DIPHONE SUBSPACE MODELS FOR PHONE-BASED HMM COMPLEMENTATION

Klaus Reinhard

University of Cambridge
Department of Engineering
Cambridge CB2 1PZ, England, U.K.
kr10000@eng.cam.ac.uk

Mahesan Niranjan

University of Sheffield
Department of Computer Science
Sheffield S1 4DP, England, UK
m.niranjan@dcs.shef.ac.uk

10.21437/Eurospeech.1999-345x

ABSTRACT

Considering the perceptual importance of phonetic transitions as minimal contextual variant units, this paper addresses the problem by modelling explicitly inter-phone dynamics covered in diphones. Subspace projections based on a time-constrained PCA (TC-PCA) are developed which focus on the temporal evolution. They reveal characteristic trajectories present in a low-dimensional spectral representation facilitating robust parameter estimation and simultaneously optimise the discriminant information. The applied multiple hypotheses rescoring scheme enables operating in very low-dimensional parameter space. Using such multiple hypotheses paradigm the complementary information effectiveness of modelling explicitly inter-phone dynamics covered in diphones can be shown using the TIMIT database, resulting in improved phone error rates.

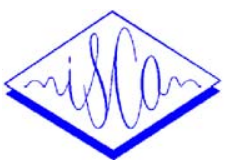
1. INTRODUCTION

One important feature of the complex temporal structure of speech signals is the systematic variation in the realisation of phones in different acoustic contexts. In fluent speech the target position towards which the articulators move may often not be realized. In the corresponding acoustic signal, it would then be hard to isolate steady state regions that could be uniquely identified with phones. Discriminatory information enabling the decoding process is not localised in the steady states, but is likely to be smoothly distributed in the sequence of transitions of the signal.

State-of-the-art systems try to model such complex contextual structure by using specialised models and enlargements of the parameter space. This leads to the dilemma most successful systems face these days. Due to the high dimensionality of the used parameterisation and the huge number of models robust parameter estimation becomes the challenging task. Clearly, a phone model for the vowel [i:] derived from all contexts would be noisy, due to the different spectral trajectories into the vowel [i:], for example in the CV transition /bee/ and /gee/. Hence we start from a slightly different premise that attempts to focus on the transition between phones. Diphone units capture these transitions being defined as half of one phone followed by half of the next phone. While the number of segments to model increases rapidly, the hope is that one has a greater chance of capturing the transitional information explicitly.

The work described in this paper is an extension of our earlier approach to model speech transitions in a subspace where the temporal ordering is preserved which may be found in [3]. The paper utilises the multi-trajectory concepts to rerank multiple hypotheses generated by a baseline HMM which is augmented by derivatives in the parameterisation. The methods of trajectory clustering and subspace selection are shown to optimise our subspace models. Using a 2-dimensional diphone subspace trajectory score in combination with the HMM score, reranking of the hypotheses is performed by optimising the weighting of the complementary between-phone information on a speaker by speaker basis.

Findings suggest that complementary information is retained when augmenting the baseline HMM with diphone information. We illustrate this on six speakers from the TIMIT database. Tuning the weighting parameter, reranked hypotheses are selected which resulted in an average relative improvement of 4.8%. The result represents an average utilisation of 31.2% of the available accuracy improvement within the multiple hypotheses.



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

2. DIPHONE SUBSPACE TRAJECTORIES

Projecting a sequence of short term spectral parameters onto a subspace with $L \leq 3$, where the temporal sequence of these vectors are preserved, makes it possible to visualise and model trajectories of important speech dynamics. The parameter requirements for such a subspace model is reduced to $L \times (n+p)$, where p is the dimensionality of the spectral representation and n is the average number of spectral frames for a speech unit. An adaptation of the well known technique for dimensionality reduction, Principal Component Analysis (PCA) [1], is used to generate projections onto a L-dimensional subspace where the temporal ordering of the data sequence is preserved. This method is called time-constrained PCA (TC-PCA) [3]. In order to preserve the temporal sequence information, the dimensionality of the speech parameterisation is expanded by one. Defining $\mathbf{y}_n = [y_0^n, y_1^n, \ldots, y_p^n]$ the extended frame parameterisation consists of $y_0^n = \tau * n$ with $n = 1 \ldots N$ being the frame number of a diphone sequence consisting of N frames. Each frame is initially parameterised by p MFCC coefficients. y_0^n represents a scalable temporal ordering. Hence a parameterisation of a diphone segment is expanded by an additional dimensionality. The extra dimension representing a scalable frame ordering as the time constraint. The scale factor τ is introduced to control the weighting imposed by this arbitrary choice of incorporating the order information.

2.1. Trajectory Mixtures

Characteristic trajectories captured in diphones are modelled in the high-dimensional spectral parameterisation. Optimising the trade-off between parameter requirements and accuracy performances, the most suitable subspace can be chosen using any best subspace projection onto a l-dimensional subspace $l < p$ by adaptation of the projection matrix. Trajectory templates for model parameters are obtained using a maximum likelihood criterion, that maximises $P(\mathbf{t}_1^N|a)$. The re-estimation formulas based on the EM algorithm were derived by Fukada et al. [2], who maximised the following auxiliary function Q:

$$\begin{aligned} Q(\bar{\Phi}|\Phi) &= \mathcal{E}[logP(\mathbf{t}_1^N, k|\bar{\Phi})|\mathbf{t}_1^N, \Phi] \\ &= \sum_{k=1}^{M} \frac{P(\mathbf{t}_1^N, k|\Phi)}{P(\mathbf{t}_1^N|\Phi)} logP(\mathbf{t}_1^N, k|\bar{\Phi}) \end{aligned} \quad (1)$$

where Φ and $\bar{\Phi}$ are the sets of the current model parameter and the re-estimated model parameter. Maximising Eq. 1 will lead to the different model parameters for k different mixture components. The results of a k-means trajectory clustering are used as initial model parameters for the EM approach.

2.2. Trajectory Scoring

When a test trajectory $\mathbf{y}$ is received, it is time warped to match the length of the template. This enables the scoring of test trajectories of different length. The model score is defined by a distance score which is the result of the best trajectory model match $\mathcal{D}_a$. The trajectory is than classified by finding the diphone model with the best score, discriminating between all competing diphones within the set of models. The scoring is performed in a arbitrary subspace. The parameterisation of the trajectories used 5 MFCCs only to avoid cepstral coefficients with an oscillating nature. The projection matrix is applied to the templates and the test trajectory before the scoring process is performed. The distance score $\mathcal{D}_a$ for a individual diphone a can be expressed as:

$$\mathcal{D}_a(\mathbf{y}, \boldsymbol{\mu}^k) = \min_k \left\{ \frac{\sum_{i=1}^{N} \|\mathbf{y}^i - \boldsymbol{\mu}^{k_i}\|^2}{\text{arclength}(\mathbf{y})} \right\}. \quad (2)$$

2.3. Optimal Subspace

Transformation plane selection is performed using the data for all competing models in parallel to obtain a plane which utilises between-diphone discriminant information. Considerations of all possible subspace plane combinations for all diphone models in parallel is not desirable due to the computational costs $\mathcal{O}(l^m)$ to select the appropriate plane. But there is still the need to find a plane which is most competitive considering other diphones optimising the between-diphone discrimination. The maximum relative distance (MARD) method is used to calculate a relative goodness of discrimination for a particular diphone model. The method finds the projection plane which results in the most discriminant transformation considering all other training trajectories. Given the training set $\mathcal{T}_a$ of diphone a the MARD plane index which corresponds to a certain time constraint can be computed using:

$$PI_{\tau_a} = \max_{\tau_a} \left\{ \sum_{d=1}^{N_{di}} \left(\frac{\sum_{j=1}^{D_{\tau_{d \neq a}}} \mathcal{D}(\mathbf{y}_\tau^j, \boldsymbol{\mu}_\tau^{a_k})}{\sum_{j=1}^{D_{\tau_a}} \mathcal{D}(\mathbf{y}_\tau^j, \boldsymbol{\mu}_\tau^{a_k})} \right) \right\}. \quad (3)$$

Because this algorithm considers for each model all available training data the computational costs is $\mathcal{O}(m^2 l)$, Yet it constitutes a cheaper alternative in comparison to an exhaustive search through all possible diphone and plane combinations.

3. MULTIPLE HYPOTHESES PARADIGM

The multiple hypotheses rescoring paradigm involves the generation of a list of N best hypotheses by a recognition system and subsequent rescoring of this hypothesised list using other knowledge sources. The constrained recognition search is particular useful for trajectory models based on diphones, which in comparison to phones have a significantly larger recognition search space. Complementary information sources can enhance such systems by using this knowledge to decide between ambiguous situations when several models are hypothesised for the same segment. Although a list of hypotheses can be considered for a variety of models, like sentences, words or phonemes, in this case the interest is focused on phoneme hypotheses because they can be related to diphone segments as an additional knowledge source.

3.1. Diphone Rescoring

A typical multiple hypotheses list of phones is provided by the baseline HMM containing information which can be converted in a new list of diphone hypotheses as follows. The probability scores are combined by averaging the two phone segment probability scores to obtain the HMM based diphone score given by:

$$\log(P^{di}_{hmm}) = \frac{\log(P^{ph1}_{hmm}) + \log(P^{ph2}_{hmm})}{2}. \qquad (4)$$

To keep the ordering in the correct sequence according to the maximum of the sum of diphone scores, the first and the last frame have to be treated differently. Here the total score of the first and last phone contribute to the initial and final score respectively.

The previously defined distance score defined is proportional to the log-likelihood of the hypothesised model. Hence, it enables a combination of HMM-based and trajectory-based scores.

$$\log(P^{di}_{sub}) \propto -\mathcal{D}(\mathbf{y}, \boldsymbol{\mu}). \qquad (5)$$

The baseline HMM and trajectory scores are then combined to rerank the diphone hypotheses. The optimum scores for each segment are computed by a linear combination of the HMM-based diphone score $\log(P^{di}_{hmm})$ and the subspace trajectory based diphone score $\log(P^{di}_{sub})$.

$$\log(P^{di}_{new}) = (1-\beta) * \log(P^{di}_{hmm}) + \beta * \log(P^{di}_{sub}). \quad (6)$$

The different choices of β result in different re-ranked lists of diphone hypotheses which can be translated back to a list of hypothesised phones. The phone lists can be evaluated and an optimal β is chosen on a speaker to speaker basis to obtain the best recognition accuracy results. If $\beta = 0$ the best result is the first hypothesis in the N-best list chosen by the HMM system. Increasing the influence of the trajectory score can allow the best hypothesis to be found. For $\beta = 1$ all segment hypotheses will be re-evaluated based on the trajectory score alone. In the case of a hypothesised diphone which is not covered by the modelled diphone repertoire the original HMM score is kept. In the next section the evaluation is reported for both baseline HMMs.

4. EXPERIMENTAL WORK

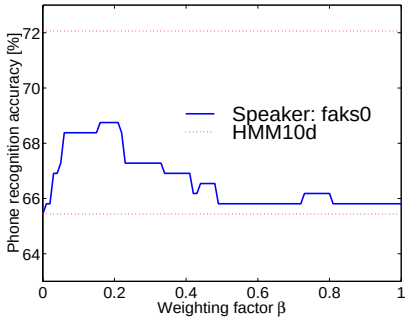
The experimental illustrations focus on the potential importance of between-phone transitions and the complementary power to enhance baseline HMMs augmented by delta coefficients. Therefore the used baseline HMM models 3-state monophones. The number of Gaussian used in the mixture distributions to model the acoustic observations are set to 10. The acoustic feature vector to represent a frame of speech used 13 mel-frequency cepstral coefficients and its first derivatives (HMM10d). Since the aim of this evaluation is to compare the acoustic modelling ability of inter-phone characteristics, a "null" grammar was used which assumes that there is no restriction in the sequence of lexical entities and all sequences of phones are equally likely. In the experiments described, the number N of examined hypotheses was set to 100 whereas the maximum number of tokens kept in each HMM state was set to 10 during the decoding of each utterance. The achieved phone recognition error rates for the individual speakers are given in Table 1, distinguishing between the 1-best and the 100-best solution available. As a result of the rescoring process using a combined HMM-DST score, the achieved phone recognition error rates are given. From the absolute values the relative improvements and the relative achievable improvements are calculated which shows how much of the offered improvements covered in the

Speaker	Baseline HMM errors	HMM-DST error	Relative Improv.	Rel. ach. Improv.
faks0	34.6%(27.9%)	31.6%	8.5%	44.4%
mdab0	50.7%(46.4%)	49.3%	2.8%	33.2%
fcmr0	54.0%(45.6%)	50.8%	5.9%	38.1%
mabw0	35.5%(28.7%)	35.8%	-1.1%	-5.6%
fcmh0	32.6%(26.2%)	29.6%	9.2%	47.0%
mcsh0	35.3%(31.6%)	34.2%	3.2%	30.1%

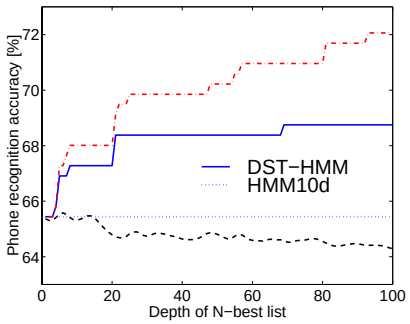
Table 1: The relative reductions in phone recognition error rates and relative achievable reduction in phone recognition error rates for baseline HMM.

multiple hypotheses could be utilised. The resulting average performance improvements over all speakers was 4.8% relative improvement, utilising 31.2% of the available accuracy.

The used diphone subspace trajectory (DST) score is obtained in a 2-dimensional subspace which results in a minimum parameter requirement for such models. The rescoring process is illustrated in Figure 1 for TIMIT speaker `faks0`. A best weighting parameter β is found by considering a list of 100 hypotheses. Increasing gradually the weighting for the DST score an optimum β is determined. It is used to rescore the N-best lists for various depths as depicted in Figure 1. The figure shows furthermore the baseline performance of the HMM and the evolution of performance increase including more hypotheses. As a reference performance, the average accuracy obtained over randomly selected hypotheses is depicted as a dashed line.



(a) Accuracy monitor for β



(b) Rescoring results using best β

Figure 1: Evaluation of the choice of β and the corresponding N-best rescoring result. A best β is selected achieving maximum phone recognition accuracy within N=100 hypotheses shown in (a). The results for rescoring with increasing number of included hypotheses are given in (b).

5. DISCUSSION

In this paper we presented a multiple hypotheses rescoring scheme based on diphone subspace trajectories which resulted in significantly improved phone recognition error rates. Such improvements could even be obtained using a baseline HMM which was augmented with derivatives which underlines the potential importance of explicitly modelling between-phone transitions. Nevertheless, these findings are limited by various factors which are directions of necessary future work. One limitation is the covered diphone repertoire which was restricted to a set of 228 diphones within TIMIT having at least 50 test tokens. The restriction influences the reranking process but the choice of small optimal β suggest that combined scores will be of similar magnitude in comparison to the HMM score alone. Further limitations can be found in the optimisation process to find the best ratio between the DST score and the HMM score. The parameter is tuned on a speaker to speaker basis but should ideally be determined by using a cross-validation set. But as already suggested by Schwartz et al. [4] such a cross-validation set should contain at least 300 utterances to reliably determine values for the weighting parameters used in rescoring schemes. Summarising, we demonstrated that diphone subspace trajectory models contain useful discriminant information. Despite the simplistic approach of using a distance score in a 2-dimensional subspace, the modelled between-phone characteristics could enhance baseline HMM augmented with derivative information. The potential enhancements to make DST models more effective will be investigated in the future.

6. REFERENCES

[1] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. New York, Wiley, 1973.

[2] T. Fukada, Y. Sagisaka, and K.K. Paliwal. Model parameter estimation for mixture density polynomial segment models. *Int. Conference on Acoustics, Speech and Signal Processing*, 2:1403–1406, 1997.

[3] K. Reinhard and M. Niranjan. Parametric subspace modeling of speech transitions. *Speech Communication*, 27:19–42, 1999.

[4] R. Schwartz, S. Austin, F. Kubala, J. Makhoul, L. Nguyen, P. Placeway, and G. Zavaliagkos. New uses for the N-best sentence hypotheses within the BYBLOS speech recognition system. *Int. Conference on Acoustics, Speech and Signal Processing*, 1:1–4, 1992.