when compared against an optimal segmentation which is privy to the true word boundary information.

Speakers come in all shapes and sizes, and so does their speech. We use a maximum likelihood based vocal tract length normalization algorithm in order to remove some of the variation due to speakers' differing vocal tract characteristics. A non-linear warping [3] in the frequency domain is done based on the second formant according to:

$$\begin{aligned} \hat{f} &= fK^{\frac{3f}{2F}} \\ K &= F_2/\bar{F}_2 \end{aligned}$$

where K is the warping factor for a particular speaker with second formant F_2, F is the Nyquist frequency, and $\bar{F}_2$ is the average second formant for all speakers.

In order to find F_2 for a speaker, an initial estimate is made by using a formant tracker. A phoneme recognizer is then used to calculate the likelihood of the signal for several different warping factors around this initial estimate. In this way the maximum likelihood estimation of F_2 is determined. We observe a performance gain of almost 2% absolute by using this ML-VTLN technique over simply processing the F_2 that the formant tracker provides.

The final preprocessing step is the calculation of the input features for our speech segments. These are derived by using a Linear Discriminant Analysis (LDA) transformation over a 9 frame window of Perceptual Linear Prediction (PLP) coefficients. The LDA transformation has the dual benefit of reducing the feature space from 117 dimensions to 48, and of optimally separating the phonetic classes.

2.2. Acoustic Modeling using Polyphones

Context-dependent acoustic models have been applied in speech recognition research for many years, and have been shown to increase the recognition accuracy significantly. The most common approach is to use triphones. Recently, several speech recognition groups have started investigating the use of larger phonetic context windows when building acoustic models [1, 6]. We also make use of a larger context in our recognizer by allowing questions in the allophonic decision tree not only referring to the immediate neighboring phones but also to phones further away (for Switchboard we used a context of two instead of the context of one as in the triphone setup).

In a two stage decision tree based clustering approach the codebooks are clustered first and, based on the clustered codebooks, in a second step the distributions are clustered. For Switchboard we ended up having 4000 codebooks and 20000 distributions. This clustering approach implements a flexible parameter tying scheme, and gave us significant improvement across many tasks, including WSJ, Switchboard, and the Spontaneous Scheduling Task. It has also proved itself across several languages (German, Spanish, English) [4]. For Switchboard, we have observed a WER reduction of 2.4% absolute.

2.3. Language Modeling

The Switchboard corpus contains approximately 2 million words of training text. Typically only about 60% of the trigrams in the test text were actually seen in the training text. Smoothing of the trigram models was therefore seen an important possible source for LM improvement. We have implemented class based models and linear interpolation algorithms to make maximum use of the Switchboard data that we have, and to integrate models being built on the NAB (North American Business News) corpus. Our evaluation language model uses a linear interpolation of 4 trigram backoff models: one standard Switchboard trigram model, one standard NAB model and two class based models built on the Switchboard corpus. The classes for the class based models were derived from NAB and Switchboard respectively. Context dependent linear interpolation did not show a significant improvement compared to the use of context independent interpolation. Using these techniques we achieved a WER reduction of 1% absolute over the standard Switchboard trigram model.

buying	sailing	Friday	mainly
adding	bowling	Monday	mostly
burning	camping	Saturday	partly
owning	dancing	Sunday	primarily
renting	setting	Sundays	purely

Table 1. SWB class based LM: sample classes

The word classes for our class-based models were built using a procedure that optimizes the bigram perplexity criterion [5]. Table 1 shows examples of some of the automatically generated classes for the Switchboard model. In order to derive effective word classes, we classified only words that have more than a minimum number of counts and introduced a prior on the number of classes. This enables us to tune the number of classes and run the class clustering procedure in reasonable time.

In the Switchboard corpus, silences of various durations are interspersed with speech (for instance, when a speaker listens to what the other speaker is saying). For language modeling purposes, we have found that the exact treatment of silence can make a significant difference in a system's performance. In a speech recognizer, short silence is usually modeled as an "optional silence" that can be inserted at any point with a context-independent probability that does not modify the context of the language model. Recent experiments have shown that treating some silences as regular LM tokens yields a WER improvement of approximately 1% absolute. Since the NAB database is not annotated with silence, we used a mapping that changed several punctuation markers to silence tokens. This allowed us to train an interpolated language model that included the silence word token.

In addition to the above, we have investigated both selective unigram cache models and maximum entropy trigger models. Even though we have achieved significant perplexity reductions with some of these techniques, they did not

experiment	WER
no adaptation baseline	38%
adapt on hypothesis	37%
adapt on Correct only	35%
adapt on Transcription	31%

Table 2. Confidence Measure Performance

reduce the word error, and so were not applied in the recent evaluation.

3. LABEL BOOSTING

Several stages of our training algorithm (LDA, Kmeans, BW) use a Viterbi search to find the best path through a training utterance. For reasons of speed, we currently generate Viterbi path labels only once for each utterance in our training set, and run the different stages of training using these labels. The accuracy of our acoustic models thus depends heavily on the accuracy of these labels. The MLLR speaker adaptation algorithm (described below) can be used to adapt the acoustic models to each speaker in our training set, and thus we can effectively generate labels with the equivalent of a speaker dependent recognizer. We have noted consistent improvements of 1-2% using this technique. This is not surprising, when we note the tremendous performance gain that adaptation can bring to the system when used with known transcriptions (table 2).

4. ACOUSTIC STABILITY CONFIDENCE MEASURE

Similar to the N-Best confidence measure described in [2], the idea behind our acoustic stability confidence measure algorithm is that we expect regions of high acoustic stability to be regions that are relatively error free, and regions of low acoustic stability to be regions that will frequently contain recognizer errors. We can isolate regions of stability for a single utterance by comparing the hypothesis of our recognizer over several different language model weights and word penalties. This in effect is a way of adding LM noise to the recognizer. The less stable words in the hypothesis will tend to change with the minimal addition of this noise. We calculate the confidence of a specific word, given a list of hypotheses with varying LM weights and penalties, as the ratio of the number of hypotheses in which the word occurs to the total number of hypotheses.

One advantage of our confidence measure over the N-Best measure is that for a very stable utterance, all our hypotheses could potentially have the same word string, and thus the confidence of the words in this hypothesis would be very high. The confidence of the words using the N-Best measure is limited, since some word must change in each hypothesis of the N-Best list.

Preliminary results show that this confidence measure technique classifies words correctly (errors as errors, and correct words as correct words) with an accuracy of approximately 70%. It has also proven useful during our unsupervised adaptation procedure, as described below.

5. MLLR UNSUPERVISED SPEAKER ADAPTATION

Although the use of our ML-VTLN algorithm helps in reducing the variance of speakers' voice characteristics, it doesn't solve the problem alone. In VTLN, we try to normalize a speaker's speech signal by stretching or compressing along the frequency axis, which roughly corresponds to changing one parameter: the vocal tract length. But many aspects of the speech signal are not normalized by this simple approach. For this reason, a form of unsupervised adaptation is used in our evaluation system. It has the advantage of performing an arbitrary linear transformation on the acoustic models.

We use a maximum likelihood linear regression (MLLR) unsupervised speaker adaptation algorithm [7] to adapt our acoustic models to specific speakers during testing. Given a set of recognition hypotheses for a speaker's utterances, the algorithm transforms the acoustic models in order to maximize the likelihood of these hypotheses. The actual number of transformations performed is determined automatically based on how much adaptation data is available by the model clustering stage of the algorithm.

This model clustering algorithm combines all Gaussians from our acoustic models into one cluster. This cluster is then split along the axis of highest variance into two clusters. These new clusters are then also split, and the procedure is iterated until the amount of training data for each Gaussian cluster reaches a minimum threshold. In order to find the number of transformations for each test speaker, we prune this cluster tree until we have a minimum number of samples in each leaf for the test speaker. We then use the MLLR algorithm to find a transformation for each of these model clusters. This automatic clustering algorithm has the principle advantage in that we do not have to specify the number of transformations that we want to perform during adaptation. This will be selected automatically based on the amount of adaptation data available.

These adapted models can then be used in a new recognition pass, thus providing better hypotheses. This procedure could in principle be iterated several times, each time tuning the models based on the new recognition hypotheses. In practice, the performance asymptotes quickly. For the Switchboard evaluation, three recognition passes where performed, including two adaptation steps. After one iteration, a WER gain of 1.4% absolute was achieved. An aditional iteration of adaptation, yielded only another 0.2%. Current results show WER improvements of 2.6% absolute over recognition without adaptation.

The fewer errors there are in a hypotheses, the better the adaptation algorithm can adapt to the given speaker. This idea is confirmed by the results shown in table 2, in which we see a large decrease in word error when the adaptation algorithm is given the correct transcription on which to adapt itself, instead of the hypothesis string. Another interesting experiment shows that if we can filter out the errors of a hypothesis such that we don't adapt on them, we again get a substantial performance increase. This is also shown in table 2, where we adapt our recognizer on correct parts of the hypothesis only.

These results naturally lead us to the use of confidence

1	[AX IX] N $\rightarrow$ EN
2	[AX IX] M $\rightarrow$ EM
3	[AX IX] L $\rightarrow$ EL
4	[AX IX] R $\rightarrow$ AXR
5	[T D] $\rightarrow$ DX / [+VOWEL] __ [AX IX AXR]
6	[T D] R $\rightarrow$ DX
7	L $\rightarrow$ 0 / __ Y [AX IX AXR]
8	[T D] $\rightarrow$ 0 / [+VOWEL] __ [TH DH]
9	[T D] $\rightarrow$ 0 / [+CONS +CONTINUANT] __ WB
10	R AX $\rightarrow$ ER / [-WB] __ [-WB]

Table 3. Sample of Variant Pronunciation Rules

measures as a way of filtering out the errorful parts of a recognizer hypothesis. Using our confidence measure to aid the unsupervised adaptation algorithm improves the recognizer by 1.4% absolute compared to using adaptation with no confidence measure. We feel that further improvements in the confidence measure is one of the most fruitful areas of research for improving our recognition rates in the near future.

6. SPEAKING MODE DEPENDENT PRONUNCIATION MODELING

In spontaneous conversational speech there is a large amount of variability due to accents, speaking styles and speaking rates (also known as the speaking mode) [8]. Because current recognition systems usually use only a relatively small number of pronunciation variants for the words in their dictionaries, the amount of variability that can be modeled is limited. Increasing the number of variants per dictionary entry is the obvious solution. Unfortunately, this also means increasing the confusability between the dictionary entries, and thus often leads to an actual performance decrease. We believe that the probability of encountering a particular pronunciation variant is a function of a speaker's speaking mode, and thus cannot be modeled adequately using static word variant probabilities.

We expand our recognition dictionary by applying a set of phonological rules in order to generate a variety of pronunciation variants. A sample of these rules is given in table 3. By the use of these rules, our dictionary grew to have an average of 1.8 variants per base entry. Based on this expanded dictionary, we perform a forced alignment pass through our training data. During this pass, we extract training data for each of the rules, by noting the speaking mode indicators associated with each rule. The speaking mode indicators include features such as measures of the speaking rate, word durations, and the fundamental frequency of the speech. These features are then used to train a set of decision trees, one tree for each rule. These trees are used to predict the mode dependent pronunciation rule probabilities.

We have implemented this technique as part of our lattice rescoring pass. Based on these trees, the words in the lattice, and the speaking mode indicators associated with these words, we can generate dynamic pronunciation probabilities for the various word variants. Preliminary results indicate a WER decrease of 1.7% absolute even using a very restricted set of indicators.

experiment	init WE	end WE	% change
Polyphonic System	46.0%	43.6%	5.2%
VTLN	46.0%	43.9%	4.5%
Label Boosting	43.6%	42.4%	2.7%
MLLR Adapt. + CM	43.4%	40.8%	6.0%
ML-VTLN	40.2%	38.4%	4.5%
SWB LM + sil	38.4%	37.4%	2.6%
Mode dep. Rules	39.0%	37.6%	3.5%

Table 4. Performance Gain Summary

7. CONCLUSION

A variety of significant enhancements to the Janus speech engine have reduced the error rate on the Switchboard LVCSR task from over 50% to 38.4%. A summary of the enhancements, together with their approximate respective improvements is shown in table 4. Note that several of these numbers come from results on the Switchboard development test set, and several from the evaluation test set.

REFERENCES

[1] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahmoo, and M.A. Picheny. Decision Trees for Phonological Rules in Continuous Speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, 1991. IEEE.

[2] F. Beaufays, Y. Konig, Z. Rivlin, A. Stolcke, and M. Weintraub. Neural Network Based Measures of Confidence. In *Proceedings of LVCSR Hub 5 Workshop*, April 1996.

[3] Ellen Eide and Herbert Gish. A Parametric Approach to Vocal Tract Length Normalization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, 1996. IEEE.

[4] Michael Finke and Ivica Rogina. Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997. IEEE.

[5] Reinhard Kneser and Herman Ney. Improved clustering techniques for class-based statistical language modeling. In *Eurospeech*, Berlin, Germany, 1993.

[6] R. Kuhn, A. Lazadrides, Y. Normandin, and J. Brousseau. Improved Decision Trees for Phonetic Modeling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 552–555, Detroit, Michigan, 1995. IEEE.

[7] L. Neumeyer, A. Sankar, and V. Digalakis. A Comparative Study of Speaker Adaptation Techniques. In *Eurospeech*, 1995.

[8] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A.Waibel, , B. Wheatley, and T. Zeppenfeld. Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode. In *International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.

RECOGNITION OF CONVERSATIONAL TELEPHONE SPEECH USING THE JANUS SPEECH ENGINE

Torsten Zeppenfeld , Michael Finke , Klaus Ries , Martin Westphal and Alex Waibel

Interactive Systems Laboratories

Carnegie Mellon University, USA

University of Karlsruhe, Germany

Recognition of conversational speech is one of the most challenging speech recognition tasks to-date. While recognition error rates of 10% or lower can now be reached on speech dictation tasks over vocabularies in excess of 60,000 words, recognition of conversational speech has persistently resisted most attempts at improvements by way of the proven techniques to date. Difficulties arise from shorter words, telephone channel degradation, and highly disfluent and coarticulated speech. In this paper, we describe the application, adaptation, and performance evaluation of our JANUS speech recognition engine to the Switchboard conversational speech recognition task. Through a number of algorithmic improvements, we have been able to reduce error rates from more than 50% word error to 38%, measured on the official 1996 NIST evaluation test set. Improvements include vocal tract length normalization, polyphonic modeling, label boosting, speaker adaptation with and without confidence measures, and speaking mode dependent pronunciation modeling.