SPECIALIZED ACOUSTIC MODELS FOR HYPERARTICULATED SPEECH

Hagen Soltau and Alex Waibel

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)
{soltau,waibel}@ira.uka.de

ABSTRACT

This study aims to improve the performance of automatic speech recognizers at hyperarticulated speech. Hyperarticulation often occur as a strategy to recover previous recognition errors in spoken dialogue systems. Contrary to this intention a significant performance degradation can be observed at hyperarticulation. In this paper we present an analysis of features that caused the performance lost. The average phone duration is nearby 20% longer. Pitch contour and fundamental frequency changes significantly at hyperarticulation. We report on adapting acoustic and transition models to hyperarticulated speech. We achieved a word error reduction about 23% at hyperarticulation.

1. INTRODUCTION

The usability of spoken dialogue and dictation systems strongly depends on the fact that a user can enter any information faster into the system using speech technology instead of typing it. But now, current state of the art speech recognizers will always exhibit some errors. The advantages of speech interfaces will be greatly reduced through the time needed for error correction [8]. To develop user friendly speech interfaces, it is important to examine, how users react to recognition errors. When humans use recognition technology it is commonly observed, that they follow similar recovery strategies as in interaction with humans. These strategies are typically attempts at speaking more clearly and accented in an effort to disambiguate the original mistake. Previous studies [5, 7] demonstrate that the speaking styles changes significantly in such situations and the recogniton accuracy decrease. As an effect of the worse recognition, a user will try to speak in a stressed way even more as previously, and the recognition performance will become more and more worse.

The paper is organized as follows: In the first section we describe our experimental setup, our database with normal and hyperarticulated speech, and our baseline recognition system that we used. We will give some details about our procedure to collect hyperarticulated speech in a spoken dialogue system scenario. After that we present an analysis of features to detect hyperarticulation. In the last section we report on adapting acoustic and transition models to hyperarticulated speech and summarize our progress to reduce the word error rate at hyperarticulation.

2. EXPERIMENTAL SETUP

2.1. The data

We have collected a german database with normal and hyperarticulated isolated speech. In order to induce hyperarticulated speech realisticly we analyzed typical errors of our current LVCSR system at first and generated a list of frequent confusions. The recording scenario consists of two sessions.

In the first session data were recorded with normal speaking style. We selected 50 word pairs for each speaker. Each word pair consists of a word and the corresponding confusable word (as per error analysis). We presented the 2 x 50 words independent of each other in the first section without any instructions. In the second session, we tried to induce hyperarticulated speech. We simulated recognition errors and presented phrases like "Word A was confused with Word B. Please repeat Word A" up to three times for each word pair. The decision if the system accepts or rejects the input was chosen randomly. To avoid monotonous spoken utterances from bored subjects we set the probability for two attempts to 20% and for three attempts to 10% only. Since we assumed that opposite features are used to disambiguate two words A vs. B and B vs. A, respectively we presented each word pair in reverse order also.

For each speaker we collected 100 normally spoken words in the first session and approximately 120 hy-

perarticulated words in the second session with this strategy. Table 1 shows the size of our data collection.

	Spk	utterances		speech	
		normal	hyper	normal	hyper
train	61	5901	7309	154 min	235 min
test	20	1926	2374	47 min	72 min
all	81	7827	9683	202 min	307 min

Table 1: Database for normal and hyperarticulated speech

2.2. Baseline recogniton system

The Recognizer used for this experiments was build using our JANUS-III Speech Recognition Toolkit. The baseline system is a 60k vocabulary semi continuous speech recognizer. For speech extraction, we derive 13 MEL-scaled cepstral coefficients (MFCC) with first and second order derivatives normalized with cepstral mean subtraction. The vector dimension is reduced to 32 by performing an linear discriminant analysis. For the acoustic model, we use 10000 context-dependent subquintphones build in a two-stage decision tree based clustering approach. The acoustic models are trained with around 90 hours of spontaneous and read speech. Vocal tract length normalization and speaker adaptation is applied during training and decoding. The performance of the recognizer is currently at 85% word accuracy with a 60k vocabulary and an oov-rate of 3.5% on a continuous test set.

2.3. Recognition of isolated speech

Since our baseline system is optimized for continuous speech, results of first recognition runs with isolated speech are very poor. We achieved 63.7% word accuracy only. To avoid segmentation errors we added a second search pass. In the second search pass we use a word list generated from the first pass and restrict hypotheses to isolated words only. By comparing the likelihoods from both passes we can automatically detect isolated speech.

Additionally, we adapted the acoustic models to isolated speech using Maximum a Posteriori Smoothing [2]. Usually, our transition models based on a 3 state left-to-right architecture with fixed transition probabilities. This is sufficient for continuous speech. It is commonly observed, that the use of duration modeling or training of transition probabilities doesn't improve word accuracy significantly. Interestingly, we found this is not true in case of isolated speech as you can see in table 2. We attribute this result to the fact that training transition models helps to adjust different speaking rates with continuous and isolated speech. With this modifications we have improved the word accuracy from 63.7% to 80.5% for isolated speech.

system	Speaking Style	
	normal	hyper
baseline	63.7%	51.5%
second search pass	67.3%	56.8%
MAP Adaptation	78.5%	71.6%
Duration Modeling	78.6%	71.7%
Transition Modeling	80.5%	74.6%

Table 2: baseline results for normal and hyperarticulated speech (results in word accuracy)

3. MODELING HYPERARTICULATED SPEECH

The results in table 2 demonstrate a strong mismatch between hyperarticulated speech and acoustic models. Even after adaptation of acoustic and transition models, the error rates are 30% higher at hyperarticulation. To model hyperarticulated speech in a HMM based framework properly, there are several issues:

1. Feature Space
   MFCC and other feature extraction methods based on separating pitch and envelope in the spectral domain and assume the fundamental frequencies are below a certain threshold. In our analysis, we found that the fundamental frequencies increase at hyperarticulation. One approach to fit the assumptions is to adapt the cepstrum filter dynamically. However, it is not feasible to use separate feature spaces for different speaking styles.

2. Pronunciation Modeling
   Another source of wrong model assumptions is the pronunciation lexicon. In some hyperarticulated excerpts we observed changes in the vowel qualities. For such cases, we should add appropriate variants to the lexicon. To that end, we generated a list of context dependent vowel substitution rules automatically which were applied to the original dictionary.

   In a first step, we decoded phone sequences using the baseline acoustic models and phone networks such as shown in figure 1. After that we aligned

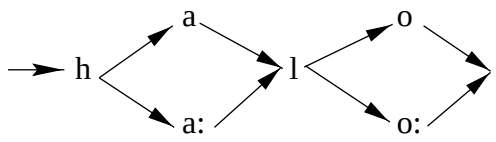

Figure 1: phone recognition along flexible phonetic transcriptions. example HALLO: /h/ /a/ /l/ /o:/). colons indicate long vowel qualities

this phone hypotheses with the original pronunciations using a dynamic programming approach and generated a list of context dependent vowel substitution rules. The most probable rules were applied to the dictionary. We added 16,000 new pronunciations in total. Because of memory limitations, we used this dictionary in the second search pass only. However, we could not observe any improvements with the new dictionary. One reason for that is maybe the increased word confusability. The increased number of pronunciation variants can induce that different words have similar pronunciations in the dictionary.

3. Acoustic Modeling
   Besides the feature extraction and the pronunciation lexicon the acoustic models can cause heavy performance degradations if they doesn't fit to the data. The standard approach to improve the acoustic model on unseen data conditions is to adapt the models with MLLR or MAP depending on the size of the adaptation data. The question that arise is if it is better to train one set of models for both normal and hyperarticulated speech or is it better to train separate models. In the latter case we need to classify utterances if they are hyperarticulated or not to use the appropriate acoustic model. In the following subsections we analyze some acoustic features to detect hyperarticulation and report about our recognition experiments using separate acoustic models for hyperarticulated speech.

3.1. Detecting Hyperarticulation

1. Phone duration
   In section two we observed that the gain obtained by using transition models is higher for hyperarticulated data than for normal data. This is a clue that the speaking rate at hyperarticulated speech differ from the rate at normal speech. To analyze phone durations, we have done a forced alignment to get a phone based segmentation. The results for different phone classes are summarized in table 3. On average, we observed that the duration is increased by 20%. More exactly, the changes are mainly at the voiced consonants and schwa sounds.

Phone Group	average phone duration		
	normal	hyper	relative
all phones	59 msec	70 msec	20%
vowels	72 msec	85 msec	17%
consonants	82 msec	102 msec	25%
consonants, voiced	75 msec	96 msec	29%
consonants, unvoiced	87 msec	107 msec	22%
schwa	82 msec	109 msec	33%

Table 3: Phone Duration Analysis

2. F0 mean
   To analyze the effect of pitch, we did a t-test (student-test) for paired samples to level alpha = 0.005 and divided the test speaker into three groups where the mean of $F0$ increased, decreased, or didn't change between normal and recovery mode. For each of this groups we computed the average word accuracy for the baseline system. The results refereed in table 4 shows that there is a correlation between $F0$ and the performance degradation at hyperarticulation.

mean of F0	speaker	Speaking Style		delta
		normal	hyper	
increasing	8	81.3%	70.7%	-10.5%
decreasing	6	82.5%	81.4%	-1.1%
changed not	6	77.8%	73.6%	-4.2%

Table 4: word accuracy as a function of F0 changes

3. F0 contour
   Besides the F0 mean, the contour can also be interesting for our purposes. The problem is that not every utterance spoken to correct a recognition error is hyperarticulated. One approach is to detect hyperarticulation by comparison the utterances spoken the first time and spoken to correct the recognition errors and assume hyperarticulation if the respective contours differ. To compare two pitch contours we segment each utterance into pieces of 200 msec and compute the gradient of the pitch contour for each segment separately. We consider only the direction of the

gradient but not the absolute value. Briefly, we assume that hyperarticulation is occured if the direction of the gradients are differ.

A more simple way to use the pitch contour information to detect hyperarticulation is to consider the end of the utterance only. Normally a increasing pitch at the end indicate questions. But it can also serve as a clue to indicate an emotional state of the speaker.

3.2. Recognition Experiments

In table 5 we summarize the recogniton results for different acoustic models and selection criteria described above. Using likelihood as selection criterion means to evaluate both acoustic models and take the hypothesis that produce the higher likelihood. The last two experiments using the data base information or using hypothesis are cheating experiments. In the latter case we match the hypotheses to the reference and take the best hypothesis. This is to see what is theoretically possible if we have the perfect selection criterion. The results show that we can can improve the performance by using separate acoustic models significantly (81.8% / 78.2% by using the likelihood criterion) in comparison with the baseline system (80.5% / 74.7%).

Acoustic model	Speaking Style	
	normal	hyper
baseline	80.5%	74.7%
shared model	81.2%	76.7%
separate models: select model by		
- F0 mean	81.7%	77.3%
- F0 contour, variant 1	82.0%	77.5%
- F0 contour, variant 2	81.6%	76.6%
- likelihood	**81.8%**	**78.2%**
- data base information	82.0%	77.9%
- hypothesis alignment	83.2%	79.4%

Table 5: recognition results for different acoustic models (results in word accuracy)

4. CONCLUSIONS

We have described an approach to handle hyperarticulated speech. We reduced the error rate for hyperarticulated speech by using separate acoustic models significantly. However, the gap between normal and hyperarticulated speech isn't closed. Surprisingly we didn't get any gains from our pronunciation modeling yet. In future work, we will focus on better learning hyperarticulated pronunciations. Another interesting question is, if they are the same or similar effects in different languages. To that end, we started a project to collect English hyperarticulated speech already.

5. ACKNOWLEDGMENTS

The authors wish to thank all members of the Interactive Systems Labs for useful discussions and active support.

6. REFERENCES

[1] F. Alleva, X. Huang, M. Hwang, and L. Jiang. Can continuous speech recognizers handle isolated speech? In *Proceedings of the Eurospeech*, Rhodos, Greece, 1997.

[2] J. Gauvain and C. Lee. Maximum a poseriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, April 1994.

[3] J. Humphries. *Accent Modelling and Adaptation in Automatic Speech Recognition.* PhD thesis, University of Cambridge, 1998.

[4] P. Lieberman and S. Blumstein. *Speech physiology, speech perception, and acoustic phonetics.* Cambridge University Press, 1988.

[5] S. Oviatt. The CHAM model of hyperarticulate adaptation during human-computer error resolution. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[6] K. Scherer. Vocal effect expression: A review and a model for future research. *Psychological Bulletin*, 99, 1986.

[7] H. Soltau and A. Waibel. On the influence of hyperarticulated speech on the recogniton performance. In *Proceedings of the International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[8] B. Suhm. *Multimodal Interactive Error Recovery for Speech User Interfaces.* PhD thesis, University of Karlsruhe, Germany, 1998.