Evaluation of Phone Lattice Based Speech Decoding

Jacques Duchateau, Kris Demuynck, Hugo Van hamme

ESAT department, Katholieke Universiteit Leuven, Belgium
Jacques.Duchateau@esat.kuleuven.be, Kris.Demuynck@esat.kuleuven.be

Abstract

Previously, we proposed a flexible two-layered speech recogniser architecture, called FLaVoR. In the first layer an unconstrained, task independent phone recogniser generates a phone lattice. Only in the second layer the task specific lexicon and language model are applied to decode the phone lattice and produce a word level recognition result. In this paper, we present a further evaluation of the FLaVoR architecture. The performance of a classical single-layered architecture and the FLaVoR architecture are compared on two recognition tasks, using the same acoustic, lexical and language models. On the large vocabulary Wall Street Journal 5k and 20k benchmark tasks, the two-layered architecture resulted in slightly but not significantly better word error rates. On a reading error detection task for a reading tutor for children, the FLaVoR architecture clearly outperformed the single-layered architecture.

Index Terms: ASR architecture, phone lattice decoding, system assessment

1. Introduction

The idea to use a phone recogniser as a first processing step in an open (and typically large) vocabulary word recognition system is more than thirty years old [1, 2, 3, 4]. But although a (phone) recogniser’s output upgraded from a single best (phone) string to a (phone) lattice over the years, it still seems to be hard to build a phone lattice based word recogniser that performs equally well as a straightforward single-layered word recogniser.

In [5], we proposed the FLaVoR architecture, a modular, two-layered recognition system architecture that allows building of recognisers for new tasks efficiently. This architecture was developed in the FLaVoR (Flexible Large Vocabulary Recognition) project1.

The key aspect of the FLaVoR framework consists of splitting up the search engine into two separate layers. The first layer performs phone recognition and outputs a dense phone network, which acts as an interface to the second layer. In this second layer, the word decoding is accomplished. This decoding may involve more complex models than a traditional single-layered architecture, in which the complexity of the search makes the use of simple models obligatory.

The two-layered architecture has several advantages. First, the modular setup allows to develop recognisers for different tasks more efficiently as the phone recogniser in the first layer may be re-used. Also, the use of a phone lattice allows to apply more complex modelling early in the recognition process, for instance phone duration modelling. Furthermore, the first layer in the FLaVoR architecture decodes only at the acoustic level, so for any segment in the phone lattice we can suppose there is real acoustic evidence in the audio. In the single-layered architecture, the lexicon and the language model are more actively looking for phones that complete the hypotheses in the search, so phone segments without real acoustic evidence may be introduced into the recognition result easily.

However, applying the lexicon and the language model only in a second stage of the recognition also has a drawback. Due to the high degree of acoustic confusability in speech, early inclusion of higher level information, as offered by the lexicon and the language model, has proved to be fruitful for reliably pruning away the most unlikely options from the search space. In other words, the phone lattice needed for good overall recognition performance may be so large that the FLaVoR architecture is of no practical use.

In [6], it was shown that the FLaVoR architecture performs equally well as the single-layered setup on the medium vocabulary (1000 word) Resource Management recognition task. In this paper, we investigate the performance of the FLaVoR architecture on two different tasks: a large vocabulary task, the Wall Street Journal (WSJ) benchmark, and a reading error detection task in a reading tutor for children. As we want to compare two recogniser architectures, we use the same modelling for both. So in this paper, we did not exploit the additional modelling mentioned above that is possible in the two-layered setup.

The paper is organised as follows: in section 2, the FLaVoR architecture is shortly reviewed for the recognition tasks at hand. Information on the experiment setup and the modelling involved is given in section 3. Section 4 then evaluates the recogniser architectures on both recognition tasks, and a conclusion is given in section 5.

2. The FLaVoR architecture

An overview of the FLaVoR architecture is given at the left of figure 1. In the first layer, a phone recogniser generates a phone lattice. The phone recogniser is based on an acoustic model and a probabilistic phone sequence model. The acoustic models for both tasks investigated are detailed in section 3. As for the phone sequence model, in both cases an N-gram was estimated from a database with read newspaper sentences. The phone recogniser for the WSJ task used a 4-gram for English, the one for the reading tutor (RT) task used a 3-gram for Dutch.

The search engine in the second layer turns the phone lattice into a word level recognition result. Similar to a single-layered system, this result may be a lattice that can be used in further processing, e.g. for generating confidence measures. In the second layer, the task dependent information is modelled. For the WSJ task, the benchmark bi-gram language models were used, both for the 5k word and the 20k word recognition task setups. For the reading tutor, the sentence that should be read is known. So we opted for a finite state transducer (FST) to produce a detailed model for the speech to be expected from the child, as

1website: http://www.esat.kuleuven.be/psi/spraak/projects/FLaVoR
Figure 1: Two-layered architecture (left) and task specific model in the reading tutor (right).

shown at the right of figure 1 for the Dutch sentence De tafel is rood (The table is red). The FST used is a composed FST: the word FSTs (top right in figure 1) and the garbage FST (middle right) are inserted at the right places in the sentence FST (bottom right).

The general idea of the two-layered architecture is that in the second layer, the system looks in the lattice for the phone transcript of the words in the lexicon. But as shown in [6], even for large (though computationally tractable) lattices, the phone transcript of some words, pronounced correctly by the speaker, are not available in the lattice. Therefore, for both recognition tasks evaluated in this paper, we help the system to find phone transcripts in the lattice. In the large vocabulary system (the WSJ task), an error model was used that allows a limited amount of substitutions, deletions and insertions when looking for a phone transcript in the phone lattice. This method was proposed and described in detail in [6]. In the RT task, the words that should be read by the child are known. So we can increase the probability of the phone sequences in those words in the N-gram that models the phone sequences. We tried this in different ways, e.g. by weighing the general phone sequence N-gram with a small task specific one, or simply by putting the probabilities of the phone sequences in the words to be read to one in the general N-gram. The conclusion was that the way in which the probabilities are increased does not matter for the overall performance of the system, as long as all probabilities for the words in the task are sufficiently high. For the evaluations in this paper, the logarithm of the probabilities for the expected phone sequences was set to zero in the general phone sequence N-gram.

In this paper, the FLaVoR architecture described in this section was compared to the corresponding single-layered system, i.e. the recognition system that uses the same acoustic modelling, lexicon and language model but that directly generates a word level result from the input audio. This means that the best score for any sentence is the same for both systems (for the large vocabulary task, the error model may exceptionally lead to differences). So differences in recognition result between both systems are due to the difference in the search strategy, as in a beam search at any point in time only the promising hypotheses are kept alive. For instance, in the first layer of the two-layered system, mainly the phone segments with sufficient acoustic evidence will be kept.

3. Recognition tasks

In this section, the acoustic databases for testing and for acoustic model estimation are described for both recognition tasks investigated in this paper.

3.1. Large vocabulary task

The performance of the presented FLaVoR architecture on large vocabulary recognition tasks was evaluated on the November 92 evaluation test set from the Wall Street Journal (WSJ) benchmark. Both for the 5k word closed vocabulary task and the 20k word open vocabulary task (which has a 1.9% out-of-vocabulary word rate), the reference bi-gram language model was used.

Training of the acoustic models is based on the standard SI-284 WSJ1 database which contains speech from 284 different speakers, 37516 sentences in total. The feature extraction is Mel spectrum based, including mean subtraction on the log spectrum, and a discriminant analysis resulting in 36 features. Speaker independent, cross word context dependent acoustic models were generated with 5443 tied HMM states, the state ty-
ing is defined by an automatically generated phonetic decision tree. In total 32725 tied gaussians (with diagonal covariance) were estimated. A flexible tying system for the gaussians resulted in tied states modelled as a weighted sum of on average 95.1 gaussians.

3.2. Reading tutor task

For detailed feedback in an automated reading tutor for children, it is important to be able to detect reading errors by the child. In a reading tutor, a reading error is defined as a word for which the final try by the child is wrong.

The test set for this task was selected from the Chorec database [7, 8], which contains reading sessions of Dutch speaking elementary school children from grade 1 to 4. For the evaluation in this paper, the reading session with 40 isolated, two-syllable words was chosen. The test set consists of 182 session recordings by different children of the same list of 40 words, so 7280 words in total.

The acoustic modelling is based on a 22 hour read speech database in Dutch, different from the Chorec database. It contains 16 kHz recordings of continuous sentences read or spoken by children aged between 5 and 11 years. Cross word context dependent acoustic models were estimated with 1343 tied HMM states and 16054 tied gaussian distributions in total (on average 94.6 gaussians per state). A straightforward Mel spectrum based signal processing scheme was adhered to, including mean subtraction on the log spectrum, discriminant analysis and Vocal Tract Length Normalisation (VTLN). When processing the Chorec data, recorded at 22050 Hz, the spectrum over 8 kHz is used for VTLN warping factors over 1.0. As the acoustic models had to be used in a system for tracking the child while reading RT tasks, we developed a VTLN system that does not result in any latency. This VTLN system is described in [9].

4. Evaluation of the FLavoR architecture

4.1. Large vocabulary task

For the evaluation of the systems on the large vocabulary WSJ task, we use the Word Error Rate (WER) as metric. The WER is the number of substitutions, deletions and insertions in the recognition result per 100 words in the reference transcription.

The results are given in table 1. We can see that the two-layered FLavoR setup performs slightly better than the baseline on both the 5k word and the 20k word task, but the differences are not significant. It seems that when the same modelling is used, both architectures produce almost the same result.

4.2. Reading tutor task

We evaluate the recognition systems on the reading tutor task by investigating the sensitivity, specificity and positive predictive value of the binary classifier that classifies every word as a reading error or not. So three measures are defined. The Miscue Detection (MD) rate is the proportion of the reading errors that was classified correctly as a reading error. The False Alarm (FA) rate is the proportion of the correctly read words that was classified incorrectly as a reading error. Finally the Feedback Error (FE) rate is the proportion of the words classified as a reading error, that is in fact not a reading error. The latter value is important in reading tutors: suppose the reading tutor gives feedback to the child for every reading error it detects, then the child will be confused in the case of words he or she read correctly. So the FE rate should be low (the results in this section show that this is difficult to obtain in general for all words in this reading task).

In figure 2, MD rate versus FA rate is shown, figure 3 plots MD rate versus FE rate. The one plot can be deduced from the other as the child error rate (the percentage of reading errors by the child) is known. The different values in these figures were obtained by changing the cost for entering the garbage model in the sentence FST shown in figure 1 at the bottom right. The higher this cost, the lower the MD rate.

It can be seen from the figures that for the reading error detection task, the two-layered architecture clearly outperforms the single-layered system. For example, for a 10% false alarm rate, the miscue detection rate increases significantly (p < 0.001%) from about 51% to about 63% by using the two-layered architecture, while the feedback error rate typically decreases by just over 10% absolute. This improvement is probably due to the fact that in the phone lattice in the two-layered system, only phone segments occur for which there is real acoustic evidence, as the full language model is not available in the first layer. Especially when detecting reading errors, which can be due to an error on a single phone, it is important to make sure that recognition results are at least based on some acoustic evidence.

5. Conclusions

In this paper, it was shown that phone lattice based speech decoding works well on real recognition tasks. On large vocabulary continuous speech recognition, evaluated on WSJ, it performs equally well as the classical single-layered architecture that uses the same modelling. On a reading tutor task, detecting
reading errors made by the child, the experiments in this paper showed that the two-layered setup outperforms the baseline significantly.

Overall, given also the evidence in [6] on a medium vocabulary task, we conclude that for recognition systems in which the acoustic modelling is based on phone models, the two-layered architecture will perform equally well or better on a wide range of recognition tasks. Moreover, the two-layered FLAVoR architecture has several additional advantages. The modular setup of the recognition system allows to re-use components when building a recogniser for a new task. The availability of the phone lattice allows to provide, at no extra cost, a more rich transcript for the word level recognition result. Also, more complex modelling, e.g. phone duration modelling, can be applied early in the recognition process, this may improve the recognition result of the two-layered architecture even further.

6. Acknowledgement

The research in this paper was supported by the IWT project SPACE (sbo/040102): SPEech Algorithms for Clinical and Educational applications, home page: http://www.esat.kuleuven.be/psi/spraak/projects/SPACE.

7. References


