USING PRINCIPLES OF LANGUAGE MODELING IN DECODING

David Andrš

University of West Bohemia, Faculty of Applied Science, Department of Computer Science and Engineering, Univerzitní 8, 306 14 Pilsen, Czech Republic
E-mail: `andrsd@kiv.zcu.cz`

Abstract: The goal of decoders in speech recognition is to get one or more hypotheses from phone estimations. Those phone estimations are computed for each part of speech signal (called *frame*).
Nowadays, decoders used in speech recognition task are based on two basic algorithms, Viterbi algorithm and A* search. Both algorithms use language models, but Viterbi algorithm can handle only bigram language models. A* search can handle language models of higher order, but the crucial part of it is the evaluation function. It directly determines how fast the search will be. Another disadvantage of Viterbi algorithm is that it perfroms quite slow on larger vocabularies.
In this paper, we propose new decoding method, which combines principles used in Viterbi algorithm and principles of language modeling used in A* search. The method is using a vocabulary which is recognised, so the resulting hypotheses are not restricted by a grammar or language models. The language models will be used in further phase – *rescoring*.
The method is using two level model – segment and word model. Words are automatically divided into segments, which consists of phonems. During decoding the recognition goes in opposite direction. Segments are built from phones and than words are built from segments.

Keywords: decoder, language models, czech

1 INTRODUCTION

In present days, decoders work with phonemes, diphones and triphones which are elemental units. Those units are trained from large speech corpus which consists of hundreds or thousands of hours. Obtaining such the corpus is very difficult and also it is very expensive and time consuming. The main problems are in having corpus with all possible difones or triphones with high frequecy of occurency – to have well trained models of such the diphones or triphones.

Diphones and triphones are used to incerease the performance of decoders – they are contextual models and it is the context which increases the performace.

1.1 Language Models

Language models are probabilistic models. They are trained from large text corpora. Their purpose is to estimate the following word from known history, usually of limited size – we denote it $n-1$. Then we speak about n-gram language models – $n-1$ words in history and one predicted word. In language modelling n is limited to 2 or 3 – then we speak about bigram and trigram language models. Trigram language model is computed by equation

$$P(w_1 w_2 w_3) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)$$

To obtain $P(w_1w_2w_3)$ is still very difficult and therefore it is approximated with frequency of occurency

$$P(w_1w_2w_3) \approx \frac{F(w_1w_2w_3)}{F(w_1w_2)}$$

1.2 Evaluating of Language Models

Language models are evaluated with two metrics – word error rate and perplexity (*Jelinek*). Word error rate says how well the model could predict the following word. Perplexity is the measure which expres the complexity of the model. In other words it is the avegare number of words which can follow every word in the lexicon (of the corpus).

2 COMPOSITE COARTICULATION MODELS

The base principle of composite coarticulation models is that words are not composed of isolated phonemes but it is composed from overlapping segments (those segments are similar to phonemes). The effect of overlapping is called *coarticulation*. If two segments overlap, we speak about diphone model. The model is *composed* from two segments.

2.1 Acoustic-Phonetic Decoding

The input of acoustic phonetic decoding is a sequence of phoneme probability estimation vectors. The goal of decoder is to find the hypothesis which fits the user utterance the best. This is performed either by Viterbi algorithm or by A^* search. The decoder could be modified to outputs N best hypotheses.

2.2 Modified decoder

The new proposed decoder has the same input data as current decoders have. The difference is in output. While current deocders outputs hypotheses, our decoder outputs only words. In other words, its goal is to extract words (when they occured in the utterance and the exact word – what was said) from phoneme probability estimation vectors.

The neccessary part of such the decoder is the lexicon of known words. Such the lexicon contains thousands or even hunderds of thousands words. Then extracing of words became very time comsuming and optimalizations must be done.

2.3 Segmentation

Segmentation is kind of optimalization to speed up extracting of words from vectors of phoneme estimations. The idea is to break words down into smaller pieces. The number of those pieces should be smaller than the number of words in the lexicon. The segmentation allows us to avoid processing longer words if only short word appears (problem of words *in* vs. *information*).

Segmentation of lexicon was done by two types of segmentation – with evolution algorithm and segmentation based on syllable-like units. The first one suffers with the following errors:

- it was very hard to add new word to the lexicon,

- improper segmentation (word *král* – in engl. king – was segmented like *krá – l*),
- very time-consumpting – the optimal segmentation is a global extreme where the nubmer of segments is small enough with high frequency of occurency and small perplexity of segments.

The segmentation based on syllable-like units is more convenient because:

- adding of new word is very easy – the word segmentation is based only on the word itself not on the other words,
- it mostly reflects morphology and therefore is very good for flective languages like Czech,
- there could be more than one segmentation of one word (this is very usefull when an error in recognition occurs – if the first model can not be used, then the other could be and no more application logic have to be done – it is simply putting the segments together).

2.4 Decoding Algorithm

As we have said, we use segmented lexicon, e.g. we use 2-level models – in the first level there are models of segments, in the second level words are composed from segments. Both levels use deterministic language models (they can determine if the tested combination is valid or not).

The algorithm performs according to the following steps:

- detection of coarticulation effects
- detection of isolated phonemes
- composing coarticulation effects and isolated phonemes into segments
- composing segments into words

The figure 1 shows coarticulation effects in Czech word *dva* – in engl. two. We can see development of three phoneme probabilities – for phonemes **d**, **v** and **a**. Time is denoted by *t*.

The complexity of the detection of coarticulation effect is very high. We have to compare each development of probability with other ones. This phase filters out unwanted speech material and reduces the size of data. Then we work with elements that are composed into segments using segment models. From those segments the words are built up.

3 CONCLUSIONS AND FUTURE WORK

We proposed a new decoding technique which uses phonemes and takes advantage of coarticulation effect in speech signal. The models used by this method could be trained from significantly smaller corpus than models used in diphone/triphone decoders.

This method is parametrized only by a lexicon of words which are recognized. The result is set of recognized words with information where they occured in the utterance. Such the result have to be rescored with language models to get the most probable hypotheses.

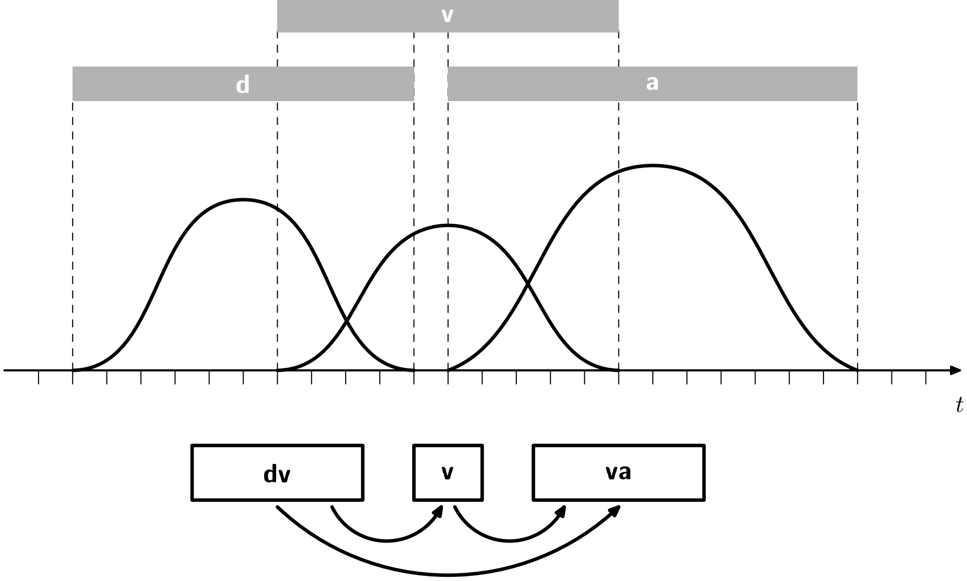

Figure 1: Effect of coarticulation and isolated phoneme in Czech word *dva* – in engl. two.

3.1 Future Work

Because the method is under heavy development and therefore no evaluation was done. One suitable metric for evaluation is error rate. The other metric used in language modeling – perplexity – is not very suitable because we process phonemes and almost every phoneme could be followed by another one, thus the perplexity is quite high. The perplexity metric could be used on word medel level, not in segment level.

REFERENCES

Jelinek, F., R.L. Mercer, R.L. Bahl and J.K. Baker. *Perplexity – a measure of difficulty of speech recognition tasks*, Journal of the Acoustical Society of America.