ON THE INFLUENCE OF HYPERARTICULATED SPEECH ON RECOGNITION PERFORMANCE

Hagen Soltau and Alex Waibel

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

ABSTRACT

Since we cannot exclude that speech recognizers fail sometimes, it is important to examine how users react to recognition errors. In correction situations, speaking style becomes more accentuated to disambiguate the original mistake. We examine the effect of speaking style in such situations on speech recognition performance. Our results indicate that hyperarticulated effects occur in correction situations and decrease word accuracy significantly.

1. INTRODUCTION

Considerable progress has been achieved in speech recognition over the last years through techniques such as vocal tract length normalization (VTLN), maximum likelihood linear regression (MLLR) or speaker adapted training. However, even in dictation applications with 95% word accuracy (WA) it is often necessary to correct word errors.

Studies [5] show that a user can lose a lot of time through error correction that he won through dictating instead of typing. Since recognition systems will always exhibit some errors it is important to examine how users react to recognition errors. Various user strategies to correct word errors have been investigated. Usually, a user repeats misrecognized words first of all. Only if this doesn't lead to success the user applies paraphrases or spells the word or use other modalities if available [5].

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 in an effort to disambiguate the original mistake. Oviatt et. al presented in [4] 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 arise is if the user reaction helps the system to find the right hypothesis. In this paper, we examine the effect of speaking style and hyperarticulation in such situations on speech recognition performance. To that end, we have collected an isolated word speech corpus with different speaking styles. The database and the strategy to induce hyperarticulated effects is described in the next section.

Our baseline system is a continuous speech recognizer for dictation tasks. Since our database is an isolated speech corpus we have adapted the acoustic models and the search engine to isolated speech. We have reduced the error rate by 41.9% over the baseline LVCSR (large vocabulary continuous speech recognition) system. In section three, we summarize our efforts to recognize isolated speech.

We discuss the effects of speaking style and hyperarticulation in section four. We examine prosodic features, pitch, and phoneme durations. The results indicate that hyperarticulated effects leads to a significant lost in word accuracy contrary to the users intention.

2. DATABASE

We have collected a german database with normal and hyperarticulated isolated speech. In order to induce hyperarticulated speech realisticly we analyzed typical errors of the LVCSR system at first and generated a list of frequent confusions. 72.8% of the errors are substitutions, 14.6% deletions, and 12.6% insertions. The most confusions are caused through inflections. The german language has a lot of words that differ only in suffix, for example $ein, eine, eines, einem$ or $einen$. The translation of all words is the indefinite article a. But there are also highly confusable words with different meanings, for example $Maus - Haus - Klaus$[1], or $Mut - Wut - Hut$[2], or $erlangen - ergangen$[3].

We extracted a list of such word pairs from the LVCSR alignments. The data collection base on this list of word pairs. The recording scenario consists of two sections. In the first section 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 times

[1] Maus = mouse, Haus = house, Klaus is a proper name
[2] Mut = courage, Wut = rage, Hut = hat
[3] erlangen = obtain , ergangen = endure

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 "Mut was confused with Wut. Please repeat Wut" 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 section and approximately 120 hyperarticulated words in the second session with this strategy. The problem is that the recording procedure needs approximately 30 min for each speaker but the real speech that we get is only between 6 to 7 min. Table 1 shows the size of our data collection. The database consists of 81 speaker in total. All results in the next sections are based on the test data.

	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

3. RECOGNITION OF ISOLATED SPEECH

The baseline system is a $60k$ vocabulary continuous speech recognizer. For speech extraction, we derive 13 MEL-scaled cepstral coefficients with first and second order derivatives normalized with cepstral mean subtraction. The vector dimension is reduced to 32 by performing an LDA. A maximum likelihood VTLN procedure is used for speaker normalization. For the acoustic model, we use 10000 senones sharing 2500 codebooks. The senones are modeled of mixtures of 16 Gaussians of diagonal covariance. Cross-word triphones are also modeled. The decoder has multiple search passes including warp factor estimation, MLLR adaption and lattice rescoring. The performance of the recognizer at present is about 85% word accuracy with a 60k vocabulary and an oov-rate of 3.5%.

3.1. Search pass

Since substitutions errors occur more often than insertion and deletion errors, isolated words are usually sufficient to correct word errors and whole word phrases are as necessary. Therefore, we restricted the decoder to handle isolated words only.

Instead of multiple search passes for continuous speech decoding we compute the best path through an HMM containing all vocabulary words by performing a viterbi alignment. To segment speech from non-speech we added a silence state at start and end to the HMM. This HMM is depicted in figure 1. We did not use word priors.

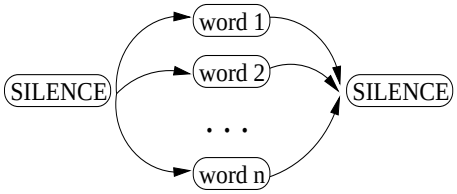

Figure 1: HMM used for the search pass

The baseline result is shown in table 2. Without adaptation of the acoustic and transition models to isolated speech we achieve only 66.4% word accuracy on normal speaking style. We expected a loss in performance because of the mismatch between isolated speech and continuous acoustic models. But we should also consider the fact that our database contain those words are rather difficult to recognize because the data collection was based on the error analysis. The results on hyperarticulated speech are discussed later in section four.

Acoustic Model	Transition Model	Speaking Style	
		normal	hyper
baseline	-	66.4%	57.7%
mllr (126)	-	76.6%	69.9%
mllr (126)	adapted	79.6%	72.9%

Table 2: MLLR adaption (results in word accuracy). Number of adapted mllr matrices into brackets

3.2. Acoustic and Transition Models

Alleva et.al. [1] observed that the acoustic models significantly differ from isolated and continuous speech and recognizing both isolated and continuous speech is not only a decoding problem.

We examined two methods to adapt acoustic models. First, the codebooks were supervised adapted with MLLR [3]. For the adaptation we used only the normally spoken data. The adaptation improved the word accuracy to 76.6%

Additionally, we trained transition probabilities of the HMM states. Originally, we have used a six state transition model with equal transition probabilities (figure 2).

Although in earlier experiments on continuous speech training the transition models did not bring improvements, here it caused an error reduction around 3%. We attribute this result to the fact that training transition models helps to adjust different speaking rates with continuous and isolated speech.

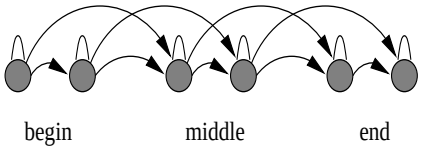

Figure 2: standard six state transition model

An analysis of the adapted transition models shows that the probabilities for self-loops significantly increased. On average, the self-loop probabilities are about 0.72 for the most phonemes except for plosives.

We examined also a second method [2] to adapt acoustic models. The problem is that we don't have enough data to train acoustic models with isolated speech only. And if we train using both corpora the problem occur that the new models are rather continuous than isolated speech models because the continuous speech corpus is much bigger than the isolated. We have about 89 hours continuous speech and only 2.5 hours isolated speech (normally spoken). Therefore, we pushed the isolated data artificially using a weighting factor. This was done by manipulating the counts for the model accumulators. All models received the complete data but we weighted isolated speech much higher.

Acoustic Model	Transition Model	Speaking Style	
		normal	hyper
mix (0.3)	adapted	74.3%	-
mix (0.1)	adapted	77.8%	-
mix (0.05)	adapted	79.6%	-
mix (0.012)	adapted	**80.5%**	74.7%
mix (0.005)	adapted	79.9%	-
mix (0.012)	-	78.5%	71.6%

Table 3: mixed training (results in word accuracy). mixing factor into brackets

The results are shown in table 3. We achieved the best results with a weighting factor of 0.012. That means that the originally ratio from isolated to continuous speech that was $2.5/89 = 1/35$ is now $(2.5{*}0.988)/(89{*}0.012) = 2/1$ approximately. We did not cluster new context dependent models. We trained only one iteration with the mixing factor. To compare with MLLR adaptation the mixed training gave us a gain from 79.6% to 80.5%. The drawback to the mixed training is the additional free parameter.

4. HYPERARTICULATED EFFECTS

To compare the results between both speaking styles we used our best isolated speech system $mix(0.012)$ with adapted transition models. First, we divided the 20 test speaker into four groups depending on their word error rates (table 4). Differences smaller than 5% were viewed being not significant. The word accuracy increased for two speaker only. But for most of the test speaker, we observed a significant lost in performance about 10%.

Speaker Group depending on WA	Spk	Speaking Style		delta
		normal	hyper	
significantly better	2	72.5%	79.6%	+7.1%
significantly worse	12	81.9%	71.4%	-10.5%
slightly better	3	81.4%	82.5%	+1.1%
slightly worse	3	79.3%	76.9%	-2.4%

Table 4: compare normal with hyperarticulated speech in word accuracy

The results indicate that the speaking style change clearly in correction mode and reduce the recognition performance often. It seems that the acoustic models don't fit hyperarticulated speech, mostly. But we observed also that the changed speaking style caused an error reduction in two cases.

4.1. Pitch Analysis

We examined which acoustic and prosodic features differ between both speaking styles.

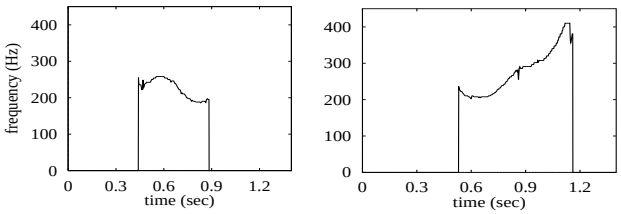

Figure 3: Pitch contours for the word "Leonard" from a female speaker, normally spoken (left side) and hyperarticulated (right side)

First, we extract pitch information from the signal. For this, we used a free available pitch tracking software. $F0$ mean and standard deviation were calculated only on voiced regions. Examples for pitch contours for both normal and hyperarticulated speech are depicted in figures 3 and 4 for a female and a male speaker, respectively. In particular, figure 4 shows clear differences between both utterances. It shows the pitch for the word $Leonard$ in

both cases but the female was asked to correct the confusion $Leonard$ (proper name) with $Leopard$ (leopard) in the second case. Contrary to normal spoken speech, in correction mode the pitch increased during the time for the word $Leonard$ significantly.

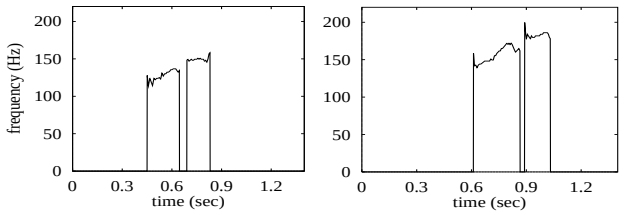

Figure 4: Pitch contours for the word "Standes" from a male speaker, normally spoken (left side) and hyperarticulated (right side)

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 correction mode.

$F0$	Speaker	Speaking Style		delta
		normal	hyper	
increasing	8	81.2%	70.7%	-10.5%
decreasing	6	82.5%	81.4%	-1.1%
changed not	6	77.8%	73.6%	-4.2%

Table 5: word accuracy as a function of $F0$ changes

The results in table 5 show that the $F0$ mean of 8 speaker significantly increased on hyperarticulated speech and that for these speaker the word accuracy decreased about 10.5%. On the opposite, the speaker with a decreasing $F0$ mean lose only 1.1% word accuracy in correction mode. The increase in error rate in the last line in table 5 can be attributed to other features than pitch.

4.2. Duration Analysis

To analyze phone durations, we have done a forced alignment and counted the frames for each phone model. The results for different phone classes are summarized in table 6 and indicate that there is a connection between loss in performance and increasing duration. In particular, the duration of unvoiced phones increased with those speaker that have significantly loss in word accuracy.

Speaker Group depending on WA	increasing duration		
	voiced	unvoiced	plosives
significantly better	3.9%	-0.4%	-4.2%
significantly worse	25.7%	31.2%	24.4%
slightly better	8.2%	3.9%	15.2%
slightly worse	17.9%	22.4%	17.3%

Table 6: relative increasing phone duration in correction mode for different phone classes

5. CONCLUSIONS

We described our efforts to adapt a continuous speech recognizer to recognize isolated speech. An error reduction of 41.9% was achieved by adapting both acoustic and transition models. Our experiments show that hyperarticulated effects occur in correction situations and decrease the word accuracy significantly.

6. ACKNOWLEDGMENTS

The authors wish to thank all members of the Interactive Systems Labs, especially Thomas Kemp and Thomas Schaaf, for useful discussions and active support.

7. 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] C.J. Leggetter and P.C. Woddland. Speaker adaption of HMMs using linear regression. Technical report, Cambridge University, England, 1994.

[4] S. Oviatt, G.-A. Levow, M. MacEachern, and K. Kuhn. Modeling hyperarticulated speech during human-computer error resolution. In *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.

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