THE JANUS SPEECH RECOGNIZER

Ivica Rogina, Alex Waibel
Interactive Systems Labs
University of Karlsruhe, Postfach 6980, 76128 Karslruhe, Germany

ABSTRACT

JANUS [17] was designed for the translation of spontaneous human-to-human speech. Before the 1994 CSR evaluation, JANUS was run with vocabularies of up to 2500 words. JANUS was also tested on the Conference Registration and the Resource Management tasks. The best error rate on the '89 Resource Management evaluation set was 5.9%. At the June 1994 Verbmobil speech component evaluation [1], JANUS scored best among eight participants on the German appointment scheduling task, a task of spontaneous human to human dialogs. In this paper we give a detailed description of the recognition engine of JANUS, focusing on the acoustic modeling and our first run with the WSJ task.

1. ACOUSTIC MODELING IN JANUS

1.1 PREPROCESSING

For the 1994 CSR evaluation we computed 16 mel scale spectral coefficients from an FFT with a window size of 256 sample points and a window shift (frame rate) of 10 ms. 16 mel spectral coefficients, 16 delta coefficients, and 16 delta-delta coefficients were used to build a 48 dimensional feature space which is then reduced to 16 LDA coefficients by a linear discriminant analysis (LDA). The LDA was computed for 150 classes (submonophones). No noise or channel compensation or speech enhancement techniques were used for the experiments reported here.

1.2 ARCHITECTURE

In a general HMM system, the emission probability for observing a speech vector $X = (x_1, .., x_n)$ given that the system is in the HMM state s can be defined as $p(x|s) = \sum_{i=1}^{n_s} c_{s,i} \cdot N(\mu_{s,i}, \Sigma_{s,i}, x)$ where N is a Gaussian densitiy and $c_{s,i}$ is the mixture weight for the i-th Gaussian in the mixture that is modeling state s. The acoustic engine of JANUS allows any degree of parameter tying, from simple discrete HMMs, over semicontinuous HMMs optionally phonetically or subphonetically tied, to fully continuous HMM with clustered senones. In the architecture configuration of JANUS we can tie different mixtures or different sets of mixture weights or complete acoustic models. By tying $c_{s_1,i}$ with $c_{s_2,i}$ for any pair of states s_1, s_2 whose mixtures are of the same size, we can smoothly modify the number of mixture weights and thus the number of acoustic models. By tying $\mu_{s_1,i}, \Sigma_{s_1,i}$ with $\mu_{s_2,i}, \Sigma_{s_2,i}$ for any pair of mixtures of the same size, we can smoothly interpolate between semicontinuous HMMs and fully continuous HMMs. So, JANUS allows the emulation of senones [2], genones [3], PICs and PELs [4]. For the 1994 evaluation, our bootstrap system used 50 context independent phonemes plus one silence phoneme and one general garbage phoneme. The only purpose of the garbage phoneme was to act as a 'garbage can' for sequences of the training set that were transcribed with noises or with incorrectly pronounced words. It was not used during testing. This system was later expanded to a system with 2885 context dependent subphones. Every phoneme is modeled with three simple state models, each of which has one transition to the successor state and one self loop.

All transitions have the same transition probability, thus they have no effect on the alignment path, neither in the forced alignment training nor during decoding. No explicite duration modeling is done, and HMM emission probabilities are the only factor to influence the durations of phonemes and words.

1.3 TRAINING

The default training procedure is as follows:

- Create labels for a given database, using an existing recognizer that was bootstrapped on previous databases (sometimes even foreign databases, if necessary). For this evaluation we used the male portion of Resource Management database.
- Create a context independent continuous density HMM from the labeled data with the k-means algorithm. Typically, we start with 150 mixtures (50 phonemes times 3 states) of 16 Gaussians each.
- Train this system with Viterbi training until no further improvement on a crossvalidation set can be achieved.
- Compute a linear discriminant analysis matrix with the best context independent system, and create new mixtures based on the LDA-preprocessed feature space.
- Train the LDA system with Viterbi training until no further improvement on a crossvalidation set can be achieved.
- Introduce separate mixture weights for every context of every model, initializing them with the corresponding mixture weights from the context independent system. (We have observed a small advantage of this approach over merging the mixture weigths that fall into the same cluster.) For the CSR evaluation this resulted in about 25000 different distributions.
- Viterbi training lets the mixture weights of different contexts diverge (the Gaussians are still shared by all contexts of the same context independent model).
- Compute one divisive context clustering tree for every context independent model using an entropy distance measure between the mixture weight distributions of different contexts [2],[14]. The evaluation system had 2885 senones.
- Initialize every generated cluster with the corresponding context independent parameters.
- Viterbi training lets the mixture weights of different clusters diverge.
- Introduce separate mixtures (Gaussian means and variances) for every cluster, initialising them with the Gaussians of the corresponding context independent model. For our evaluation system, this means create 2885 mixtures, each of which is initialized with the Gaussians from the corresponding context independent mixture. The mixture weights remain untouched.
- Viterbi training lets the Gaussian parameters of different clusters diverge.

Optionally we can add gender-dependence at any stage of the training process by splitting a system in two, and training each of the two with data from its gender. For lack of time, we did not use gender dependent systems for the 1994 CSR evaluation. Other techniques like deleted interpolation smoothing [5], corrective training [6], cross-word triphone modeling, dictionary learning [7], connectionist nonlinear discrimiant ananlysis [8], learning vector quantization (LVQ-2) [9], and mixture size optimization [10] are available and have been separately explored for our speech translation tasks, but have also not been used in the 1994 CSR evaluation. We have never trained the HMM-transition probabilities. For the evaluation, all transitions were treated equally. We also do not use duration modeling, which, together with the lack of noise modeling, resulted in hypotheses that had very long TH phones to cover noise seqences or breathing noises.

3 THE DECODER IN JANUS

The decoder is a Viterbi style two pass decoder: the first pass is a standard Viterbi search implemented roughly as described in [11]. The second pass is a word-dependent N-best seach [12] using the backtrace information from the first pass for efficient pruning [13]. First and second pass use a bigram language model. The output of the second pass is not a list of hypotheses but a word-graph from which the hypothesis with the best score is extracted using trigrams.

4 EVALUATION ON THE WSJ/NAB TASK

4.1 THE DEVELOPMENT

Our initial context independent system was bootstrapped with 50 monophone models that were trained on 2890 utterances (the male subset) from the resource management database. This system's performance was 80%

word errors on the 1992 si-dev-05 test set. We trained a recognizer with all the training steps that were described in 1.3 using only the SI-84 training set. All architecture decisions were made with this data. We ended up with 2885 context dependent models that performed best on the 1992 si-dev-05 development test set when compared to systems with other numbers of models. Fig. 1 shows the error rates on si-dev-05 at different stages of the training process.
We have observed an error reduction by 13% to 17% when comparing systems that use an LDA feature space with systems that use spectral coefficients only. The biggest improvement resulted from going from semicontinuous HMMs to fully continuous HMMs, which gave us an error reduction of 30% to 32%. Increasing the training data from SI-84 to SI-284 reduced the error by 17% to 19%. However, all the decisions about the architecture of the recognizer were based on the SI-84 training set, while a larger training set would also suggest a larger parameter space.
After that we did not change the architecture or the number of parameters any more, and continued Viterbi training on the rest of the SI-284 training set, until there was no further improvement on si-dev-05. This was the final system that was used for the official evaluation.

We have observed big differences in performance on different testsets. The followin table shows the word error rates of evaluation system (all tested with the corresponding bigram grammar):

Test set	'92 5k dev-test	'92 20k dev-test	'93 20k dev-test	'94 20k dev-test
Errors	9.3%	13.7%	31.2%	25.2%

These results show that JANUS is not yet robust enough for switching test sets.

4.2 RESULTS WITH THE OFFICIAL EVALUATION SYSTEM

The JANUS speech recognizer had an error rate of 22.8% on ARPA's official 1994 CSR evaluation set on Hub 1, condition C1. Our system did not benefit from the 1994 development test data and grammar file as this data was only received one week before the evaluation test runs. We could only use it to calibrate the weight that balances the contributions of the acoustic and language model scores. A number of feature of the system remain to be optimized on the development corpus.

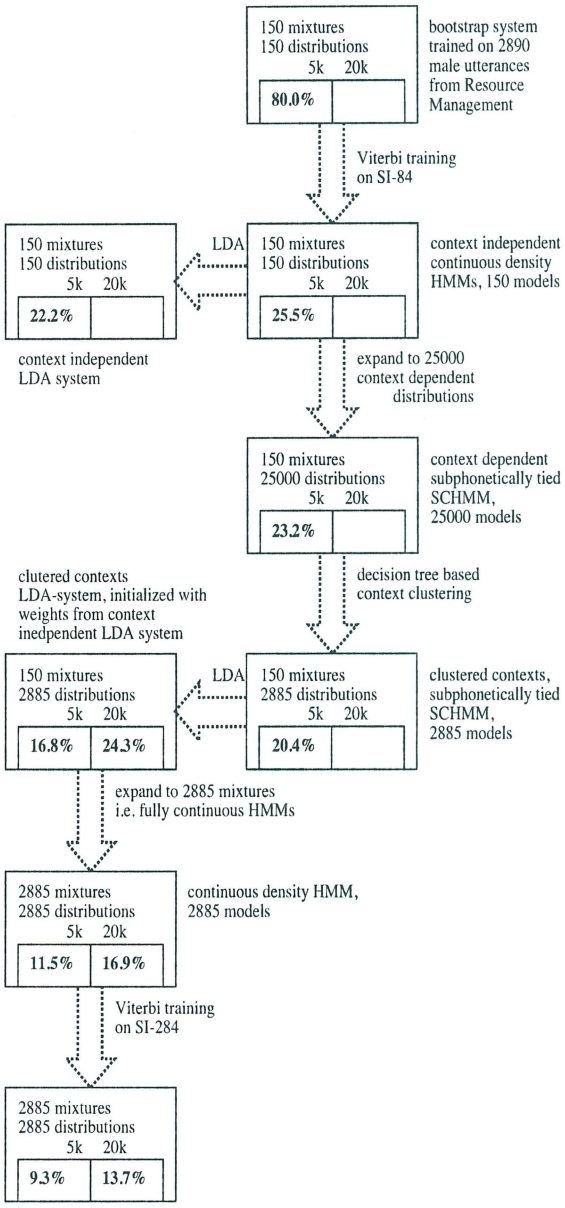

Fig. 1: development tests on the 1992 development test set, using the WSJ0 bigram grammar

4.3 UNOFFICIAL AFTER-EVALUATION RESULTS

About 25% of the SI-84 training set (i.e. about 5% of the SI-284 training set) were corrupted due to NFS problems while unpacking and preprocessing. We did not notice these errors until after the evaluation. We

started a new training which included gender dependent models. At the training stage where the evaluation system had a 16.8% error on the 1992 WSJ development set si-dev-05, the improved system had a word error rate of 14.7%, which is a reduction by about 13%.

5. CONCLUSION AND FUTURE PLANS

The JANUS speech recognizer has proven to give good recognition results as shown on the 1994 Verbmobil evaluation. However, one week with the development data was not enough to tune JANUS to the NAB task. We expect great improvements from successfully applying gender dependent acoustic modeling, optimizing the architecture (context decision trees, number of models, size of mixtures), and channel normalization. The adaptation to a new task is always hard and tricky work. A large amount of fine-tuning work has to be done to reach good performance. Although JANUS scored worst in the 1994 CSR evaluation, we feel optimistic that with some more tuning and the above mentionned techniqes, JANUS will soon compare more favorably.

REFERENCES

[1] Wahlster W., Engelkamp J., "Wissenschaftliche Ziele und Netzpläne für das VERBMOBIL Projekt, DFKI Sarbrücken, April 1992 (in German)

[2] Hwang, M.Y.: "Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition", Ph.D. Thesis, Carnegie Mellon University, 1993

[3] Digalakis V., Murveit H.: "An Algorithm for Optimizing the Degree of Tying in a Large Vocabulary Hidden Markov Model Based Speech Recognizer", ARPA Spoken Language Technology Workshop, March 1993

[4] Scattone F., Baker J., Gillick L., Orloff J., Roth R.: "Dragon's Large Vocabulary Speech Recognition System", ARPA Spoken Language Technology Workshop, March 1993

[5] Jelinek F., Mercer R.L.: "Interpolated Estimation of Markov Source Parameters from Sparse Data", Pattern Recognition Practice, E.S. Gelsema and L.N. Kanal Eds. Amsterdam, Morth-Holland, 1980

[6] Bahl L.R., P.F. Brown, de Sousa P.V., Mercer R.L.: "A New Algorithm for the Estimation of Hidden Markov Model Parameters", ICASSP 1988, pp 493ff

[7] Sloboda T.: "Dictionary Learing: Performance Through Consistency", to appear in: The Proceedings of the ICASSP 1995

[8] Maier M.: "Dimensionalitätsreduktion von Sprachsignalen mit statistischen und neuronalen Methoden", Diploma Thesis, Karlsruhe University, Jan. 1994

[9] Schmidbauer O., Tebelskis J.: "An LVQ based Reference Model for Speaker-Adaptive Speech Recognition", Proceedings of the ICASSP 1992

[10] Kemp T.: "Data-Driven Codebook Adaptation in Phonetically Tied SCHMMs", to appear in: The Proceedings of the ICASSP 1995

[11] Ney, H.: "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition". IASSP 1984, vol. 2, pp 263-271

[12] Schwartz, R. and Austin, S.: "A Comparison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypotheses", ICASSP 1991, vol. 1, pp 701-704

[13] Austin, S. and Schwartz, R. and Placeway, P.: "The Forward–Backward Search Algorithm", ICASSP 1991, vol. 1, pp 697-777

[14] Hon, H.W.: "Vocabulary-Independent Speech Recognition: The VOCIND System", Ph.D. Thesis, Report CMU-CS-92-108, Carnegie Mellon University, March 16, 1992

[15] Schultz T., Rogina I.: "Acoustic And Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition", to appear in: The Proceedings of the ICASSP 1995

[16] Rogina I., Waibel A.: "Learning State-Dependent Stream Weights for Multi-Codebook HMM Speech Recognizers", Proceedings of the ICASSP 1994

[17] Woszczyna M., Aoki-Waibel N., Buo F.D., Coccaro N., Horiguchi K., Kemp T., Lavie A., McNair A., Polzin T., Rogina I., Rose C.P, Schultz T., Suhm B., Tomita M., Waibel A.: "JANUS 93: Towards Spontaneous Speech Translation", Proceedings of the ICASSP 1994