Chapter 6

Summary and conclusion

This chapter summarises the results given in Chapter 5 and we draw some conclusions from it. In particular, the best results obtained with the various techniques tested, are repeated here, and the improvement over the baseline systems are given and compared with each other. The shortcomings and possible future work, related to that done in this dissertation, are also discussed here.

6.1 Summary of results

For the purposes of comparing the various techniques, we define the improvement as

$$Improvement = \frac{new - old}{old} \cdot 100\%, \tag{6.1}$$

where *new* indicates the improved system performance, and *old* indicates the original or baseline system. The improvement is multiplied by -1 when it is calculated for insertions, deletions and substitutions, as these are to be minimised, not maximised. Using this convention, any improvement with a negative value, indicates that performance

Table 6.1: Summary of results showing the segmentation performance obtained with the HMM based segmentation system, for different experimental conditions.

System no.	RE	LM	Mixs. (#)	Accuracy (%)	Correct (%)	Hits (#)	Insertions (#)	Deletions (#)
1	No	0.0	8	71.23	87.07	31.40	5.40	4.64
2	No	2.0	7	**75.62**	84.03	30.33	**2.84**	5.72
3	Yes	0.0	7	70.27	**88.04**	**31.76**	6.07	**4.29**
4	Yes	2.0	4	75.47	85.28	30.77	3.31	5.27

degraded, rather than improved.

Table 6.1 summarises the segmentation performance obtained with the HMM based segmentation system, for different experimental conditions. In this table, "RE" indicates embedded re-estimation of the HMM models, and "LM" indicates the use of a language model with the corresponding values in the table indicating the language model scale factor (0.0 if no language model is used). From this table it can be seen that HMM segmentation performance is higher, if accuracy is the criterium, when no embedded re-estimation is used during the training process. This is primarily due to the fact that phoneme recognition performance is maximised and not segmentation performance. The use of a language model also improves segmentation performance significantly. This is due to the fact that fewer insertions occur. Again it should be noted that even though the best performances were obtained with 8, 7, and 4 mixtures, 2 to 3 mixtures would probably have been sufficient to model the acoustic space due to the fact that the TIMIT database only consists of two genders (male and female).

Table 6.2 shows the results obtained for BRNN based segmentation. Here the best performance is obtained with 60 forward hidden nodes, and 60 backward hidden nodes, indicated just by "hidden nodes" in the table. The threshold in the postprocessor, used to decide when the probability of a boundary is high enough to indicate the presence of a phoneme boundary, is found to be best at 0.35. This table shows that a segmentation

Table 6.2: Summary of results showing the segmentation performance obtained with the BRNN based segmentation system (with 60 forward and backward hidden nodes), for different experimental conditions. Also shown is the improvement over the HMM based segmentation systems.

Hidden nodes	Threshold	Accuracy (%)	Correct (%)	Hits (#)	Insertions (#)	Deletions (#)
60	0.35	80.12	86.20	31.14	2.14	4.91
Improvement over HMM based segmentation systems						
HMM system						
1		12.59	-1.00	-0.83	60.37	-5.82
2		6.06	**2.58**	**2.67**	24.65	**14.16**
3		**14.13**	-2.09	-1.95	**64.74**	-14.45
4		6.27	1.08	1.20	35.35	6.83

accuracy of 80.12% is obtained with the BRNN based segmentation system, in contrast to the 71.23%, 75.62%, 70.27%, and 75.47% accuracies in Table 6.1 of HMM based segmentation systems 1, 2, 3, and 4, respectively. The corresponding improvement in accuracy of the BRNN based segmentation system over the HMM based segmentation systems is also shown in Table 6.2 as 12.59%, 6.06%, 14.13% and 6.27%, respectively. The BRNN based segmentation system thus significantly outperforms all of the HMM based segmentation systems. This can be partially explained by the fact that the neural network is able to use all of the context information in the prediction of the phoneme boundaries, in effect learning a better "language model" than that of the HMM systems. The neural network is able to efficiently discriminate between a phoneme boundary and no phoneme boundary, while the HMM systems can only discriminate between two different phonemes.

Table 6.3 shows the baseline recognition performance of the HMM recognition system, for different experimental conditions. In this table, "WP" indicates the use of a fixed

Table 6.3: Summary of results showing the baseline recognition performance, for different experimental conditions.

System no.	RE	LM	WP	Cor (%)	Acc (%)	Hits (#)	Del (#)	Sub (#)	Ins (#)
1	No	0.0	0.0	61.76	52.40	4456	**713**	2046	675
2	No	0.0	-6.0	57.56	53.96	4153	1237	1825	260
3	No	4.0	0.0	62.62	60.58	4518	1178	**1519**	**147**
4	No	5.0	5.0	**64.24**	**61.44**	**4635**	986	1594	202
5	Yes	0.0	0.0	63.92	52.99	4612	**651**	1952	789
6	Yes	0.0	-6.0	60.18	55.45	4342	1052	1821	341
7	Yes	4.0	0.0	65.18	62.38	4703	1006	**1506**	**202**
8	Yes	5.0	5.0	**67.05**	**63.23**	**4838**	820	1557	276

word transition penalty (the constant bias term, w_b with $c = 1$, in Equation (4.12)). It can be seen that the use of embedded re-estimation during the training process improves recognition performance. When a language model and fixed word transition penalty are used separately, the performance can be increased. The language model significantly decreases the amount of substitutions and insertions, as it models the probability of one phoneme following another. The use of a word transition penalty has a similar effect, decreasing the likelihood of a transition from one phoneme to another, resulting in a reduction in the number of insertions. When both a language model and word transition penalty are used, the best performance is obtained. These two parameters must be jointly optimised on the test set.

Table 6.4 shows the phoneme recognition performance when segmentation information is included into the decoding process. In this table, "AWP" indicates the use of the two adaptive word transition penalty terms (the first two terms in Equation (4.12)), with the corresponding values in the table indicating the adaptive word transition penalty scale factor (same factor is used for both terms). In systems 1 to 4, the transition

probabilities of the HMMs are modified from the segmentation information, using a linear combination of segmentation probabilities and HMM transition probabilities (indicated by "Trans. (Linear)"). In systems 5 to 8 a non-linear combination is used ("Trans. (Non-linear)"). Systems 9 to 12 indicate the use of only the adaptive word transition penalty ("AWP"). Finally, systems 13 to 20 are similar to systems 1 to 8, where the adaptive word transition penalty is used in addition to the HMM transition probability modification. Results are shown for the cases when no language model or fixed word transition penalty is used, when only a language model or fixed word transition penalty is used, and when both are used simultaneously. It can be seen that phoneme recognition performance improves in all the cases when either or both a language model and fixed word transition penalty are used. The linear combination of HMM transition probabilities and segmentation probabilities is slightly better than the non-linear combination. This is due to the fact that both the HMM transition probability and segmentation probabilities are used in linear combination, instead of only the maximum of the two for non-linear combination. Linear combination is thus less susceptible to "noisy" estimates of segmentation probability and in effect performs a smoothing function as a result of the averaging procedure. The use of only adaptive word transition terms is also superior to the other systems. When both the HMM transition probability modification and adaptive word transition penalty are used, the performance is still not as good as that obtained when only the adaptive word transition penalty is used. This may be attributed to the fact that the two techniques may oppose each other in a small way.

Table 6.5 gives the improvements obtained with the use of segmentation information, as shown in Table 6.4, over that of the baseline recognition systems 5 to 8, given in Table 6.3, for the different cases when no language model or fixed word transition penalty is used (combined systems 1, 5, 9, 13 and 17, over baseline system 1), when only a fixed word transition penalty is used (combined systems 2, 6, 10, 14, and 18, over baseline system 2), when only a language model is used (combined systems 3, 7, 11, 15, and 19, over baseline system 3), and when both a language model and fixed word transition penalty are used (combined systems 4, 8, 12, 16, and 20, over baseline

Table 6.4: Summary of results showing the recognition performance with segmentation information included, for different experimental conditions.

System no.	Type	LM	WP	AWP	Cor (%)	Acc (%)	Hits (#)	Del (#)	Sub (#)	Ins (#)
1	Trans.	0.0	0.0	0.0	63.40	55.41	4574	739	1902	576
2		0.0	-4.0	0.0	61.00	56.34	4401	989	1825	336
3	(Linear)	3.0	0.0	0.0	65.29	62.54	4711	963	**1541**	**199**
4		5.0	8.0	0.0	**67.78**	**63.58**	**4890**	**730**	1595	303
5	Trans.	0.0	0.0	0.0	64.27	52.47	4637	**620**	1958	851
6		0.0	-5.0	0.0	61.29	55.56	4422	934	1859	413
7	(Non-linear)	4.0	0.0	0.0	65.36	62.37	4716	976	**1523**	**216**
8		5.0	5.0	0.0	**67.08**	**63.02**	**4840**	805	1570	293
9	AWP	0.0	0.0	9.0	61.98	59.40	4472	859	1884	186
10		0.0	1.0	9.0	62.33	59.60	4497	819	1899	197
11		3.0	0.0	3.0	65.07	63.01	4695	963	**1557**	**149**
12		4.0	17.0	10.0	**68.58**	**64.93**	**4948**	**610**	1657	263
13	Trans. and	0.0	0.0	8.0	61.66	59.22	4449	882	1884	176
14	AWP	0.0	3.0	9.0	62.56	59.58	4514	785	1916	215
15		3.0	0.0	1.0	65.03	62.73	4692	986	**1537**	**166**
16	(Linear)	4.0	17.0	8.0	**68.84**	**64.91**	**4967**	**597**	1651	284
17	Trans. and	0.0	0.0	9.0	61.91	59.21	4467	841	1907	195
18	AWP	0.0	5.0	11.0	62.80	59.33	4531	727	1957	250
19		3.0	0.0	3.0	65.35	63.10	4715	943	**1557**	**162**
20	(Non-linear)	4.0	15.0	9.0	**68.30**	**64.80**	**4928**	**616**	1671	253

system 4). Here it can be seen that improvement over the best baseline recogniser (baseline system 4), from the use of linear HMM transition probability modification is 0.55%, non-linear HMM transition probability modification -0.33%, adaptive word transition penalty only 2.69%, adaptive word transition penalty and linear HMM transition probability modification 2.66%, adaptive word transition penalty and non-linear HMM transition probability modification 2.48%. It is clear that the technique developed in this dissertation (the adaptive word transition penalty), is superior to that proposed by others (the HMM transition probability modification). It is also interesting to note that improvements can be obtained when no language model or word transition penalty is used by the recognition system. Improvement is also significant when only a word transition penalty is used. A slight improvement is still obtained when only a language model is used, and no word transition penalty.

6.2 Statistical significance

In order to determine whether the improvement obtained is statistically significant, a test of significance, or one-tailed hypothesis test of the difference of proportions, must be performed [89]. Let the null hypothesis be denoted by H_0, representing the hypothesis that there is no statistical difference between the baseline system and the modified system (with the neural network). Let the alternative hypothesis, H_1 represent the hypothesis that there is a statistical significance. The null hypothesis at a certain level of significance, α, is then rejected if the z score lies outside a certain range, indicating that the improvement is statistically significant. Otherwise the null hypothesis is accepted, indicating that there is no statistical significance. Hypothesis H_0 is thus accepted if

$$Z = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}} < \alpha, \tag{6.2}$$

where $\tilde{P} = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$ is used as an estimate of the population proportion p, $n_1 = n_2 =$

Table 6.5: Summary of results showing the improvement of the recognition performance with segmentation information included, for different experimental conditions. The improvement is shown only for embedded re-estimation cases.

Improvement over HMM baseline recognition systems							
System no.	Type	Cor (%)	Acc (%)	Hits (#)	Del (#)	Sub (#)	Ins (#)
1	Trans.	-0.81	**4.57**	-0.82	-13.52	**2.56**	**27.00**
2		**1.36**	1.61	**1.36**	5.99	-0.22	1.47
3	(Linear)	0.17	0.26	0.17	4.27	-2.32	1.49
4		1.09	0.55	1.07	**10.98**	-2.44	-9.78
5	Trans.	0.55	-0.98	0.54	4.76	**-0.31**	-7.86
6		**1.84**	**0.20**	**1.84**	**11.22**	-2.09	-21.11
7	(Non-linear)	0.28	-0.02	0.28	2.98	-1.13	-6.93
8		0.04	-0.33	0.04	1.83	-0.83	**-6.16**
9	AWP	-3.04	**12.10**	-3.04	-31.95	**3.48**	**76.43**
10		**3.57**	7.48	**3.57**	22.15	-4.28	42.23
11		-0.17	1.01	-0.17	4.27	-3.39	26.24
12		2.28	2.69	2.27	**25.61**	-6.42	4.71
13	Trans. and	-3.54	**11.76**	-3.53	-35.48	**3.48**	**77.69**
14	AWP	**3.95**	7.45	**3.96**	25.38	-5.22	36.95
15		-0.23	0.56	-0.23	1.99	-2.06	17.82
16	(Linear)	2.67	2.66	2.67	**27.20**	-6.04	-2.90
17	Trans. and	-3.14	**11.74**	-3.14	-29.19	**2.31**	**75.29**
18	AWP	**4.35**	7.00	**4.35**	**30.89**	-7.47	26.69
19		0.26	1.15	0.26	6.26	-3.39	19.80
20	(Non-linear)	1.86	2.48	1.86	24.88	-7.32	8.33

N, P_1 is the percentage accuracy of the modified system, divided by 100 (to get a proportion), and P_2 is the percentage accuracy of the baseline system, divided by 100. The standard deviation, $\sigma_{P_1-P_2}$ is given by

$$\sigma_{P_1-P_2} = \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_1}\right)}. \tag{6.3}$$

The above equations can be rewritten in terms of the improvement in accuracy, as

$$Improvement < \frac{\alpha \cdot \sigma_{P_1-P_2}}{P_2} \cdot 100\%, \tag{6.4}$$

where *Improvement* is defined in Equation (6.1). This equation indicates the maximum percentage improvement allowed in order to accept hypothesis H_0, and the difference between the results of the two experiments to be statistically insignificant.

6.3 Conclusion

From the Section 6.1 it can be seen that the most important segmentation results are that of the segmentation accuracy of HMM segmentation system number 2 in Table 6.1, as well as the RNN segmentation system in Table 6.2. The most important results of the phoneme recognition experiments, are that of phoneme recognition systems 4 and 12 in Table 6.4, where a neural network is incorporated into the recognition process. The improvement of the RNN based segmentation system over HMM based segmentation system number 2, is shown in Table 6.2. The phoneme recognition accuracy improvements of the modified phoneme recognition systems, over baseline system 8 in Table 6.3, are summarised in Table 6.5.

To determine whether the segmentation accuracy improvement is statistically significant, the accuracies (divided by 100) of the BRNN based segmentation system in

Table 6.6: Minimum percentage improvement required, at various levels of significance, so that the improvement in segmentation accuracy is statistically significant.

z	α	Improvement required (%)
1.28	0.1	0.45
1.645	0.05	0.58
1.96	0.025	0.69
2.33	0.01	0.82

Table 6.2 is used as P_1, and the accuracy (divided by 100) of baseline system 2 in Table 6.1 is used as P_2, and the minimum percentage improvement required to be statistically significant, at a specific significance level, can thus be calculated. Here $N = 48446$ is the total number of phoneme boundaries in the TIMIT full test set. Table 6.6 shows the improvement required to be statistically significant, for various levels of significance.

In the Section 6.1 it is shown that the best accuracy obtained with the baseline HMM segmentation system number 2 is 75.62%. The best segmentation performance obtained with the BRNN based segmentation system is 80.12%. When this improvement of 6.06% is compared with the values in Table 6.6, it is clear that the BRNN based segmentation system significantly outperforms the HMM based segmentation system.

When the accuracies (divided by 100) of phoneme recognition systems 4 and 12 in Table 6.4 are used as P_1, and the accuracy (divided by 100) of baseline system 8 in Table 6.3 is used as P_2, then the minimum percentage improvement required to be statistically significant, at a specific significance level, can be calculated. Here $N = 7215$ is the total number of phonemes in the TIMIT core test set. Table 6.7 shows approximate values for phoneme recognition systems 4 and 12, at various levels of significance.

In the Section 6.1 it is shown that the best accuracy obtained with the baseline system

Table 6.7: Minimum percentage improvement required, at various levels of significance, so that the improvement in phoneme recognition accuracy is statistically significant.

z	α	Improvement required (%)
1.28	0.1	1.62
1.645	0.05	2.08
1.96	0.025	2.48
2.33	0.01	2.94

is 63.23%. When the HMM transition probabilities are combined linearly with the neural network outputs, the best accuracy is 63.58%, with a corresponding improvement of 0.55%. When this improvement is compared with the values in Table 6.7, it is clear that this technique, used by others, does not give a statistically significant improvement, in the experiments we conducted. When an adaptive word penalty is used (the technique developed in this dissertation) the best accuracy is 64.93%, with a corresponding improvement over the baseline system of 2.69%. This improvement is statistically significant up to a significance level of 0.025 (97.5% confidence). The technique developed in our work thus not only outperforms that used by others, but also gives a statistically significant improvement in the phoneme recognition accuracy.

6.4 Shortcomings and future work

The work presented in this dissertation is partially limited due to the significant amount of computational power needed. In particular, the following future work needs to be carried out, that could not be done here due to the limited computational resources and time constraints, namely to

- find the best neural network architecture, where experiments are performed to determine the optimal number of forward and backward hidden nodes separately

(in this dissertation, the forward and backward hidden nodes are set equal to each other, resulting in many "unused" weights),

- determine the best features for segmentation and recognition separately (in this dissertation the same features were used for segmentation and recognition, which may not be optimal),
- investigate the real-time implementation of such a system (in this dissertation, segmentation is performed off-line first, before recognition is performed, that uses the segmentation information),
- investigate the use of better segmentation postprocessors, e.g. based on segmentation lattices and dynamic programming, instead of just a threshold function, and
- explore performance of NN and HMM segmentation and recognition when the test corpus is clean (i.e., noise-free), but consists of non-read speach or speakers not included in the training sets.