INTEGRATING DYNAMIC SPEECH MODALITIES INTO CONTEXT DECISION TREES

Christian Fügen, Ivica Rogina

Interactive Systems Labs, University of Karlsruhe, Germany
e-mail: {fuegen,rogina}@ira.uka.de

ABSTRACT

Context decision trees are widely used in the speech recognition community. Besides questions about phonetic classes of a phone's context, questions about their position within a word [Lee88] and questions about the gender of the current speaker [RC99] have been used so far. In this paper we additionally incorporate questions about current modalities of the spoken utterance like the speaker's dialect, the speaking rate, the signal to noise ratio, the latter two of which may change while speaking one utterance. We present a framework that treats all these modalities in a uniform way. Experiments with the Janus speech recognizer have produced error rate reductions of up to 10% when compared to systems that do not use modality questions.

1. INTRODUCTION

1.1. Context Decision Trees in Janus

As described in [FR97] and [Rog97], Janus uses decision trees [Ode92] to assign acoustic models to polyphone segments. The base algorithm of the decoder is described in [Wos98]. Like many other decoders, Janus organizes the vocabulary in a pronunciation-based tree where words with the same prefixes share a common part of that tree. The tree is computed only once before decoding and it is not feasible to modify the tree on the fly while decoding an utterance. This is no problem for static features of phonemes like e.g. the word-position. Other modalities of speech, like the gender of the current speaker or the speaker's dialect, can be handled by using specialized decision trees or, which is effectively the same, by even training speaker or dialect-dependent recognizers. This is significantly more difficult for features that are not constant during a decoding. In order to not have to reorganize the decoder-tree every time a speech modality changes, we transform the decision tree (figure 1) by shifting questions about dynamic features down to the leaves of the tree and eventually replace every leaf node by a lookup table that contains a reference to an acoustic model for every combination of speech modalities (figure 2). Although this transformation is just a technicality, it was necessary to allow the decoder to run in near-real-time.

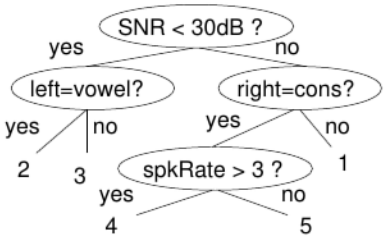

Figure 1: Original Decision Tree

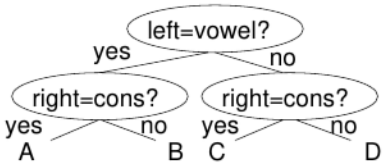

Figure 2: Transformed Decision Tree

SNR < 30 ?	yes		no	
spkRate > 3 ?	yes	no	yes	no
A	2	2	4	5
B	2	2	1	1
C	3	3	4	5
D	3	3	1	1

In [Füg99], we describe the transformation process in detail. The major advantage of incorporating modality questions into the decision tree over training separate recognizers for different modalities, is the possibility to make a data-driven decision for different areas of the acoustic parameter space, whether to share parameters or to use separate models.

In this work we continue a research presented in [FW97], in which word pronunciations were chosen depending on the current speaking mode, by adding the paradigm of effective parameter sharing.

2. MODALITIES

In the following, we will describe the modalities that we have investigated during the work presented in this paper.

2.1. Word Position

Word-position-dependent acoustic models have been used since [Lee88]. These attributes are constant during a decoding and even over several decoding passes with the same recognizer. We have found that distinguishing between word-boundary and in-word positions gives the greatest improvement. Additionally distinguishing between word-beginning and word-end positions does not improve the recognition accuracy significantly.

2.2. Gender

It has been shown [PGF+95] that gender-dependent recognizers can outperform gender-independent recognizers when there is enough training data for each gender to train a reliable acoustic model. But even in these cases, many acoustic units, esp. the non-sonorant units, need not to be modeled gender-dependently.

2.3. Dialect

Since even linguists often don't agree on the exact definitions of what is an accent, a dialect, or a different language, we decided to run a data-driven agglomerative clustering algorithm that merged different geographic areas which the speakers of our database declared to be their home into dialect classes. For every elementary region we trained a Gaussian mixture model and used an entropy distance measure to compute distances between the acoustic appearance of different regions or region classes.

The initial situation with 17 elementary dialect regions [BO99] was clustered into three main regions (North, Mid, South) and is displayed in figure 3.

We then took the most likely nine intermediate classes of regions that were created during the clustering process as the basis for a question that was available to the recognizer's decision tree.

2.4. Signal to Noise Ratio

We expected that the decision tree building process should be able to create more specific models for very clean speech while noisy speech could be handled with fewer models. Since we can only ask binary questions in our decision trees, we decided to split the continuous

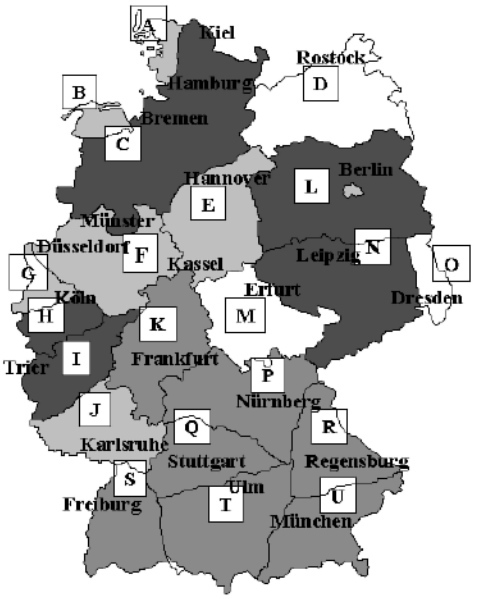

Figure 3: Clustered dialect regions: NorthE (C, L, N), North (NorthE, H, I), MidW (F, G), Mid (MidW, A, B, E, J, Berlin), SouthS (Q, S, T, U), SouthW (SouthS, P, R), South (SouthW, K), MidSouth (Mid, South) and Bavaria. D, O, M not available

SNR values into 15 intervals and only ask questions like "is the SNR within interval x?"

2.5. Speaking Rate

The speaking rate, too, is a continuous parameter which we handled by defining a set of intervals. We used the MRATE [MFL98] algorithm to compute the current speaking rate which could change in the middle of an utterance.

3. EXPERIMENTS

Using modality questions in a decision tree has some advantages over using separate recognizers for different values of modalities. While information about the within-word-position of a phone can easily be obtained during the decoding, this is not the case for speaker-related information like the speaker's gender. Many speech databases, however, do have a sufficient amount of training data for both genders such that training gender-dependent systems is justified.

A problem arises when we add other modalities. The combinatorial explosion of possible combinations of genders, speaking rates, signal qualities, and dialect-

regions makes it infeasible to train a separate acoustic model for each combination.

The best solution to this problem is to let the training process itself decide on which combinations have both, enough training data and enough uniqueness to be usefully modeled separately, or actually which sets of combinations should be modeled as one acoustic model.

3.1. The Test Environment

As a test environment we used a German spontaneous human-to-human dialog database [WE92] which was used for evaluations of German spontaneous speech recognizer in which the Janus system scored best among several participants from Germany.

The database contains 35 minutes of speech from speakers from different German regions with different dialects. The speakers were asked not to speak dialect, but still, here and there some speaker slipped a dialect word, and in most cases, speakers do have a very noticeable accent sufficient for humans to identify their home region. We expected a recognizer that could model accents separately, to perform better than one that uses the same models for all accents.

Our recognizer was developed with the Janus Recognizer Toolkit. It computes 32 LDA coefficients from 13 per-utterance mean-subtracted mel cepstral, delta, and deltadelta coefficients. Each polyphone is modeled with three HMM-states. Every system evaluated in this paper has approx. 2000 simultaneously usable acoustic models.

3.2. Results

Gender

To verify the hypothesis that a data-driven creation of a decision tree using modality questions can outperform both, the no-modality gender-independent system and the no-modality gender-dependent system, we trained three recognizers: GI is the standard baseline gender-independent recognizer, GD is its gender-dependent counterpart, and Mod is the system using modality questions. In all cases the size of the parameter space was determined automatically by the clustering algorithm [Rog97] and yielded systems of approximately the same number of parameters. Figure 4 illustrates the word error rates. When not applying a cepstral mean subtraction, the total error reduction of the modalities system becomes even more than 10%.

Dialects

Figure 5 shows the error rate improvements for different dialects. We observed a consistent improvement for each dialect. The overal error rate was reduced from 28.2% to 26.7%.

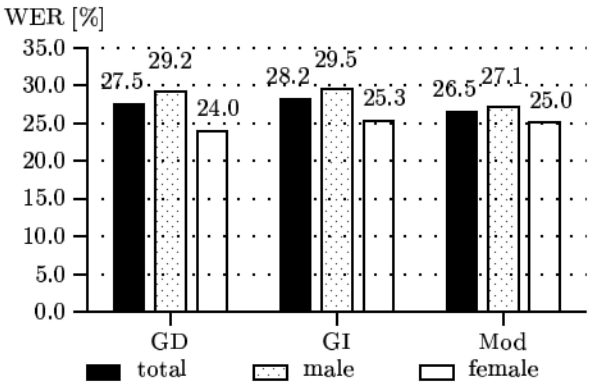

Figure 4: Modality-system compared to gender-(in)dependent systems

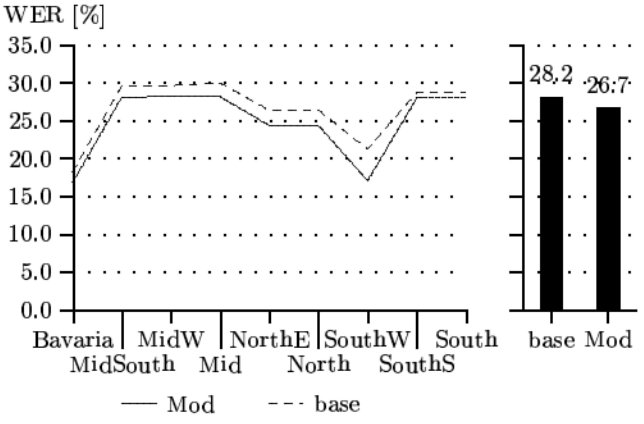

Figure 5: WER reductions for different dialect regions

Signal-to-Noise-Ratio

The experiments with the SNR modalities gave only small improvements as shown in figure 6. We didn't expect a significantly better improvement because the database consists of mainly clean speech recordings without any major noises.

Speaking Rate

The improvement with the speaking rate modality (Figure 7) was not as consistent over all speaking rates as with the dialects but the total error was reduced, too, from 28.2% to 27.3%.

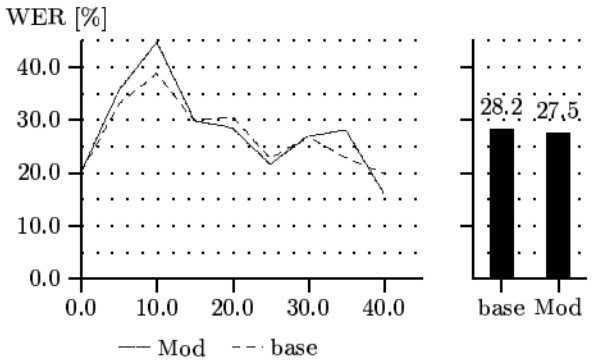

Figure 6: WER reductions for different SNRs

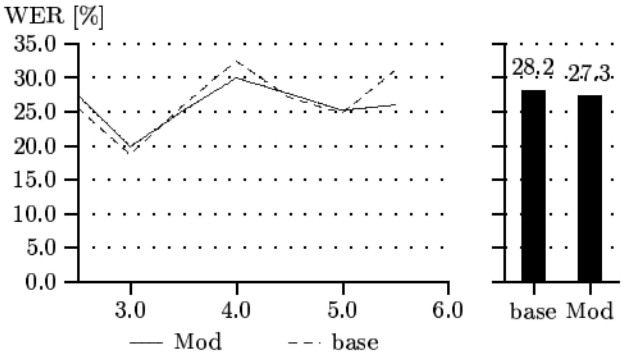

Figure 7: WER reductions for different speaking rates

4. CONCLUSION AND OUTLOOK

We have presented an algorithm that allows to use both, static (word-position, gender, dialect) and dynamic (speaking rate, signal-to-noise ratio) speech modalities in acoustic model decision trees in a uniform way.

We have trained recognizers with speech modalities and compared them to recognizers without these modalities and got a consistent improvement for all modalities.

We expect this approach to be useful not only for decision trees that find appropriate acoustic models but also for other parameters that are commonly used in the speech recognition process like for example the weighting factor of the language model probability vs. the acoustic model probability or word insertion penalties.

Acknwledgement

We would like to express special thanks to Michael Finke for his fruitful discussions and for providing us with the idea for the presented work.

5. REFERENCES

[BO99] **S. Burger and D. Opperman.** Regional Variants of German: Categories of Pronunciation Deviation from Standard German. In *Proceedings of the ICPhS*, 1999.

[FR97] **M. Finke and I. Rogina.** Wide Context Acoustic Modeling in Read vs. Spontaneous Speech. In *International Conference on Acoustics, Speech, and Signal Processing*, Band 3, S. 1743–1746. IEEE, April 1997, München.

[Füg99] **C. Fügen.** Integration von situationsabhängigen Modalitäten in Kontextentscheidungsbäume. Diploma Thesis, University of Karlsruhe, Germany, November 1999.

[FW97] **M. Finke and A. Waibel.** Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition. 1997.

[Lee88] **K. Lee.** *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System.* CMU-CS-88-148, Carnegie Mellon University, Pittsburgh, PA, April 1988.

[MFL98] **N. Morgan and E. Fosler-Lussier.** Combining Multiple Estimators of Speaking Rate. In *International Conference on Acoustics, Speech, and Signal Processing.* IEEE, 1998.

[Ode92] **J. J. Odell.** The Use of Decision Trees with Context Sensitive Phoneme Modelling. Diploma Thesis, Department of Engineering, Cambridge University, Cambridge, UK, August 1992.

[PGF+95] **D. S. Pallett, F. J. G., W. M. Fisher, J. S. Garofolo, B. A. Lund, A. Marin and M. A. Przybocki.** 1994 Benchmark Tests for the ARPA Spoken Language Programm. January 1995.

[RC99] **W. Reichl and W. Chou.** A Unified Approach of Incorporating general Features in Decision Tree Based Acoustic Modeling. In *International Conference on Acoustics, Speech, and Signal Processing.* IEEE, 1999.

[Rog97] **I. Rogina.** Automatic Architecture Design by Likelihood-Based Context Clustering with Crossvalidation. In *Eurospeech*, September 1997, Rhodes.

[WE92] **W. Wahlster and J. Egelkamp.** Wissenschaftliche Ziele und Netzpläne für das VERBMOBIL Projekt. DFKI Saarbrücken, April 1992.

[Wos98] **M. Woszczyna.** *Fast Speaker Independent Large Vocabulary Continuous Speech Recognition.* PhD Thesis, University of Karlsruhe, Germany, 1998.