Recognition of Speech with Non-random Attributes

Lukáš Burget[1,2], Jan Černocký[1]

[1] Faculty of Information Technology, Brno University of Technology
Božetěchova 2, Brno, 612 66, Czech Republic
[2] OGI School of Science & Engineering at Oregon Health & Science University
20000 NW Walker Road, Beaverton, OR, 97006, USA
{burget,cernocky}@fit.vutbr.cz

Abstract. Most of current speech recognition systems are based on Hidden Markov Models assuming that speech features are sequence of stationary stochastic processes. However, there are certain speech attributes, such as background noise type or speaker voice color, that do not have stochastic character. This fact is often ignored, by designers of robust speaker independent recognition system. In this work, we investigate how the performance of a noisy speech recognition can be improved provided that we have prior knowledge about type and level of noise. Next, recognizer that is using separate models, each trained on a particular type and level of noise, is proposed for more appropriate modeling of speech.

1 Introduction

For last two decades, usage of methods based on the statistical models and specially on Hidden Markov Models (HMM) [1] [2] have dominated in speech recognition. Each HMM specifies probability density function of speech feature vector sequence for a given word (or subword). Each HMM state represents some part of a word, for which the feature vector sequence can be considered as a stationary stochastic process. Each state then holds parameters of an output probability density function of feature vectors related to the part of the word. Transitions between states specify changes in statistical properties of features in time for a given word. For continuous speech recognition, transitions between models (we will call these transitions recognition network) specify possible words sequences. In other words, when using HMM, we consider speech to be sequence of stationary stochastic processes, where individual observations are statistically independent, only obeying distribution given by related HMM states. However, observations are not independent in the real speech and, moreover, there are processes or speech attributes that cannot be considered stochastic at all.

When building a robust speaker independent recognition system, it is usual practice, that parameters of each model are estimated from training data containing speech of many speakers, corrupted by noises of different types and SNR

levels. Figure 1. shows recognition network of such system used for our experiments. HMM's with usual left-to-right topology, estimated this way, assume that speech attributes, such as type and level of background noise or voice color and accent of speaker, have a random character and that they can be randomly changed between speech observation. This is obviously not true. Moreover, we often cannot estimate even prior probabilities of these attributes, since we have no prior knowledge of the environment, in which the recognizer will work and who will be its user. Therefore, these attributes cannot be even considered as having stochastic character.

In this work, first, we investigate how the recognition of noisy speech can be improved provided that we have prior knowledge about type of noise and level of SNR. Individual recognizers are trained each on data corrupted by one particular type and level of noise. Since we have the knowledge about these attributes of noise in the test data, each test utterance can be recognized by the recognizer trained on matched training data. Note, that we still assume that a part of speech related to one HMM state is stationary stochastic process. However, its probability distribution depends on the non-stochastic attributes. In our experiments we assume that only non-stochastic attribute is the noisy condition.

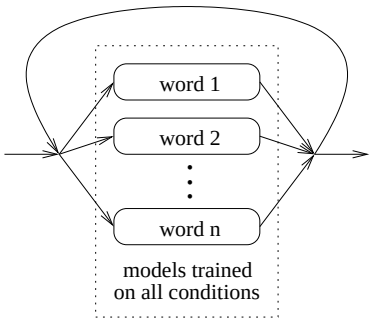

Fig. 1. Recognition network used for baseline system.

For real application, first, we need to recognize what is the noisy condition of input utterance. This can be done by evaluating all recognizers trained under different noisy conditions and selecting that one with the highest total likelihood. In this case, we consider noisy condition being constant during one utterance and all noisy conditions having equal prior probabilities for an utterance. This recognition scheme can be regarded as one recognizer with recognition network shown in figure 2.

Table 1. Baseline system.

Condition	Clean	Seen noises			Unseen noises		
SNR level [dB]	-	20	10	00	20	10	00
No. of Gauss.							
1	96.96	97.40	93.33	42.58	97.34	94.55	57.10
2	98.17	97.86	94.68	47.29	98.11	95.61	60.27
3	98.63	98.25	95.26	48.82	98.42	96.23	61.49
5	98.89	98.69	95.86	**49.67**	98.72	96.44	**62.11**
8	98.86	98.73	**95.97**	48.67	98.68	96.31	59.60
11	98.99	**98.78**	95.76	48.61	**98.78**	96.44	59.45
17	**99.12**	98.76	95.94	48.92	98.77	**96.56**	60.12

2 Experimental setup

Speech data from TI Connected Digits database [4] were used for both training and testing in all experiments. Limited number of clean speech utterances were selected for training (616 utterances from 4 male and 4 female speakers) to be able to find the point where models become overtrained when increasing number of model parameters. Four types of noise (subway, car, exhibition, babble) from AURORA2 TI Digits database [5] were artificially added to speech data with SNR 20 dB and 10 dB. The same 616 utterances were used to create data for all noisy conditions. Together 616 * (1 + 4 * 2) = 5544 utterances were used for training.

Test data were prepared in a similar manner. Here, 912 utterances from 12 male and 12 female speakers were used, 4 noises used for training and four unseen noises (train station, airport, restaurant, street) were added to test data. Additionally, SNR 0 dB condition was generated for both seen and unseen noises. Together 912 * (1 + 8 * 3) = 22800 utterances were used for testing.

As speech features, 15 Mel Frequency Cepstral Coefficients [3] augmented with their first and second order derivatives (delta, double-delta) were used for all experiments (23 bands in Mel filter bank, 25 ms window length, 10 ms frame rate). Continuous HMMs were used for all experiments with output probability density function modeled by Gaussian mixture. Whole word models with left-to-right topology (16 states for digits, 3 states for silence) were used. Recognition networks for whole recognizers (transitions between word models) used for individual experiments are shown in figures 1, 2, 3.

3 Baseline system

As mentioned above, the usual practice is that parameters of each model are estimated from all training data containing speech with different noisy conditions. The recognition network of such recognizer used for our *baseline system* can be seen in figure 1. Table 1 shows recognition accuracies of this system for different numbers of model parameters (number of Gaussians in Gaussian mixture) and

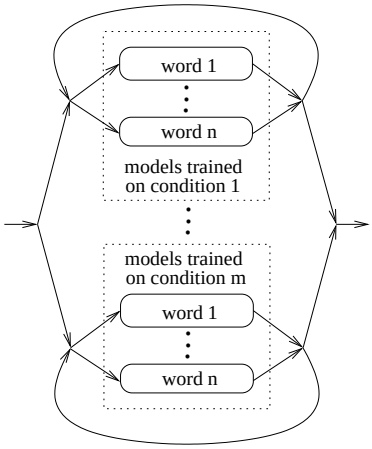

Fig. 2. Recognition network used highest likelihood system.

for different levels of SNR for both the seen and the unseen conditions. All values in the table (except those for clean condition) are averaged accuracies of four seen or four unseen types of noise. In average, the best results were obtained for 17 Gaussians per state, however, for the most noisy conditions the best result (bold values in table 1) was obtained for other number of Gaussians. Further increasing of number of Gaussians caused degradation in recognition performance as models became overtrained. Accuracies for unseen noisy condition are almost always higher than corresponding accuracies for seen conditions. This fact is not so surprising if we realize that it is extremely difficult to recognize speech corrupted by noises such as babble noise or car noise, which are present in seen conditions.

4 Recognition using model trained on matched condition

For the second experiment, we have used 9 recognizers each trained on one noisy condition. Each test utterance was then recognized by system trained on corresponding condition. This is of course cheating, since we have provided to the system the correct information about noisy condition. For conditions with SNR 0 dB, which are not used for training, recognizer trained on data with the corresponding type of the noise and SNR 10 dB was selected as the recognizer trained on the closest condition. We will refer to this system as to *matched system*. Accuracies for this experiment can be seen in table 2. Obviously, there are no accuracies, in the table, for the unseen conditions, since, for these conditions, we have no corresponding recognizer trained.

Individual systems usually reach the best performance with less Gaussians (5 Gaussians for clean condition, 1 for 0 SNR) compared to *baseline system* (17 Gaussians for clean condition, 5 for 0 SNR). For the *baseline system*, we can not take the advantage of the fact that, for certain condition, the system performs

Table 2. Matched system.

Condition	Clean	Seen noises		
SNR level [dB]	-	**20**	**10**	**00**
No. of Gauss.				
1	98.47	98.06	95.30	**70.66**
2	98.89	98.70	96.30	70.40
3	99.05	98.97	96.56	68.80
5	**99.09**	**99.07**	96.64	65.83
8	99.09	98.95	**96.75**	62.18
11	98.27	98.85	96.58	58.80
17	94.06	98.68	96.11	54.38

Table 3. Optimal matched system.

Condition	Clean	Seen noises		
SNR level [dB]	-	**20**	**10**	**00**
Accuracy [%]	99.09	99.07	96.75	70.66
WER Imp. [%]	-3.41	25.00	17.24	42.56

best with certain number of Gaussians. There is only one system with chosen number of Gaussians used for recognition of utterances of all conditions. For our *matched system*, however, each recognizer, trained for particular condition, can use different number of Gaussians, such, for which it performs the best. Regarding this, we can pick up the best accuracy for each condition from table 2 (bold values) as accuracies of *matched system* with optimal number of Gaussian in each recognizer. Accuracies for this "optimal" system can be also seen in table 3 together with relative word error rate improvement over the best *baseline system* (system with 17 Gaussians). We did not observe any improvement for clean condition (slight relative degradation -3.4% corresponds to one more misrecognized word), however, there is significant improvement for other conditions.

5 Selecting recognizer with the highest total likelihood

First, we must be able to recognize what is the noisy condition of a test utterance, if we want to use *matched system* for real application. Each recognizer in *matched system* trained on one particular condition, can be regarded as a compound HMM, which can be evaluated to obtain the likelihood of observation sequence for given noisy condition. We select the condition, which corresponds to system giving the highest likelihood. Considering Viterbi approximation of model evaluation, we can simply run the recognition for every system an select the result of such system giving the highest Viterbi score. This recognition scheme can be also regarded as one recognizer with recognition network shown in figure 2. We will refer to this system as to *highest likelihood system*.

In the first experiment with *highest likelihood system*, for every noisy condition, one recognizer with optimal (see section 4) number of Gaussians was trained in the same manner as for *matched system*. Table 4 shows accuracies of this system together with relative word error rate improvement over the best *baseline system*. This system seems to be good approximation for *matched system*, which is even outperformed for 20 dB and 10 dB SNR conditions. Note that, in contrary to *matched system*, we can obtain also results for unseen conditions, since the system itself decides which recognizer is the most appropriate for given utterance condition.

Table 4. Best likelihood system using systems with optimal number of Gaussians.

Condition	**Clean**	**Seen noises**			**Unseen noises**		
SNR level [dB]	-	**20**	**10**	**00**	**20**	**10**	**00**
Accuracy [%]	99.09	99.11	96.82	67.62	98.96	96.01	72.80
WER Imp. [%]	-3.41	28.23	21.67	36.61	15.45	-16.00	31.8

Table 5. Best likelihood system using all systems with 1 to 8 Gaussians.

Condition	**Clean**	**Seen noises**			**Unseen noises**		
SNR level [dB]	-	**20**	**10**	**00**	**20**	**10**	**00**
Accuracy [%]	99.18	99.05	96.85	67.50	99.00	96.22	73.09
WER Imp. [%]	6.82	23.39	22.41	36.37	18.70	-9.88	32.52

Table 6. Relaxed system.

Condition	**Clean**	**Seen noises**			**Unseen noises**		
SNR level [dB]	-	**20**	**10**	**00**	**20**	**10**	**00**
Accuracy [%]	99.18	98.91	96.71	69.78	98.81	96.25	74.37
WER Imp. [%]	6.82	12.10	18.97	40.84	3.25	-9.01	35.75

In the second experiment with *highest likelihood system*, for every of nine noisy conditions included in training set, five recognizers were trained with 1, 2, 3, 5 and 8 Gaussians. In this case, the system is not only allowed to select recognizer with respect to on which condition it is trained but also with respect to how many parameters it uses. Results from this experiment (table 5) are comparable with those obtained in previous experiment.

6 Noisy conditions forming a different pronunciation variant of a word

In previous experiments, systems were forced to use models trained on one particular condition for whole utterance. This is consistent with our test data, since

only one type of noise was artificially added to each test utterance with constant SNR level. Type and level of background noise is, however, changing in the real world. Therefore, we tried to relax this constraint the way, that noisy condition is expected to be constant only during a word. Recognition network of system that fulfills this new requirement is shown in figure 3. Now, when leaving one word model, we can enter another model trained even on different noisy condition. We can think, that words pronounced under different noisy conditions represent different pronunciation variants of the same word. We will refer to this system as to *relaxed system*.

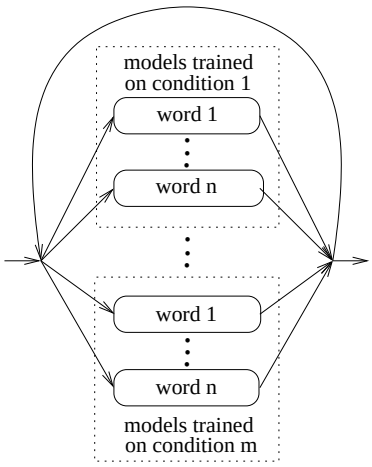

Fig. 3. Recognition network used by relaxed system.

In the experiment for *relaxed system*, for each word and each noisy condition, one model was trained. Since, the same model as in the first experiment for *highest likelihood system* were used, results for *relaxed system* shown in table 6 can be directly compared with those in table 4. Comparing these two systems, we can see notable degradation for some noisy conditions. This could be caused by relaxing constraint fulfilled in our test set. On the other hand, relaxing of these constraint is quite helpful for 0 SNR conditions unseen during training. Note that the relative word error rate improvement shown in table 6 is again with respect to *best baseline system*.

7 Discussion and conclusions

When training independent recognizer for each particular noisy condition and recognizing speech with appropriate recognizer we observed significant improvement, for recognition of noisy speech, with respect to recognizer conventionally trained on all conditions. For real application, first, the noisy condition must be

determined. Recognizers trained on each noisy condition were used as models for recognition of noisy condition. In this case, we must, however, perform recognition at least for each noisy condition, which is much more computationally expensive (in our experiment about 5 times slower) than running only one recognizer trained on all conditions. Assuming that recognition of type and level of noise is much easier task than recognition of speech, there is another alternative: to use different and more efficient recognizer for recognition of noisy condition. Such system could be even less computationally expensive than conventional system, since optimal number of parameters (Gaussians) is much smaller for recognizers trained per condition.

All subsets of training data with particular noisy condition contain the same speech utterances only corrupted by a particular type and level of noise. It means that there is no more information about speech in training data from different conditions. We also carried out experiments where each training utterance contained speech not repeated in any other utterance even with different noisy condition. Result from this experiment were not so promising as those presented here. The reason could be that recognizer trained on all data had more information about the speech and that was more important than the constraint of unchangeable condition during an utterance. However, this may not be a disadvantage, because often we have only clean data for training and noise can be added artificially to all data in the same manner as in our experiments.

8 Acknowledgments

This research has been partially supported by Grant Agency of Czech Republic under project No. 102/02/0124 and by EC project Multi-modal meeting manager (M4), No. IST-2001-34485. Jan Cernocky has been also supported by a postdoctoral grant from the Grant Agency of Czech Republic, no. 102/02/D108.

References

1. F. Jelinek. "Statistical Methods for Speech Recognition", MIT Press, 1998.
2. B. Gold and N. Morgan. "Speech and Audio Signal Processing", New York, 1999.
3. S. B. Davis and P. Mermelstein. "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on Acoustics, Speech & Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.
4. R.G. Leonard. "A database for speaker-independent digit recognition", Proc. ICASSP'84, pp. 42.11.1-4.
5. H. G. Hirsch, D. Pearce. "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000, "Automatic Speech Recognition: Challenges for the Next Millennium", Paris, France, September 2000.