Ensemble machine learning methods for acoustic modeling of speech

Yunxin Zhao

Professor, Department of Computer Science

University of Missouri, Columbia MO, USA

Introduction

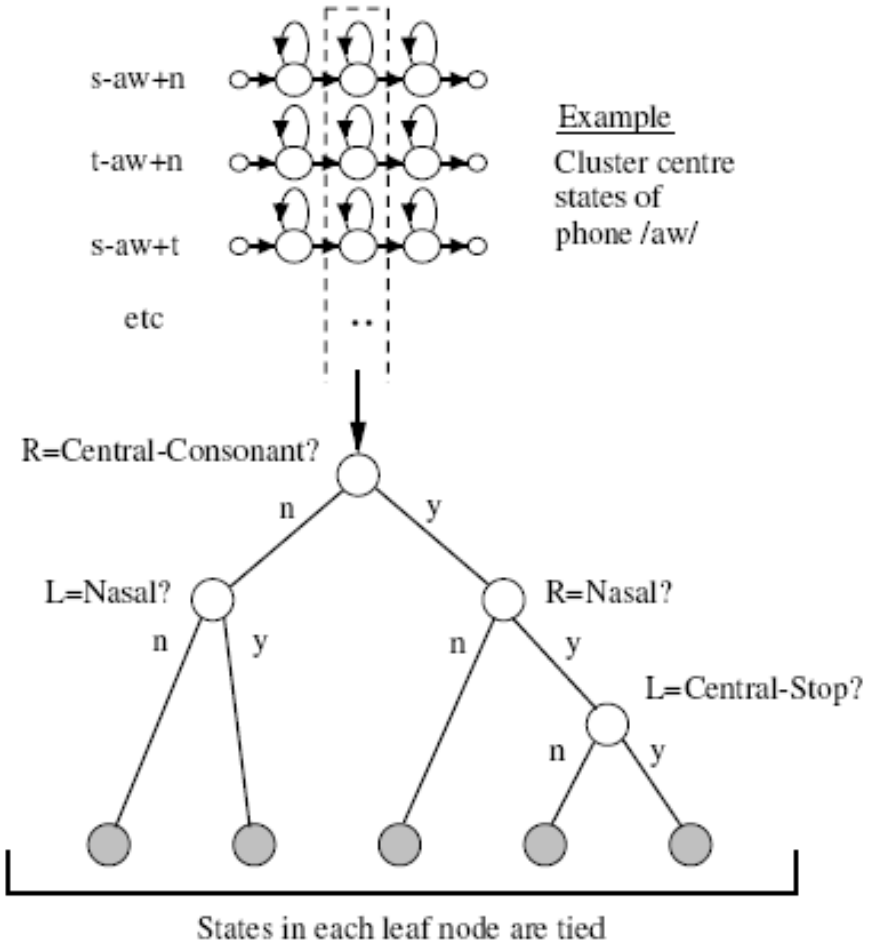
- Automatic speech recognition has unarguably made great advances over the past several decades, but improving the accuracy and robustness of natural, spontaneous speech recognition remains a long standing challenge.
- Speech sounds have large variability, caused by variations in speaking styles, speech rate, speech disfluency, illgrammatical syntax, dialects, speakers' voice characteristics, diverse acoustic conditions, and so on.
- The knowledge sources of acoustic model, lexicon, and language model all need improvements, and among which acoustic modeling plays a crucial role.

Statistical modeling of speech signals as popularized by HTK [1]

Hidden Markov model (HMM)

Phonetic decision tree (PDT)

Data in leave nodes are modeled by Gaussian mixture densities (GMD)



Improving statistical acoustic models

- one model for one speech sound unit

- Increase granular details of conditioning variables beyond the left-right neighboring phones, including
  - longer phonetic contexts (e.g., pentaphones) [2]
  - prosodic features of speech rate, pitch, lexical stress, et al. [3]
- Improve allophonic clustering beyond the phonetic decision trees (PDT), including
  - associating each allophone HMM state with multiple tied states within a PDT (soft clustering) [4]
  - optimize PDT construction by k-step lookahead or stochastic full lookahead [5].
  - apply probabilities of split variables in PDT construction [6].

- Optimize model parameter estimation beyond maximum likelihood estimation by discriminative training, including
  - minimum classification error (MCE) [7]
  - maximum mutual information (MMIE) [8]
  - minimum phone error (MPE) [9]
  - large margin [10]

Randomization for ensemble classifier design

- **Bagging**: uniform random sampling with replacement is applied on a training dataset to produce bootstrap replicated datasets, and from which multiple classifiers are trained and combined through majority voting [11].
- **Boosting**: employ importance-based data sampling to construct multiple classifiers sequentially, with higher importance assigned to data that are more frequently misclassified by the classifiers constructed so far, and importance-based weights are used to combine the classifiers [12].

- **Random subspace**: To construct a tree classifier, first sample *m* split variables randomly out of a total of *M* variables, and then use the sampled variables and a standard greedy algorithm to grow the tree. The tree classifiers are combined by majority voting [13].
- **Random forests**: To construct a tree classifier, first generate a dataset by bagging, and then at each node, select *m* split variables randomly out of a total of *M* variables and choose the best split determined in these *m* variables. The tree classifiers are combined by majority voting [14].

Ensemble classification error

- Correlation and strength of the base classifiers [14]

  Pr(generalization error of the ensemble) $\leq \bar{\rho}(1 - s^2)s^2$

  $\bar{\rho}$ is the pair-wise correlation between the base classifiers,

  S is the strength (accuracy) of the base classifiers.

- Bias-variance-noise decomposition [15]

  Classification error = bias + variance + noise

  Ensemble classifier allows base classifiers to overfit to training data to reduce bias error, and use decision averaging to control variance error.

Improving ASR accuracy

-multiple decoding output integration

- Combine word hypothesis outputs from multiple ASR systems as in ROVER [16].
- Use different optimization criteria, e.g. MLE, MMIE, to train multiple acoustic models from one set of training data, run separate decoding searches, and combine word hypotheses.
- Randomize the construction of PDTs (randomly select a split from the n-best splits at each node) to generate multiple acoustic models from one set of training data, run separate decoding searches, and combine word hypotheses [17].

- Pros

  Integrating multiple decoding outputs improves recognition accuracy when the multiple systems or models are good individually and complementary collectively.

- Cons
  - Combining N decoding outputs requires N times the decoding time of the conventional single system.
  - Cannot effectively exploit the opportunity of reducing the biases of the individual systems or models to reduce classification errors since the word errors are not localized and one error may spawn additional errors down the search path.

Improving statistical acoustic models
- multiple models for each speech sound unit [18]

- Use random forests of phonetic decision trees to generate multiple acoustic models for each allophone-state unit and combine the acoustic model scores for each speech frame.
- The diversity of multiple acoustic models is exploited at local scales, and decoding search is improved in every step to produce more accurate output word hypotheses.
- For each phone-state unit, the individual acoustic models can overfit to training data to drive down bias, since combining the models will control the variance.

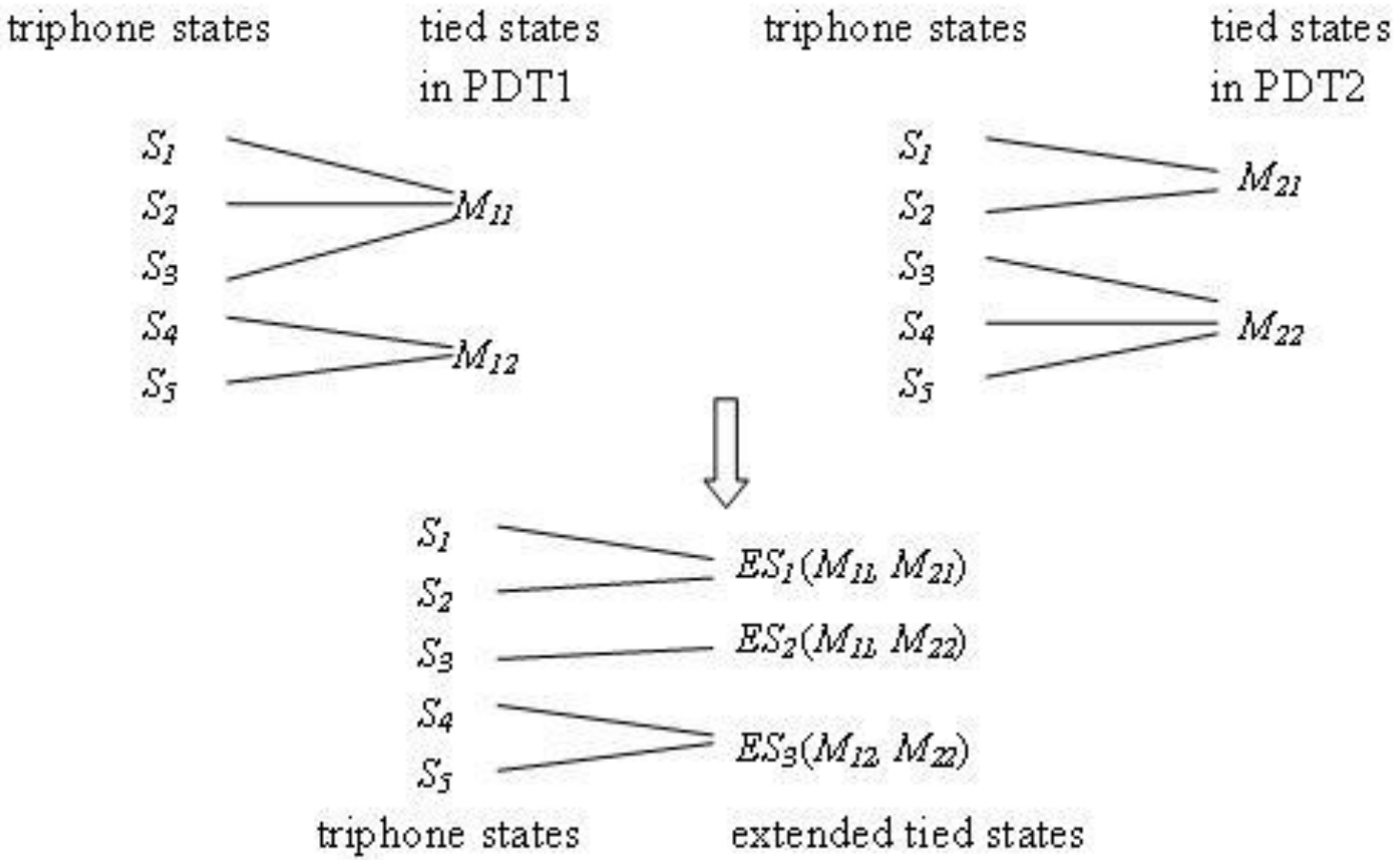
Construction of Random Forests of PDTs

(sampling split variables = sampling phonetic questions)

- Apply multiple sweeps of randomization on split variables.
- In any sweep, randomly sample a subset of *m* phonetic questions and train a set of PDTs as in conventional acoustic model training, producing one set of acoustic models.
- From multiple sweeps of phonetic question sampling, obtain multiple sets of PDTs and hence multiple sets of acoustic models .

- For each phone state, generate RF tied states by tying the triphone states that are in the same tied states in every PDTs of the random forest.
- In decoding search, combine the acoustic scores from the Gaussian mixture densities of the PDT tied states within each RF-tied state.
- Only one decoding search is necessary, and hence computation overhead is small.

Illustration of RF tied states



Acoustic model of a RF tied state:

$$P(\underline{x}_t \mid ES_l) = F(\underline{x}_t \mid M_{l_1 1}, \cdots, M_{l_K K})$$

linearly combination

$$P(\underline{x}_t \mid ES_l) = \sum_{k=1}^{K} w_{tlk}\, p(\underline{x}_t \mid M_{l_k k})$$

where $p(\underline{x}_t \mid M_{l_k k})$'s are the Gaussian mixture density scores of the PDT tied states. The combination weights w_{tlk} can be estimated by different methods.

Combining acoustic scores at each frame

- Non-trainable combiners:
  - Maximum score
  - Average score
  - Average *n*-best score
- Trainable combiners:
  - Maximum likelihood based weight estimation
  - Confidence score based weight estimation
  - Relative entropy based weight estimation

Experiments on telehealth automatic captioning [19]

- Speech features: standard 39 components of MFCCs+Δ+ΔΔ;
- Acoustic models: triphone HMM models trained by mock conversation speech data of five doctors collected in Missouri Telehealth Network;
- Language model: word-class mixture trigram language models trained with both in-domain and out-of-domain datasets;
- Task vocabulary size: 46,480.
- Baseline word accuracy: 78.96%.

Performance vs. forest size and combiner

Word accuracies (%) averaged over five speakers,
GMD size = 16, question subset size *m*=150 for RFs

	Forest size *K*			
	10	20	50	100
MAX	80.35	80.41	79.95	79.70
Average	80.39	80.57	80.71	80.80
MLE	80.47	80.81	80.90	80.92
P-value	80.43	80.69	80.85	80.90
R-entropy	80.39	80.64	80.88	80.91
MLE+R-entropy	80.39	80.72	80.95	80.96

Performance vs. GMD mixture size

Word accuracy vs. GMD mixture size,

question subset m=150 for RFs

	Number of Gaussian components per GMD			
	8	16	20	24
Baseline	77.65	78.96	78.68	78.15
RF method, K=10	78.08	80.47	81.57	81.70
RF method, K=20	78.06	80.81	81.86	81.92

Measure the quality of an acoustic model ensemble:

- The sampling subset size m can control the quality of an acoustic model ensemble.
- Large m -> the individual PDTs are strong and the correlations among the PDTs are high.
- Small m -> the individual PDTs are weak and the correlations among the PDTs are low.
- An appropriate subset size m can help balance the strength and the correlation of the PDTs.

Measure correlations between acoustic models

- Use Viterbi alignment to segment training data into phone states.
- Within each segment, compute the allophone state posterior probability vectors for each input speech frame by the model sets i and j:

$$\underline{p}_i(t) = [P_i(1 \mid \underline{x}_t), \cdots, P_i(N \mid \underline{x}_t)]'$$

$$\underline{p}_j(t) = [P_j(1 \mid \underline{x}_t), \cdots, P_j(N \mid \underline{x}_t)]'$$

Measure correlation between acoustic model sets

- Compute absolute correlations between the model sets i and j for each triphone state n within each phone-state segment u:

$$Corr_{u,n}(i,j) = \left| \frac{\sum_{t \in \Omega_u} [P_i(n \mid \underline{x}_t) - \overline{P}_i(n \mid \underline{x}_t)][P_j(n \mid \underline{x}_t) - \overline{P}_j(n \mid \underline{x}_t)]}{\sqrt{\sum_{t \in \Omega_u} [P_i(n \mid \underline{x}_t) - \overline{P}_i(n \mid \underline{x}_t)]^2 \sum_{t \in \Omega_u} [P_j(n \mid \underline{x}_t) - \overline{P}_j(n \mid \underline{x}_t)]^2}} \right|$$

- Average $Corr_{u,n}(i,j)$ over n and u to produce the correlation between the model sets i and j.

Word accuracy vs. model correlation and question subset size *m* (one speaker)

Total questions = 216, forest size $K = 50$,
MLE was used for weight estimation.

Subset size *m*	15	20	100	150	200	210
Correlations	0.74	0.76	0.79	0.82	0.89	0.93
Word accuracy (%)	76.35	78.69	79.00	79.00	79.00	78.36

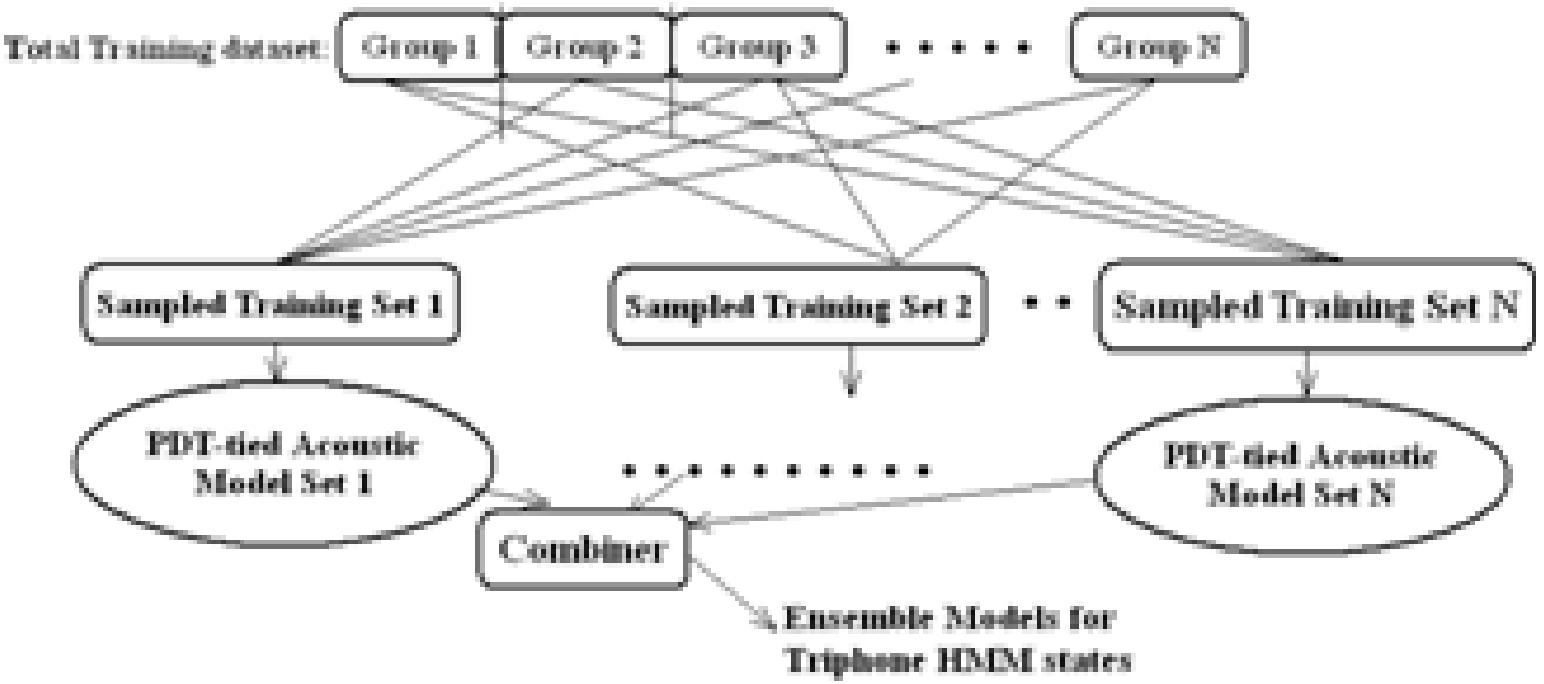
Performance vs. question subset size *m* (five speakers)

Total questions = 216, forest size K = 50,
MLE was used for weight estimation.

Subset size *m*	15	20	100	150	200	210
Word accuracy (%)	77.68	80.38	80.92	80.96	80.90	80.65

Random Forests for PDTs

— sampling data by N-fold CV partitions [20]



Performance vs. CV partition size (five speakers)

GMD mixture size = 16, uniform combination weights

CV folds	1	5	10	20
Word accuracy (%)	79.24	80.88	81.47	81.36

Future directions:

- Incorporate more split variables for PDT constructions to improve the strength of the individual models and reduce the correlations among them.
- Investigate additional sampling methods and their combinations for different types of training data to enhance ensemble diversity.
- Combine ensemble acoustic modeling with diversification methods in features [21], model parameter estimation [22], and language modeling [23].
- Investigate issues that have or have not been studied for the conventional acoustic models in the paradigm of ensemble acoustic modeling (e.g. speaker adaptation).
- Efficiently compute Gaussian density scores for real-time decoding search.

References

[1] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling, " *Proc. ARPA Human Language Technology Workshop*, pp. 307–312, 1994.

[2] T. Pfau, M. Beham, W. Reichl, and G. Ruske, "Creating large subword units for speech recognition," *Proc. EUROSPEECH*, pp. 1191–1194, Rhodos, Greece, 1997.

[3] E. Shriberg and A. Stolcke, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, Vol. 32, No. 1-2, 2000.

[4] X. Luo and F. Jelinek, "Probabilistic classification of hmm states for large vocabulary continuous speech recognition," *Proc. ICASSP*, pp. 353-356, 1999.

[5] J. Xue and Y. Zhao, "Novel lookahead decision tree state tying for acoustic modeling," *Proc. of ICASSP ,* pp. IV-1133-1136 , Honolulu, Hawaii, April 2007.

[6] R.-S. Hu and Y. Zhao, "Knowledge-based adaptive decision tree state tying for conversational speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 15, no. 5, pp. 2160-2168, Sept. 2007.

[7] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Proc*essing, vol. 5, no. 3, pp. 257-265, 1997.

[8] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer, Speech and Language,* vol. 16, No 1, pp. 25-47, Jan. 2002

[9] Povey, D. Woodland, P.C. "Minimum phone error and I-smoothing for improved discriminative training,**"** *Proc. ICASSP*, vol. 1, pp. 105-108, 2002.

[10] Y. Dong, L. Deng, X. He, A. Acero, "Large-Margin Minimum Classification Error Training for Large-Scale Speech Recognition Tasks**,"** *Proc. ICASSP*, vol. IV, pp. 1137-1140, 2007.

[11] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[12] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

[13] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832--844, 1998.

[14] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp.5-32, 2001.

[15] R. Kohavi and D. H. Wolpert, "Bias plus variance decomposition for zero-one loss functions," *Proc. 13th ICML*, pp. 275-283, 1996.

[16] J. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER), " *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding,* 1997.

[17] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," *Proc. ICASSP*, pp. I-197 – I-200, 2005.

[18] J. Xue and Y. Zhao, "Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing,* vol. 16, no. 3, pp. 519-528, March 2008.

[19] Y. Zhao, X. Zhang, R-S. Hu, J. Xue, X. Li, L. Che, R. Hu, and L. Schopp, "An automatic captioning system for telemedicine," *Proc. ICASSP*, pp. I-957 – I-960, 2006.

[20] X. Chen and Y. Zhao, "Data sampling based ensemble acoustic modeling," to appear in *ICASSP* 2009, Taipei, Taiwan.

[21] Q. Zhu, A. Stolcke, B.Y. Chen, N. Morgan, "Using MLP features in SRI's conversational speech recognition system," *Proc. Interspeech*, pp. 2141–2144, Lisbon, Portugal, 2005.

[22] T. Shinozaki, M. Ostendorf, "Cross-validation EM training for robust parameter estimation," *Proc. ICASSP*, vol. IV, pp. 437–440, 2007.

[23] P. Xu, and F. Jelinek, "Random forests in language modeling," *Proc. of EMNLP*, Barcelona, Spain, July, 2004.