COMPENSATING FOR HYPERARTICULATION BY MODELING ARTICULATORY PROPERTIES

Hagen Soltau, Florian Metze, and Alex Waibel

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

ABSTRACT

In spoken dialogue systems, hyperarticulation occur as an effect to recover previous recognition errors. It is commonly observed that users of automatic speech recognition systems apply similar recovery strategies as in human-human interactions. Previous studies have shown that current speech recognizers don't cover hyperarticulated speech well. As an effect of higher word error rates at hyperarticulated speech, humans try to reinforce this speaking style which results in even more recognition errors. In this study, we investigate the use of articulatory features to compensate hyperarticulated effects. The underlying idea is, that acoustic models for articulatory features are more robust against variations in the speaking style compared to pure phone models. We present a streaming architecture which integrates articulatory features in a standard HMM based system. Using this approach, we achieved an error reduction of 25.1% for hyperarticulated speech and even 8.9% for normal speech without any use of hyperarticulated training data.

1. INTRODUCTION

The usability of spoken dialogue and dictation systems strongly depends on the fact that an 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, an 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 [12].

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. In [7] an user study is presented, 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 such an user reaction helps the system to find the correct word hypothesis. In [10] we showed that the recognition rates are significantly worse at hyperarticulation contrary to the users intention.

Hyperarticulated speech exhibits differences in speaking rate, pitch contour, or formant frequencies in order to stress a certain part of the utterance. To model these changes in the acoustic space, we examined in [11, 2] how to integrate dynamic questions about the speaking style in a context decision tree. We achieved an error reduction of 9% by using these speaking mode dependent acoustic models. The model splits related to hyperarticulation were clearly phone dependent, e.g. consonants seems to exhibit more significant changes compared to vowels. On the other hand, we investigated whether these effects can be compensated by phone substitutions, e.g. confusions of the nasal sounds /m/ and /n/ can be stressed at hyperarticulation. However, the training of hyperarticulation pronunciations using decision trees didn't work well, which indicate that *different* sounds are produced under these conditions, which don't match to any of the pretrained phoneme models.

Motivated by these previous results, we investigate in this study how hyperarticulated effects can be modeled by articulatory features. In the next sections, we give details how we extracted articulatory features and describe the system architecture which integrate theses models in a standard HMM system using streams.

2. ACOUSTIC MODELS SUPPORTED BY ARTICULATORY FEATURES

The assumption for using articulatory features (AF) to compensate hyperarticulated effects is that people don't substitute a whole phone in order to contrast a previous recognition error. For example, the nasal sounds /m/ and /n/ can be described as *+voiced,+nasal,+labial* and *+voiced,+nasal,+velar* respectively. As a consequence, the hyperarticulated speech might exhibit differences according to the place of articulation in order to disambiguate such a recognition error but doesn't exhibit changes according to the *voice* and *nasal* attributes. Acoustic models based on articulatory features would therefore allow a more precise modeling of hyperarticulated effects. As a first step, we included articulatory features into a pure phone based HMM system by using multiple streams to compute acoustic scores as described in [8, 5].

2.1. Extraction of articulatory features

By mapping the phonemes to bundles of articulatory features [6], we generated time segmentations for the features based on viterbi alignments using the phone based HMM system. To avoid incorrect or unprecise alignments, we used only the time segmentations aligned to the middle state of the three state HMMs, since the viterbi paths don't exhibit sharp phone boundaries usually. These converted training data were then used to train a GMM for each of the binary articulatory features. Additionally, an anti-model was trained for each articulatory model using all data which do not belong to that articulatory feature. An example of the extracted features is shown in figures 1 and 2 where the difference of the likelihood of the *fricative* and *plosive* models and anti-models are plotted over the time. The figures contain the curves for the word *doubts*, once spoken normally and once hyperarticulated. The hyperarticulation occured in order to disambiguate *doubts* from *doubt* which is clearly represented in the figure. While the *fricative* attribute becomes stressed under hyperarticulation, the *plosive* feature is suppressed to distinguish from the previous */t/* sound.

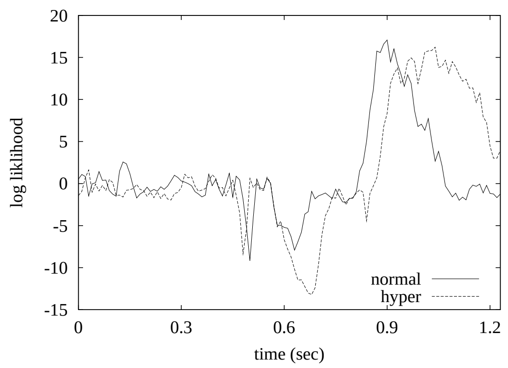

Figure 1: curve of the *fricative* likelihood for the word *doubts*, spoken normally and hyperarticulated

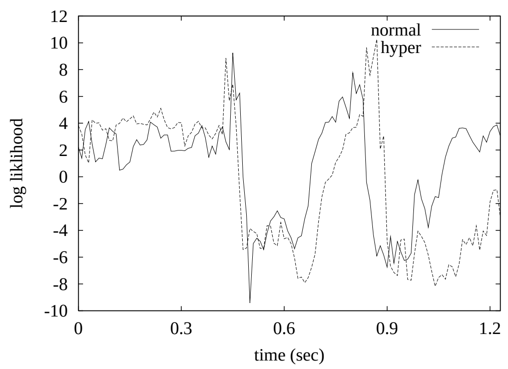

Figure 2: curve of the *plosive* likelihood for the word *doubts*, spoken normally and hyperarticulated

2.2. System Architecture

Starting from a standard context dependent phone based HMM recognizer, we attach for each phone model the corresponding GMMs for the articulatory features. For example, a HMM state belonging to */m/* has the associated models for *+voiced,+nasal,+labial* and anti-models for *-unvoiced,-velar* etc. These bundles of models will be combined using an exponential weighting of the resulting likelihoods, e.g. the acoustic log-likelihood of such a bundle will be computed as the weighted sum of the log-likelihoods from their underlying articulatory attributes and phones. During decoding, a word will be treated as a sequence of such bundles of phone and articulator models. This architecture is illustrated in figure 3, where two articulatory features were integrated to model the sound */e/* for the word *hello*. Please note that during decoding each state of the HMM is associated to the corresponding AF model although the AF models were trained only at the middle states.

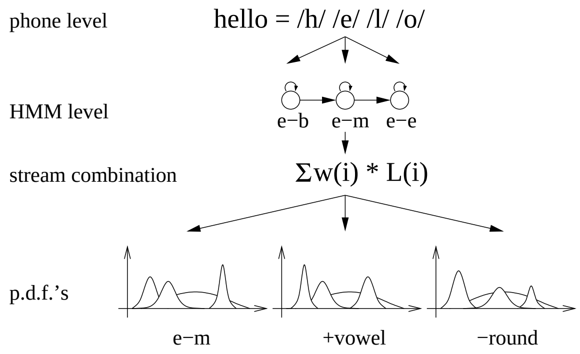

Figure 3: streaming architecture, example for the cascade of models to build the word *Hello*

An advantage of that approach is, that already existing phone based HMM systems can be *enriched* by the AF models without a retraining of the whole system. However, we cannot exploit the asynchrony between different articulatory features since the AF and phone models are rather tight coupled in the stream architecture.

3. EXPERIMENTAL SETUP

3.1. HMM baseline system

The baseline system with a $40k$ vocabulary is trained on different corpora such as broadcast data and english Verbmobil (ESST) data to transcribe colloquial speech recorded at informal group meetings [1]. The acoustic models base on 4000 polyphone states with a total of $132k$ gaussians. Several state of the art normalization and adaptation techniques are used to cover speaking style and channel variations. The front-end consists of mel-filtered cepstral coefficients with a context window of 7 frames followed by a LDA transform to reduce the dimension to 40 and a single semi-tied full covariance. Speaker incremental cepstral mean and variance normalization is used to reduce channel variations.

3.2. Training of the AF models

The AF classifier base on gaussian mixture models using the same front end as used for the baseline system. The models are trained on the ESST data only using the time boundaries of the center states from the phone alignments. Each binary AF classifier has 48 diagonal gaussians. The number of additional parameters to extract the AFs is only a small fraction compared to the phone models.

3.3. Hyperarticulated data

We have collected an 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 confuseable 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.

4. EXPERIMENTS

4.1. Classification of articulatory features

The results of the AF classifier are shown in table 1 and 2 according to place and manner of articulation. Due to the way we collected the hyperarticulated data, the results base on the same set of speakers and words for both speaking styles.

AF	Speaking Style	
	normal	hyper
plosive	92%	89%
nasal	88%	83%
fricative	95%	95%
approximant	88%	86%
stop	91%	88%

Table 1: classification accuracy per frame for manner of articulatory attributes for each speaking style

However, due to expected variations in the pronunciation of hyperarticulated speech, we have a "moving target" problem for the AF references. The results in table 1 for the manner of articulation features don't exhibit large differences across the speaking style. But the situation is completely different if we take a look of the place of articulation features. In particular, the models for the *interdental* and *palatal* attributes seems to fit better for hyperarticulated speech.

AF	Speaking Style	
	normal	hyper
labial	87%	85%
bilabial	90%	85%
interdental	76%	92%
alveolar	69%	63%
palatal	79%	92%
velar	92%	85%

Table 2: classification accuracy per frame for place of articulatory attributes for each speaking style

Since the reference base on phone alignments, worse classification results don't imply necessarily that the classifier works badly. But what we can conclude is that differences occur at certain articulatory features for hyperarticulated speech and some features are more robust against variations in the speaking style [1].

4.2. Decoding Experiments

Some initial experiments are summarized in table 3, where we started with the meeting recognizer and adapted then the transition and acoustic models in order to reduce the speaking rate[2] and channel mismatch using the normally spoken part of the training data.

system	Speaking Style	
	normal	hyper
baseline	33.4%	49.2%
+ trained transition models	32.7%	46.3%
+ adapted acoustic models	18.9%	29.9%

Table 3: initial experiments for normal and hyperarticulated speech (results in word error rates)

Results obtained by using the adapted models will serve as the baseline for the AF system. Hyperarticulation cause a drastic error increase of more than 58% compared to the normally spoken utterances. Interesting is also that the re-training of the transition model has much more influence of the hyperarticulation part, which indicate a stronger mismatch in the phone durations.

For the system using AF streams we picked the most robust features, mainly concerning manner of articulation:

[1] The variation of the articulation features might also depend on the type of disambiguation in context of error repairs.

[2] The meeting system is trained on spontaneous data while this data is isolated speech.

system	Speaking Style	
	normal	hyper
adapted acoustic models	18.9%	29.9%
+ articulatory features	17.2%	22.4%

Table 4: HMM models supported by streams of articulatory features (results in word error rates)

plosive, *fricative*, *lateral*, *approximant*, *stop*, *consonantal*. The results in table 4 are very encouraging. We got an error reduction of 7.5% absolute for the hyperarticulated speaking style. Additionally, these models don't hurt normal speech, in opposite there is also an error reduction of 1.7% absolute, which demonstrate the potential of articulatory features.

4.3. Specialized models

In a further series of experiments, we investigated the use of hyperarticulated training data by generating specialized models for normal and hyperarticulated speech by computing MLLR regression classes. During decoding, the approriate set of acoustic models will be chosen based on a likelihood criterion. If we don't use the AF streams, we got an improvement from 29.9% to 27.9% on hyperarticulated data. Using the AF streams, the gain of using hyperarticulated training data is only 0.5% (from 22.4% to 21.9%).

system	Speaking Style	
	normal	hyper
specialized models	18.5%	27.9%
+ articulatory features	16.7%	21.9%

Table 5: specialized acoustic models for hyperarticulated speech (results in word error rates)

5. CONCLUSIONS

We investigated the integration of articulatory features in state of the art phone based HMM recognizer in order to compensate hyperarticulated effects. The results indicate that the extraction of articulated features using gaussian mixture models is robust against speaking style variations for features according to manner of articulation. Features related to the place of articulation exhibit differences at hyperarticulation, in particular the models for *interdental* and *palatal*, which are trained on spontaneous speech, seems to fit better for hyperarticulated speech. We observed a drastic error increase of more than 58% under hyperarticulation for the pure phone based system. By incorporating streams of articulatory features, we reduced the error rate from 29.9% to 22.4% for hyperarticulated speech while improving the recognition of normal speech as well (from 18.9% to 17.2%).

In the future, we will focus on more sophisticated models which will also capture the asynchrony of articulatory features and examine how hyperarticulated effects can be modeled by a kind of trajectory variation in an articulatory feature space.

6. REFERENCES

[1] Alex Waibel et.al. Advances in meeting recogniton. In *Proceedings of the First International Conference on Human Language Technology Conference (HLT 2001)*, San Diego, USA, 2001.

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

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

[5] K. Kirchhoff, G. Fink, and G. Sagerer. Conversational speech recognition using acoustic and articulatory features. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

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

[7] 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.

[8] I. Rogina and A. Waibel. Learning state-dependent stream weights for multi-codebook hmm speech recognition systems. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Adelaide, Australia, 1994.

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

[10] 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.

[11] H. Soltau and A. Waibel. Acoustic models for hyperarticulated speech. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000.

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

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

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