An Analysis of Deep Neural Networks in Broad Phonetic Classes for Noisy Speech Recognition

F. de-la-Calle-Silos, A. Gallardo-Antolín, C. Peláez-Moreno

Department of Signal Theory and Communications.
Universidad Carlos III de Madrid.
Leganés (Madrid), Spain.

November, 25th, 2016

Outline

Outline

Introduction

1. Deep Neural Networks (DNN) have become very popular for acoustic modeling due to the improvements found over traditional Gaussian Mixture Models (GMM).
2. Not many works have addressed the robustness of these systems under noisy conditions.
3. In this paper we further investigate how these improvements are translated into the different broad phonetic classes and how does it compare to classical Hidden Markov Models (HMM)
4. A combination of the different DNN systems and classical HMM is also proposed.
5. Our hypothesis is that the traditional GMM/HMM systems have a different type of error than the Deep Neural Networks hybrid models.

Introduction

1. Deep Neural Networks (DNN) have become very popular for acoustic modeling due to the improvements found over traditional Gaussian Mixture Models (GMM).
2. Not many works have addressed the robustness of these systems under noisy conditions.
3. In this paper we further investigate how these improvements are translated into the different broad phonetic classes and how does it compare to classical Hidden Markov Models (HMM)
4. A combination of the different DNN systems and classical HMM is also proposed.
5. Our hypothesis is that the traditional GMM/HMM systems have a different type of error than the Deep Neural Networks hybrid models.

Introduction

1. Deep Neural Networks (DNN) have become very popular for acoustic modeling due to the improvements found over traditional Gaussian Mixture Models (GMM).
2. Not many works have addressed the robustness of these systems under noisy conditions.
3. In this paper we further investigate how these improvements are translated into the different broad phonetic classes and how does it compare to classical Hidden Markov Models (HMM)
4. A combination of the different DNN systems and classical HMM is also proposed.
5. Our hypothesis is that the traditional GMM/HMM systems have a different type of error than the Deep Neural Networks hybrid models.

Introduction

1. Deep Neural Networks (DNN) have become very popular for acoustic modeling due to the improvements found over traditional Gaussian Mixture Models (GMM).
2. Not many works have addressed the robustness of these systems under noisy conditions.
3. In this paper we further investigate how these improvements are translated into the different broad phonetic classes and how does it compare to classical Hidden Markov Models (HMM)
4. A combination of the different DNN systems and classical HMM is also proposed.
5. Our hypothesis is that the traditional GMM/HMM systems have a different type of error than the Deep Neural Networks hybrid models.

Introduction

1. Deep Neural Networks (DNN) have become very popular for acoustic modeling due to the improvements found over traditional Gaussian Mixture Models (GMM).
2. Not many works have addressed the robustness of these systems under noisy conditions.
3. In this paper we further investigate how these improvements are translated into the different broad phonetic classes and how does it compare to classical Hidden Markov Models (HMM)
4. A combination of the different DNN systems and classical HMM is also proposed.
5. Our hypothesis is that the traditional GMM/HMM systems have a different type of error than the Deep Neural Networks hybrid models.

Why Deep Neural Networks?

1. Improved performance.
2. DNNs have a larger number of hidden layers leading to systems with many more parameters than the traditional HMMs:
   - ☺ Less influenced by training and testing mismatch.
   - ☹ Can easily suffer from overfitting: pretraining, dropout, maxout.
3. Hybrid ANN/HMM architectures:

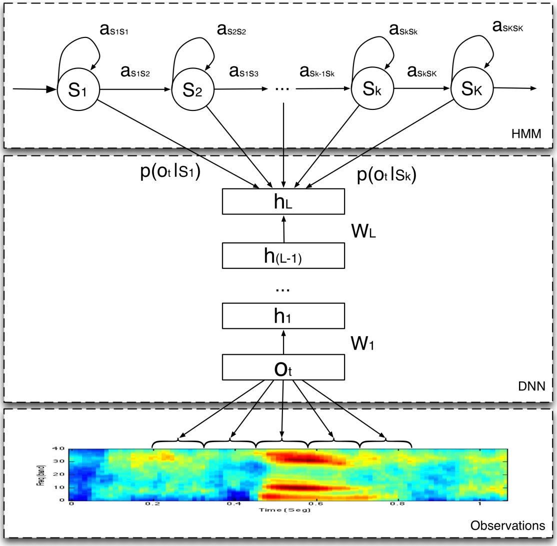
Why Deep Neural Networks?

1. Improved performance.
2. DNNs have a larger number of hidden layers leading to systems with many more parameters than the traditional HMMs:
   - ☺ Less influenced by training and testing mismatch.
   - ☹ Can easily suffer from overfitting: pretraining, dropout, maxout.
3. Hybrid ANN/HMM architectures:

Why Deep Neural Networks?

1. Improved performance.
2. DNNs have a larger number of hidden layers leading to systems with many more parameters than the traditional HMMs:
   - ☺ Less influenced by training and testing mismatch.
   - ☹ Can easily suffer from overfitting: pretraining, dropout, maxout.
3. Hybrid ANN/HMM architectures:

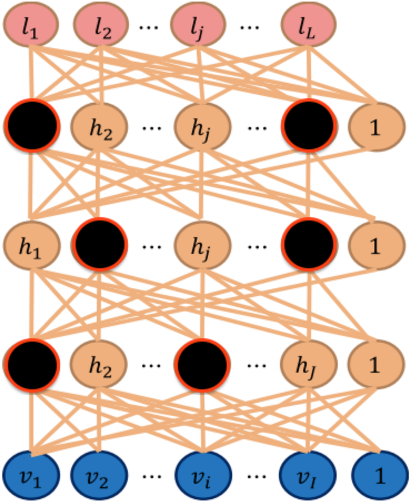
ASR Hybrid Model



Outline

Dropout

Randomly omitting a certain percentage of the hidden units on each training iteration



Dropout

Randomly omitting a certain percentage of the hidden units on each training iteration

$$\mathbf{h}^{(l+1)} = m^{(l)} \star \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right), \quad 1 \leq l \leq L \qquad (1)$$

where $m^{(l)}$ is a binary vector of the same dimension of $\mathbf{h}^{(l)}$ whose elements are sampled from a Bernoulli distribution with probability p: Hidden Drop Factor (*HDF*) .

1. Only applied in the training stage whereas on testing all the hidden units become active.
2. Can be seen as an ensemble of DNNs.
3. Similar to bagging.

Dropout

Randomly omitting a certain percentage of the hidden units on each training iteration

$$\mathbf{h}^{(l+1)} = m^{(l)} \star \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right), \quad 1 \leq l \leq L \qquad (1)$$

where $m^{(l)}$ is a binary vector of the same dimension of $\mathbf{h}^{(l)}$ whose elements are sampled from a Bernoulli distribution with probability p: Hidden Drop Factor (*HDF*) .

1. Only applied in the training stage whereas on testing all the hidden units become active.
2. Can be seen as an ensemble of DNNs.
3. Similar to bagging.

Dropout

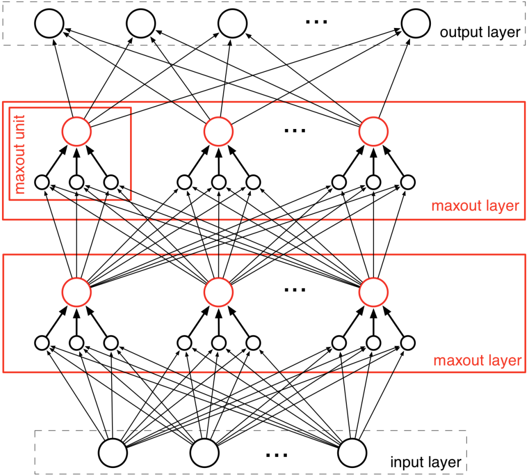
Randomly omitting a certain percentage of the hidden units on each training iteration

$$\mathbf{h}^{(l+1)} = m^{(l)} \star \sigma\left(\mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right), \quad 1 \leq l \leq L \qquad (1)$$

where $m^{(l)}$ is a binary vector of the same dimension of $\mathbf{h}^{(l)}$ whose elements are sampled from a Bernoulli distribution with probability p: Hidden Drop Factor (*HDF*) .

1. Only applied in the training stage whereas on testing all the hidden units become active.
2. Can be seen as an ensemble of DNNs.
3. Similar to bagging.

Maxout



A Maxout Network of 2 hidden layers and a group size of $g = 3$.
The hidden nodes in red perform the max operation.

Maxout (DMN)

Each hidden unit takes the maximum value over the g units of a group

$$h_i^{(l+1)} = \max_{j \in 1,\dots,g} z_{ij}^{(l+1)}, \quad 1 \leq l \leq L \tag{2}$$

where $z_{ij}^{(l+1)}$ is the lineal pre-activation values from the l layer:

$$\mathbf{z}^{(l+1)} = \mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)} \tag{3}$$

1. DMNs reduce the number of parameters over DNNs: the weight matrix of each layer i is $1/g$ of its equivalent DNN.
2. Also maxout units can approximate any convex function.

Maxout (DMN)

Each hidden unit takes the maximum value over the g units of a group

$$h_i^{(l+1)} = \max_{j \in 1,\ldots,g} z_{ij}^{(l+1)}, \quad 1 \leq l \leq L \tag{2}$$

where $z_{ij}^{(l+1)}$ is the lineal pre-activation values from the l layer:

$$\mathbf{z}^{(l+1)} = \mathbf{W}^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)} \tag{3}$$

1. DMNs reduce the number of parameters over DNNs: the weight matrix of each layer i is $1/g$ of its equivalent DNN.
2. Also maxout units can approximate any convex function.

Outline

Results: Corpus

1. Experiments were performed on the TIMIT corpus.
2. 462 speakers training set, 50 speakers development set for tuning.
3. Results are reported using the 24-speaker core test set.
4. Added noise (white, street, music and speaker) using FANT.
5. Kaldi toolkit: GMM-HMM.
6. Kaldi + Tensorflow: DNN-HMM.

Results: Corpus

1. Experiments were performed on the TIMIT corpus.
2. 462 speakers training set, 50 speakers development set for tuning.
3. Results are reported using the 24-speaker core test set.
4. Added noise (white, street, music and speaker) using FANT.
5. Kaldi toolkit: GMM-HMM.
6. Kaldi + Tensorflow: DNN-HMM.

Results: Corpus

1. Experiments were performed on the TIMIT corpus.
2. 462 speakers training set, 50 speakers development set for tuning.
3. Results are reported using the 24-speaker core test set.
4. Added noise (white, street, music and speaker) using FANT.
5. Kaldi toolkit: GMM-HMM.
6. Kaldi + Tensorflow: DNN-HMM.

Results: Corpus

1. Experiments were performed on the TIMIT corpus.
2. 462 speakers training set, 50 speakers development set for tuning.
3. Results are reported using the 24-speaker core test set.
4. Added noise (white, street, music and speaker) using FANT.
5. Kaldi toolkit: GMM-HMM.
6. Kaldi + Tensorflow: DNN-HMM.

Results: Corpus

1. Experiments were performed on the TIMIT corpus.
2. 462 speakers training set, 50 speakers development set for tuning.
3. Results are reported using the 24-speaker core test set.
4. Added noise (white, street, music and speaker) using FANT.
5. Kaldi toolkit: GMM-HMM.
6. Kaldi + Tensorflow: DNN-HMM.

Results: Corpus

1. Experiments were performed on the TIMIT corpus.
2. 462 speakers training set, 50 speakers development set for tuning.
3. Results are reported using the 24-speaker core test set.
4. Added noise (white, street, music and speaker) using FANT.
5. Kaldi toolkit: GMM-HMM.
6. Kaldi + Tensorflow: DNN-HMM.

Results: Clean Conditions

Recognition results in terms of PER(%) for the TIMIT development and core test sets in clean conditions.

Method	Dev (PER %)	Eval (PER %)
Mono	31.90	32.57
Triphone	24.70	26.68
Triphone LDA + MLLT + SAT	20.40	21.77
DNN random [5 x 1024]	19.80	21.25
DNN pretrain [5 x 1024]	19.17	20.69
DNN pretrain + dropout [5 x 1024]	18.49	19.46
DMN [5 x 400]	17.73	18.54

Outline

Analysis in broad phonetic classes

1. High impact of the DNN on ASR is its enhanced overall performance.
2. These new systems could be fused with the others to even obtain better robustness.
3. The combined systems should individually present different error behaviors and strengths.
4. We split the overall results into broad phonetic classes: vowels, semivowels, nasals consonants, fricative consonants, affricates consonants, stop closures and silence segments.
5. Tested in different noise conditions (white, street, music and speaker) at 15 *dB SNR*

Analysis in broad phonetic classes

1. High impact of the DNN on ASR is its enhanced overall performance.
2. These new systems could be fused with the others to even obtain better robustness.
3. The combined systems should individually present different error behaviors and strengths.
4. We split the overall results into broad phonetic classes: vowels, semivowels, nasals consonants, fricative consonants, affricates consonants, stop closures and silence segments.
5. Tested in different noise conditions (white, street, music and speaker) at 15 *dB SNR*

Analysis in broad phonetic classes

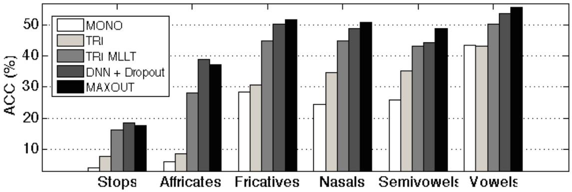
1. High impact of the DNN on ASR is its enhanced overall performance.
2. These new systems could be fused with the others to even obtain better robustness.
3. The combined systems should individually present different error behaviors and strengths.
4. We split the overall results into broad phonetic classes: vowels, semivowels, nasals consonants, fricative consonants, affricates consonants, stop closures and silence segments.
5. Tested in different noise conditions (white, street, music and speaker) at 15 *dB SNR*

Analysis in broad phonetic classes

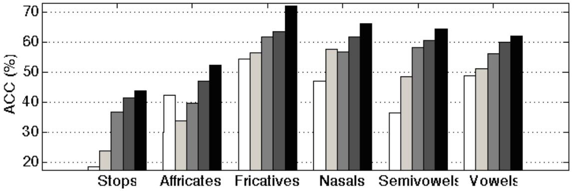
1. High impact of the DNN on ASR is its enhanced overall performance.
2. These new systems could be fused with the others to even obtain better robustness.
3. The combined systems should individually present different error behaviors and strengths.
4. We split the overall results into broad phonetic classes: vowels, semivowels, nasals consonants, fricative consonants, affricates consonants, stop closures and silence segments.
5. Tested in different noise conditions (white, street, music and speaker) at 15 *dB SNR*

Analysis in broad phonetic classes

1. High impact of the DNN on ASR is its enhanced overall performance.
2. These new systems could be fused with the others to even obtain better robustness.
3. The combined systems should individually present different error behaviors and strengths.
4. We split the overall results into broad phonetic classes: vowels, semivowels, nasals consonants, fricative consonants, affricates consonants, stop closures and silence segments.
5. Tested in different noise conditions (white, street, music and speaker) at 15 *dB SNR*

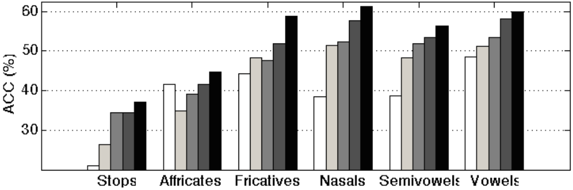


(a) White Noise

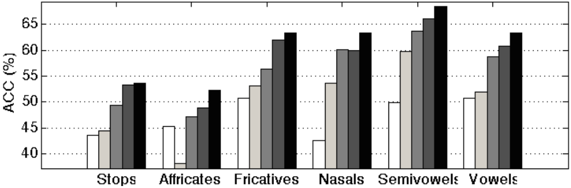


(b) Street Noise

Comparison of the performance in broad phonetic classes of the different systems in terms of PER [%] for TIMIT test set in different noisy conditions at 15 dB SNR.



(a) Music Noise



(b) Speaker Noise

Comparison of the performance in broad phonetic classes of the different systems in terms of PER [%] for TIMIT test set in different noisy conditions at 15 dB SNR.

Analysis in broad phonetic classes: conclusions

1. Performance is tightly related to the particular phonetic class.
2. Stops and affricates are the least resilient.
3. Relative improvements of DNN variants are distributed unevenly.
4. DNN and DMN based systems, is significantly dependent on the phonetic classes, being stops and affricates the most difficult ones.
5. Due reduced number of instances of affricates causes a erratic behavior of the different systems.
6. Stops match the performance ordering of the systems with exception on white noise where DNNs are slightly better than DMNs.

Analysis in broad phonetic classes: conclusions

1. Performance is tightly related to the particular phonetic class.
2. Stops and affricates are the least resilient.
3. Relative improvements of DNN variants are distributed unevenly.
4. DNN and DMN based systems, is significantly dependent on the phonetic classes, being stops and affricates the most difficult ones.
5. Due reduced number of instances of affricates causes a erratic behavior of the different systems.
6. Stops match the performance ordering of the systems with exception on white noise where DNNs are slightly better than DMNs.

Analysis in broad phonetic classes: conclusions

1. Performance is tightly related to the particular phonetic class.
2. Stops and affricates are the least resilient.
3. Relative improvements of DNN variants are distributed unevenly.
4. DNN and DMN based systems, is significantly dependent on the phonetic classes, being stops and affricates the most difficult ones.
5. Due reduced number of instances of affricates causes a erratic behavior of the different systems.
6. Stops match the performance ordering of the systems with exception on white noise where DNNs are slightly better than DMNs.

Analysis in broad phonetic classes: conclusions

1. Performance is tightly related to the particular phonetic class.
2. Stops and affricates are the least resilient.
3. Relative improvements of DNN variants are distributed unevenly.
4. DNN and DMN based systems, is significantly dependent on the phonetic classes, being stops and affricates the most difficult ones.
5. Due reduced number of instances of affricates causes a erratic behavior of the different systems.
6. Stops match the performance ordering of the systems with exception on white noise where DNNs are slightly better than DMNs.

Analysis in broad phonetic classes: conclusions

1. Performance is tightly related to the particular phonetic class.
2. Stops and affricates are the least resilient.
3. Relative improvements of DNN variants are distributed unevenly.
4. DNN and DMN based systems, is significantly dependent on the phonetic classes, being stops and affricates the most difficult ones.
5. Due reduced number of instances of affricates causes a erratic behavior of the different systems.
6. Stops match the performance ordering of the systems with exception on white noise where DNNs are slightly better than DMNs.

Analysis in broad phonetic classes: conclusions

1. Performance is tightly related to the particular phonetic class.
2. Stops and affricates are the least resilient.
3. Relative improvements of DNN variants are distributed unevenly.
4. DNN and DMN based systems, is significantly dependent on the phonetic classes, being stops and affricates the most difficult ones.
5. Due reduced number of instances of affricates causes a erratic behavior of the different systems.
6. Stops match the performance ordering of the systems with exception on white noise where DNNs are slightly better than DMNs.

Analysis in broad phonetic classes: conclusions

1. For the remaining phonetic classes, we can conclude that the improvements due to DNN and DMN learning algorithms are translated to all of them but not with the same intensity.
2. The most benefited phonetic class is fricatives since the relative loss of the best HMM-based system from the best DNN-based (DMN) is the highest (13 for white noise, 14 for street, 19 for music and 11 for speaker).

Analysis in broad phonetic classes: conclusions

1. For the remaining phonetic classes, we can conclude that the improvements due to DNN and DMN learning algorithms are translated to all of them but not with the same intensity.
2. The most benefited phonetic class is fricatives since the relative loss of the best HMM-based system from the best DNN-based (DMN) is the highest (13 for white noise, 14 for street, 19 for music and 11 for speaker).

Outline

System Combination

1. Combination of the different systems can improve the recognition rates since the types of errors are different for each system.
2. Two combinations proposed:
   1. Comb 1: DNN with dropout system + DMN-based one
   2. Comb 2: DNN with dropout + DMN + triphone MLLT.
3. Systems are fused using Recognition Output Voting Error Reduction (ROVER) by Average Confidence Scores.

System Combination

1. Combination of the different systems can improve the recognition rates since the types of errors are different for each system.
2. Two combinations proposed:
   1. Comb 1: DNN with dropout system + DMN-based one
   2. Comb 2: DNN with dropout + DMN + triphone MLLT.
3. Systems are fused using Recognition Output Voting Error Reduction (ROVER) by Average Confidence Scores.

System Combination

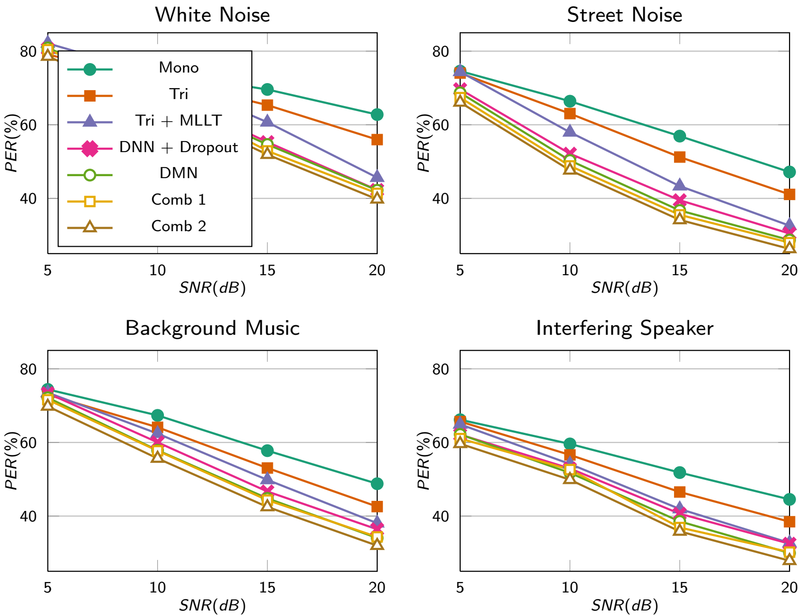
1. Combination of the different systems can improve the recognition rates since the types of errors are different for each system.
2. Two combinations proposed:
   1. Comb 1: DNN with dropout system + DMN-based one
   2. Comb 2: DNN with dropout + DMN + triphone MLLT.
3. Systems are fused using Recognition Output Voting Error Reduction (ROVER) by Average Confidence Scores.

System Combination

1. Combination of the different systems can improve the recognition rates since the types of errors are different for each system.
2. Two combinations proposed:
   1. Comb 1: DNN with dropout system + DMN-based one
   2. Comb 2: DNN with dropout + DMN + triphone MLLT.
3. Systems are fused using Recognition Output Voting Error Reduction (ROVER) by Average Confidence Scores.

System Combination

1. Combination of the different systems can improve the recognition rates since the types of errors are different for each system.
2. Two combinations proposed:
   1. Comb 1: DNN with dropout system + DMN-based one
   2. Comb 2: DNN with dropout + DMN + triphone MLLT.
3. Systems are fused using Recognition Output Voting Error Reduction (ROVER) by Average Confidence Scores.


White Noise
Street Noise
Background Music
Interfering Speaker
Mono
Tri
Tri + MLLT
DNN + Dropout
DMN
Comb 1
Comb 2
PER(%)
SNR(dB)
80
60
40
5
10
15
20

System Combination: Conclusions

1. DNN with dropout + DMN provides better accuracies than DMN alone for all of the noises.
2. **Improvements are small**, but performance is still significantly dependent on the phonetic classes.
3. The inclusion of triphone-based ASR system improves the recognition rates obtained by the first combination and any of the other systems.
4. Traditional GMM-HMM-based systems produce different types of errors than the DNNs hybrid models.

System Combination: Conclusions

1. DNN with dropout + DMN provides better accuracies than DMN alone for all of the noises.
2. **Improvements are small**, but performance is still significantly dependent on the phonetic classes.
3. The inclusion of triphone-based ASR system improves the recognition rates obtained by the first combination and any of the other systems.
4. Traditional GMM-HMM-based systems produce different types of errors than the DNNs hybrid models.

System Combination: Conclusions

1. DNN with dropout + DMN provides better accuracies than DMN alone for all of the noises.
2. **Improvements are small**, but performance is still significantly dependent on the phonetic classes.
3. The inclusion of triphone-based ASR system improves the recognition rates obtained by the first combination and any of the other systems.
4. Traditional GMM-HMM-based systems produce different types of errors than the DNNs hybrid models.

System Combination: Conclusions

1. DNN with dropout + DMN provides better accuracies than DMN alone for all of the noises.
2. **Improvements are small**, but performance is still significantly dependent on the phonetic classes.
3. The inclusion of triphone-based ASR system improves the recognition rates obtained by the first combination and any of the other systems.
4. Traditional GMM-HMM-based systems produce different types of errors than the DNNs hybrid models.

Outline

Conclusions

1. Analysis of the errors that both HMM and DNN-based systems produce on broad phonetic.
2. The main conclusion was that the improvements are more significant in sonorants (vowels, semivowels, nasals), followed by stops.
3. DMN provide improved robustness due to their flexibility in the activation function
4. Combination of GMM-HMM and DNN-based systems improves the results in comparison to the individual ASR systems.
5. Future directions: larger databases, other DNN alternatives, CNN, combine DNNs by joining in a last layer.

Conclusions

1. Analysis of the errors that both HMM and DNN-based systems produce on broad phonetic.
2. The main conclusion was that the improvements are more significant in sonorants (vowels, semivowels, nasals), followed by stops.
3. DMN provide improved robustness due to their flexibility in the activation function
4. Combination of GMM-HMM and DNN-based systems improves the results in comparison to the individual ASR systems.
5. Future directions: larger databases, other DNN alternatives, CNN, combine DNNs by joining in a last layer.

Conclusions

1. Analysis of the errors that both HMM and DNN-based systems produce on broad phonetic.
2. The main conclusion was that the improvements are more significant in sonorants (vowels, semivowels, nasals), followed by stops.
3. DMN provide improved robustness due to their flexibility in the activation function
4. Combination of GMM-HMM and DNN-based systems improves the results in comparison to the individual ASR systems.
5. Future directions: larger databases, other DNN alternatives, CNN, combine DNNs by joining in a last layer.

Conclusions

1. Analysis of the errors that both HMM and DNN-based systems produce on broad phonetic.
2. The main conclusion was that the improvements are more significant in sonorants (vowels, semivowels, nasals), followed by stops.
3. DMN provide improved robustness due to their flexibility in the activation function
4. Combination of GMM-HMM and DNN-based systems improves the results in comparison to the individual ASR systems.
5. Future directions: larger databases, other DNN alternatives, CNN, combine DNNs by joining in a last layer.

Conclusions

1. Analysis of the errors that both HMM and DNN-based systems produce on broad phonetic.
2. The main conclusion was that the improvements are more significant in sonorants (vowels, semivowels, nasals), followed by stops.
3. DMN provide improved robustness due to their flexibility in the activation function
4. Combination of GMM-HMM and DNN-based systems improves the results in comparison to the individual ASR systems.
5. Future directions: larger databases, other DNN alternatives, CNN, combine DNNs by joining in a last layer.

Questions?