Speech Segmentation Aspects of Phone Transition Acoustical Modelling

Simon Dobrišek, France Mihelič, and Nikola Pavešić

University of Ljubljana, Faculty of Electrical Engineering,
Laboratory of Artificial Perception, Tržaška 25, SI-1000 Ljubljana, Slovenia,
simond@fe.uni-lj.si,
WWW home page: http://luz.fe.uni-lj.si/

Abstract. The paper presents our experiences with the phone transition acoustical models. The phone transition models were compared to the traditional context dependent phone models. We put special attention on the speech signal segmentation analysis to provide a better insight into certain segmentation effects when using the different acoustical models. Experiments with the HMM-based models were performed using the HTK toolkit, which was extended to provide proper state parameter tying for the phone transition models. All the model parameters were estimated on the GOPOLIS speech database. The annotation confusions concerning two-phone speech units are also discussed.

1 Introduction

A training process of an HMM-based speech recognition system incorporates an alignment of a speech model with a speech signal. The alignment results in a certain segmentation of the speech signal. Speech model parameters are usually estimated using an iterative training algorithm. The new estimation of the model parameters is therefore strongly affected by the segmentation obtained by the current model parameters.

If the acoustical modelling of speech is based on the phone models then even a slight change of the segmentation can change the estimation of the model parameters considerably. This happens due to the positions of the segment borders which are placed in the non-stationary transition parts of the signal. On the other hand, if the acoustical modelling of speech is based on the phone transition models, the segment borders are expected to be placed in the relatively stationary signal regions, and therefore the model parameters should not be affected that much by the change of segmentation.

We decided to investigate the differences in the acoustical modelling using traditional phone models and phone transition models.

2 Segmentation Analysis

A speech signal segmentation produced by a given acoustical model characterises this model in comparison the other models. A question that arises here is how to compare different signal segmentations and how to present the analysis results.

One possible solution is to extend the well-known problem of alignment of two strings of symbols to the problem of alignment of two sequences of labelled signal segments. We propose a variant of the string edit distance where the primitive cost function is composed of the primitive edit cost function and the additional distance function of a pair of signal segments. One of the most obvious distance functions to choose is

$$d_s(s1, s2) = \frac{|t_{b1} - t_{b2}| + |t_{e1} - t_{e2}|}{(t_{e1} - t_{b1}) + (t_{e2} - t_{b2})},$$

where t_{bi} and t_{ei} assign the beginning and end time ($t_{ei} > t_{bi}$) of the two signal segments ($i = 1, 2$). If the returned value is below 1.0 then the two segments ($s1$,$s2$) overlap at least for a short period of time.

The proposed variant of the primitive cost function returns the total number of edit operations at the optimal alignment which is comparable to the Levenshtein distance. It can be shown that the mentioned segmental-based string edit distance gives phonetically more consistent recognition score statistics, when comparing it to the traditional Levenshtein distance based speech recogniser assessment.

For all the pairs of segments in the segmental-based string edit distance which are declared to be a match or a substition, we can additionally define a function f_s which returns some information about how much the two segments are shifted relatively to each other. An example of such a function would be

$$f_s(s1, s2) = \frac{t_{b1} - t_{b2} + t_{e1} - t_{e2}}{(t_{e1} - t_{b1}) + (t_{e2} - t_{b2})}.$$

The presented segmental-based string edit distance in combination with the above function f_s provides a useful tool for obtaining some interesting speech signal segmentation statistics. All the segmentation histograms in the paper were generated using the described approach.

3 Speech unit annotation confusions

Biphones and diphones are both two-phone speech units which have the ability to capture co-articulatory effects. Biphones are understood to be just left or right context dependent (mono)phones. On the other hand, diphones represent the transition regions that strech between the two "centres" of the subsequent phones.

These two speech units are obviously different when we consider them together with the speech signal segmentation. The biphone acoustical models should produce similar signal segmentation as the (mono)phone acoustical models. On the other hand, the segmentation produced by the diphone models is expected to be shifted in time when we compare it to the segmentation produced by the biphone models.

If we observe biphones and diphones only from the aspect of phonetic transcriptions, without considering any signal segmentation, than it can be easily

shown that the difference between these two speech units is not that obvious any more.

For an example, let us observe the phonetical transcriptions of an isolated spoken command *Left!*. The possible different canonical transcriptions of the uttered command are the following:

	established annotation	optional annotation
left biphones:	`sil sil-l l-eh eh-f f-t t-sil`	`sil sil~l l~eh eh~f f~t t~sil`
right biphones:	`sil+l l+eh eh+f f+t t+sil sil`	`sil~l l~eh eh~f f~t t~sil sil`
diphones:		`sil~l l~eh eh~f f~t t~sil`

Without changing the definition of biphones the symbols '-' and '+' can be replaced by any other symbol, even by the symbol '~', which we use to annotate the transition nature of diphones. Let us assume that we allow, if required, splitting of the beginning or end silence segment into two parts that can be associated with the two subsequent speech units. It can be seen that all the above transcriptions that use the symbol '~' can be associated with signal segmentations which have a diphone transition nature.

4 Acoustical Modelling

Continuing the discussion from the previous section an interesting question arises about what signal segmentation can we expect when dealing with biphone and diphone acoustical models, respectively.

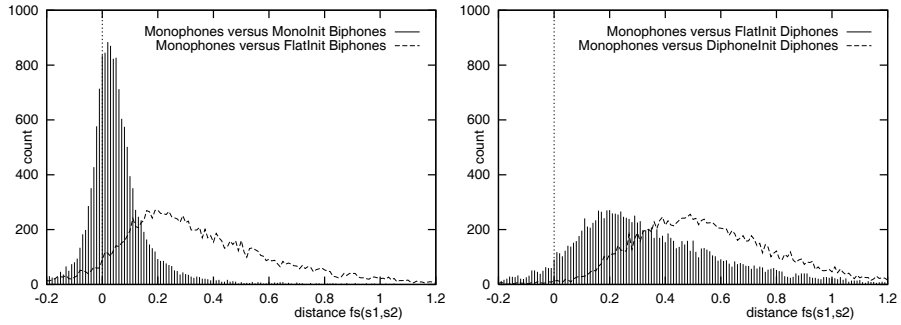

Fig. 1. The biphone and diphone segmentation histograms.

Several experiments were performed using the GOPOLIS speech database [2] and the HTK toolkit [4] which was extended to allow proper state parameter tying for the phone transition models as well. All the HMM models were built using the traditional approach.

First, we observed the differences in the signal segmentation produced by the monophone initialised biphone models and the flat started biphone models. After the initial training session which took approximately the same number of iterations through the whole available training speech data for both differently initialised models with approximately the same number of parameters,

the segmentation statistics was generated. The manual signal segmentation was compared to the segmentation produced after the forced alignment procedure.

From the histograms in Figure 1 it can be seen that the monophone initialised biphone models tend to preserve the segmentation given by the monophone models. However, the flat started biphone models produce a signal segmentation which is of a rather phone transition nature. Therefore one will conclude that the flat started biphone models are actually diphone models.

The diphone models were initialised from the diphone inventory of the first Slovenian text-to-speech system [3]. The flat started diphone models performed similary to the flat started biphone models.

The comparison between the diphone and biphone models was extended to the comparison between the bi-diphone and triphone models, where bi-diphones are just context dependent diphones. We encountered similar segmentation effect and confusion problems with the symbolical representation of these two speech units.

All the acoustical models mentioned in the paper were also incorporated into a continuously spoken word recogniser [1]. An interesting observation was that the diphone models had achieved considerably higher average log likelihood values per signal frame even when compared to the triphone models. Both models had approximately the same number of parameters. That was a good sign that indicated a high word recognition score. And the score was indeed higher for the diphone models [1].

5 Conclusions

The main conclusion would be that all even-numbered-phone speech unit models tend to produce signal segmentations which correspond to the transition signal regions. On the other hand, the monophone and odd-numbered-phone speech unit models tend to produce segmentations which correspond to the phone nuclei signal regions. This means that it is probably not wise to freely combine model parameters of these different speech units.

References

1. Dobrišek, S. (1999). Analysis and Recognition of Phones in Speech Signals. *Ph.D. Thesis in preparation*, (In Slovenian). University of Ljubljana, Faculty of Electrical Engineering, Ljubljana Slovenia.
2. Dobrišek, S., Gros, J., Mihelič, F., and Pavešić, N. (1998), Recording and labelling of the GOPOLIS Slovenian speech database. *Proc. 1st Int. Conf. on Language Resources & Evaluation*, Vol. 2, ESCA, pp. 1089-1096.
3. Gros, J., Pavešić, N., Mihelič, F. (1997), Text-to-Speech Synthesis: A Complete System for the Slovenian Language. *Jurnal of Computing and Information Technology*, Vol. 5(1), pp. 11-19.
4. Young, S., Odell, J., Ollason, D., Vatchev, V., and Woodland, P. (1997), *The HTK Book*. Cambridge University, Entropic Cambridge Research Laboratory Ltd.