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

In spoken dialogue systems, hyperarticulation occur as an effect to recover previous recognition errors. It is commonly observed that in particular real users apply similar recovery strategies as in human-human interactions. Previous studies have shown that current speech recognizer cannot handle hyperarticulated speech. As an effect of higher word error rates at hyperarticulated speech, humans try to reinforce this speaking style which result in even more recognition errors. In this paper, we present approaches to build robust acoustic models for hyperarticulated speech. One key point is that the changes of acoustic features at hyperarticulation is a phone dependent effect. The idea is to use the likelihood criterion to decide, which phones should be treated separately. This can be done by incorporating dynamic questions about hyperarticulation into the clustering stage. Based on such phonetic decision tree, we can generate appropriate acoustic models. With this method, we achieved a word error reduction about 9% relative at hyperarticulation.

1. INTRODUCTION

The usability of spoken dialogue and dictation systems strongly depends on the fact that a user can feed any information into the system faster using speech technology instead of typing. One critical issue in building intelligent human computer interfaces is failure tolerance. However current state of the art speech recognizer will always exhibit some errors. In case of recognition errors, a user will switch to other modalities (handwriting, gestures, typing) or just try to repeat the misrecognized phrase. As a consequence, the advantages of speech interfaces will be greatly reduced through the time needed for error correction [9].

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. Oviatt et. al presented in [6] a user study in which the reactions on word errors were examined. They observed that the duration of utterances increase, both speech segments and number and duration of pauses. Word repetitions were spoken more clearly than in the original spoken utterance. The question that arises is if such an user reaction helps the system to find the correct word hypothesis. In [8] we demonstrated that the recognition rates are worse at hyperarticulation contrary to the users intention. In particular, we observed that higher $F0$ values at hyperarticulation are correlated to worse recognition results.

In principle, the problems at hyperarticulation can be attributed to different components of a speech recognizer, namely the pronunciation models, duration models, and last but not least the acoustic models. In this paper, we focus to reduce the mismatch between the acoustic models and the speech patterns that occur at hyperarticulation. One key point is that the changes of acoustic features at hyperarticulation is a phone dependent effect. For example, the phone duration is increased by 44% for plosives, but only 16% for vowels. However, standard adaptation techniques doesn't make use of such knowledge. We therefore constructed phonetic context trees which make explicit use of questions about hyperarticulation.

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 report about constructing hyperarticulated models and analyze phone dependent hyperarticulated effects.

2. EXPERIMENTAL SETUP

2.1. The data

We have collected a English 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 but similar to real error rates. 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 hyperarticulated words in the second session with this strategy. In total, we've got recordings from 45 subjects. For testing purposes, 11 speaker were excluded.

2.2. The Speech Recognition Engine

The Recognizer used for this experiments was build using our JANUS-III Speech Recognition Toolkit. The baseline system is a 30k 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 20 by performing an linear discriminant analysis. For the acoustic model, we use 800 context-dependent sub-quintphones build in a two-stage decision tree based clustering approach. The acoustic models are trained with around 52 hours of spontaneous and read speech. Vocal tract length normalization is applied during training and decoding. Cepstral mean subtraction is used to compensate channel differences.

3. CONSTRUCTING ACOUSTIC MODELS FOR HYPERARTICULATED SPEECH

In a first run we used the system described above to generate baseline results. As we can see in table 1, the performance is quite poor. To reduce the channel and speaking style [1] mismatch we adapted the acoustic models using MLLR with 233 classes. In all this experiments, we used the same size of adaptation data for normal as for hyperarticulated speech. After the adaptation, the word error rate dropped down to 23.9% for normal speech and 33.9% for hyperarticulated speech. That means, that there is still a performance gap of more than 40% at hyperarticulated speech. In a second experiment, we examined to extenuate the speaking rate mismatch between continuous and isolated speech by training phone dependent transition models. As shown in table 1, there were only small improvements for normal isolated speech, but indeed a significant error reduction for hyperarticulated speech.

system	Speaking Style	
	normal	hyper
baseline	33.8%	45.0%
adapt acoustic models	23.8%	33.9%
train transition models	23.0%	29.8%

Table 1: initial experiments for normal and hyperarticulated speech (results in word error rates)

3.1. Phone dependent effects

In preliminary experiments, we tried to train separate acoustic models for hyperarticulated speech. The problem that occurred was that the acoustic characteristics change only for certain phones at hyperarticulation. Now, if we train separate acoustic models for normal and hyperarticulated speech but only certain speech states are affected by hyperarticulation, this result in a insufficient data sharing across the models. This is even true for adaptation techniques like MLLR and MAP since the regression tree base usually on how close are the acoustic components of the original models and how much adaptation data is available. As a consequence of this, the splitting of the speech states in a normal and hyperarticulated part base on how much data are available but not if the acoustic characteristics differ for this speech state. But what we want is to separate the models

[1] The original system was trained with continuous speech, but we use here isolated speech.

if the acoustic characteristics differ at hyperarticulation and not only if enough data is available. To demonstrate that hyperarticulation is indeed a phone dependent effect, we compared some acoustic features for both normal and hyperarticulated speech.

- Formant Analysis
  The formant values were extracted by computing the roots of the predictor polynomials from the speech signal. The first two formants of the vowels AA, IH, and UW are shown in table 2. We did a t-test at level $\alpha = 0.05$ to see if there are significant differences at hyperarticulation. This was the case only for the vowel *UW*.

vowel	formant	Speaking Style		t-test
		normal	hyper	at alpha=0.05
AA	F1	840	868	differ not
AA	F2	1797	1914	differ not
IH	F1	676	645	differ not
IH	F2	2179	2173	differ not
UW	F1	476	591	differ
UW	F2	1867	2117	differ

Table 2: formant frequencies (Hz) for vowels AA, IH, UW

- Phone Duration Analysis
  The larger gains for hyperarticulated speech obtained by training transition models (see table 1) indicated already that the phone durations change at this speaking mode. The phone durations were estimated by a forced alignment with the adapted acoustic models. In particular, the phone durations for the voiced plosives /B/, /D/, and /G/ are increased by more than 40 %. On the other hand, the durations of vowels changed not very much.

phone class	average phone duration		
	normal	hyper	relative
all phones	86 msec	110 msec	28%
vowels	101 msec	117 msec	16%
consonants	100 msec	132 msec	32%
plosives	79 msec	114 msec	44%
fricatives	124 msec	156 msec	26%
nasal	95 msec	127 msec	33%
glottal	148 msec	181 msec	22%

Table 3: average phone duration at normal and hyperarticulated speech

3.2. Integrating Hyperarticulation into decision trees

In the last section, we have seen some clues that not all phones are affected by hyperarticulation. In order to train appropriate models we have to find out which phones or speech states should be treated separately. Instead of using formant or duration values as a splitting criterion, we use the likelihood to decide if a speech state is affected by hyperarticulation. This can be done by incorporating dynamic questions about hyperarticulation into the clustering stage. This allows us to model this speaking mode in a common framework [2]. When building the phonetic decision tree to generate context dependent models, we now use additional questions concerning hyperarticulation. As a consequence, we split automatically these models into a normal and hyperarticulated part, only if there are really different acoustic features given the current context.

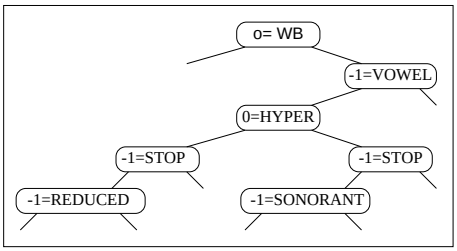

Figure 1: excerpt from the decision tree for /Z/

Fig 1 shows the results of the clustering procedure where such questions were used. Questions like "-1=?" will ask about the left context, "0=?" about the identity and so on. Left/right branches correspond no/yes answers. In this new decision tree, 15% of all nodes are now depend from the speaking mode. This confirm also that only certain speech states are affected by hyperarticulation.

phone class	hyperarticulated questions
vowels	3.5%
consonants	20.8%
- nasals	23.8%
- plosives	21.6%
- fricatives	24.6%
- approximants	9.8%

Table 4: splits relating to manner of articulation

In table 4 we analyzed which phones are mainly separated. It seems, that the acoustic space of vowels

doesn't change in a error recovery mode in contrast to the consonants. Only 3.5% of the vowel models depend on the hyperarticulated speaking mode, but more than 20% of the consonants.

phone class	hyperarticulated questions
bilabial	8.0%
labiodental	0.0%
alveolar	24.3%
retroflex	0.0%
velar	41.7%

Table 5: splits relating to place of articulation

The distribution of hyperarticulation dependencies according to the place of articulation is shown in table 5. Mainly alveolar and velar sounds exhibit acoustic changes at this speaking mode.

questions	Speaking Style		error
	normal	hyper	increase
context	23.0%	29.8%	29.6%
speaking mode	23.3%	27.1%	16.3%

Table 6: decision tree experiments (results in word error rates)

We build a context dependent system with this new decision tree by standard viterbi training. We used the same number of parameters as we used for the baseline system. Compared to the standard tree, the error rate has decreased from 29.8% to 27.1% at hyperarticulation with only a small performance degradation of 0.3% at normal speech. The performance gap between both speaking modes is now only 16.3% relative.

4. CONCLUSIONS

To build spoken dialogue systems for real world applications, it is necessary to model, how users react to recognition errors. Hyperarticulation cause a performance degradation of more than 30% relative. Only certain phones (mainly nasals and fricatives) are affected by hyperarticulation. By extending phonetic context decision trees with dynamic questions about hyperarticulation we achieved a word error reduction of 9% relative.

5. ACKNOWLEDGMENTS

The authors wish to thank all members of the Interactive Systems Labs for useful discussions and active support. Special thanks to Nils Hammer and Daniel Schneider for data collection.

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] C. Fuegen and I. Rogina. Integrating dynamic speech modalities into context decision trees. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

[3] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler. Acoustic characteristics of american english vowels. *The Journal of the Acoustical Society of America*, 97, 1995.

[4] J. Hirschberg, D. Litman, and M. Swerts. Prosodic cues to recognition errors. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Keystone, USA, 1999.

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

[6] 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.

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

[8] 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.

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

[10] D. van Kuijk and L. Boves. Acoustic characteristics of lexical stress in continuous telephone. *Speech Communication*, 27, 1994.

[11] C. Williams and K. Stevens. Emotions and speech: Some acoustic correlates. *The Journal of the Acoustical Society of America*, 52, 1972.