PHONETICALLY SEEDED SCHMM OF VARIABLE LENGTH FOR SPEAKER INDEPENDENT RECOGNITION OF ISOLATED WORDS

Florian Schiel

Lehrstuhl für Datenverarbeitung,
Technische Universität München, Germany

ABSTRACT – Most applications of HMMs of words use a constant number of states for each model, which are trained by the well-known Baum-Welch-Algorithm. We propose a new approach with phonetically seeded word models of variable length, that are trained by a Viterbi-like algorithm. Our basic motivation was the idea that every state of a word SCHMM should model a phonetic unit of the word. We can achieve this by creating seed models, where the mixture coefficients of each state describe the features of a phonetic segment of the word. A phonetic segment should be a segment where the feature *loudness spectrum* (one of 3 feature vectors) behaves quite stationary. A stationary signal leads to a stationary result of the Vector Quantization (VQ) of the loudness spectrum. It is sufficient to investigate the two best results of the semi-continuous VQ over time for stationary behaviour. We define a segment as a time interval, where at least one of a pair of codebook (CB) symbols (m_1, m_2) is observed in the first or second position of the VQ result. Furthermore there is information about the position of the vowels by detecting maxima in the smoothed loudness function according to [Zwicker, E., 1982]. The segmentations found in different training utterances are matched into one common seed model by an algorithm that tries to identify corresponding phonetical segments correctly. The seed model is re-trained with all training utterances by a Viterbi-like algorithm similar to the segmental-k-means. Although the number of training utterances was only 12 - 36 (6 speakers) we achieved very good results compared with standard HMM of fixed length.

1. INTRODUCTION

Most practical applications of speech recognition are designed to be used by more than one user and require high recognition rates, despite of being trained with a reduced database. On the other hand it is often sufficient to recognize isolated words of an amount of less than 1000 lexical words. Under these circumstances classification by semi-continuous Hidden Markov Models (SCHMM) for whole words achieve good performance even in speaker independent mode.

Most applications of HMMs of words use a constant number of states for each model, which are trained by the well-known Baum-Welch-Algorithm. We propose a new approach with phonetically seeded word models of variable length, that are trained by a Viterbi-like algorithm. In this way we achieve models with more states for longer words and vice versa. Another advantage is that we can use the Viterbi algorithm for testing and hence it is possible to connect word models for recognition of fluent speech.

In the following section the extraction of the features and the data used for training and testing is briefly described. In section 3 we present the algorithm for a phonetic segmentation, which is the base for the creation of seed models of variable length. Section 4 is concerned with the Viterbi-like training of the seed models and in the 5th section some results of testing are briefly discussed.

2. PREPROCESSING AND DATA

The task is to recognize isolated words in fluent German speech. The signal is preprocessed into a so-called loudness spectrum S according to [Zwicker, E., 1982] of 20 bark every 10 msec ([Ruske, G. & Beham, M., 1991]). From that the total loudness L and the derivation of the loudness spectrum D are calculated. This yields the 3 feature vectors S, L and D.

For each feature vector $\underline{k}(t)$, $k = S, L, D$ a codebook of 64 prototypes is calculated by LBG ([Linde, Y. & Buzo, A. & Gray, R.M., (1988)]). Each prototype $m = 1 \ldots 64$ consists of a mean vector $\underline{s}_k$ and covariance matrix $S_k = [\sigma_{kij}]$, $i, j = 1 \ldots 64$, but the covariances σ_{kij}, $i = j$ are set to zero.

The semicontinuous vector quantization (VQ) calculates the weighted Euclidean distance to the R next prototypes $m = 1 \ldots R$ and – by inversion and normalization to 1 – a weighting $P_k(t, m)$ every time interval $t = 1 \ldots T$.

$$P_k(t,m) = \frac{\frac{1}{(\underline{k}(t)-\underline{s}_k(m))' S_k(m)^{-1} (\underline{k}(t)-\underline{s}_k(m))}}{\sum_{i=1}^{R} \frac{1}{(\underline{k}(t)-\underline{s}_k(i))' S_k(i)^{-1} (\underline{k}(t)-\underline{s}_k(i))}} \tag{1}$$

In the training R is 64 (all), in the test R is variable. We do not use the correct Bayesian probability as required of the theory ([Huang, X.D. & Jack, M.A., 1988]), because this leads to models, which are 'overtrained' to the training database and therefore not efficient for speaker independent mode. The VQ of eqn. 1 causes smooth distributions of the mixture coefficients and improves the recognition results ([Streicher, M., 1991]).

The word models are semicontinuous hidden Markov models (SCHMM) according to [Huang, X.D. & Jack, M.A., 1988]. The emission probability $E(z)$ of a state z within a SCHMM is calculated as a product code of the emission probabilities in the three features:

$$E(z) = E_S(z, \underline{P}_S(t))\, E_L(z, \underline{P}_L(t))\, E_D(z, \underline{P}_D(t)) \tag{2}$$

The emission probability of one feature k is ([Huang, X.D. & Jack, M.A., 1988])

$$E_k(z, \underline{P}_k(t)) = \sum_{m=1}^{R} P_k(t,m)\, Q_k(z,m) \tag{3}$$

where $Q_k(z, m)$ is the mixture coefficient of the feature k in the state z for the VQ symbol m.

The database used for training consists of 100 German sentences with an amount of 341 words. Each sentence is spoken twice by 3 male and 3 female speakers. Therefore the minimum number of utterances of each word is 12, the maximum was set to 36. The word boundaries were segmented by hand. The database used for speaker independent testing has the same structure but is spoken of two unknown male speakers.

3. PHONETIC SEGMENTATION AND SEED MODELS

Motivation

Our basic motivation was the idea that every state of a word HMM should model a phonetic unit of the word. We can achieve this by creating seed models, where the mixture coefficients of each state describe the features of a phonetic segment of the word. After that the seed models can be trained by an Viterbi-like algorithm to re-estimate the mixture coefficients and to estimate the transition probabilities. Hence we need an algorithm to detect the number and boundaries of phonetic segments within the training data and to average the observed feature vectors of corresponding segments into the states of a word SCHMM.

Phonetic segmentation

A phonetic segment should be a time interval in which the feature S behaves quite stationary. A stationary signal leads to a stationary result of the Vector Quantization (VQ) of the feature S. We found that it is sufficient to investigate the two best results of the VQ over time. So, we define a phonetic segment as

> phonetic segment : a time interval, where at least one of a pair of codebook symbols (m_1, m_2) is observed in the first or second position of the VQ result

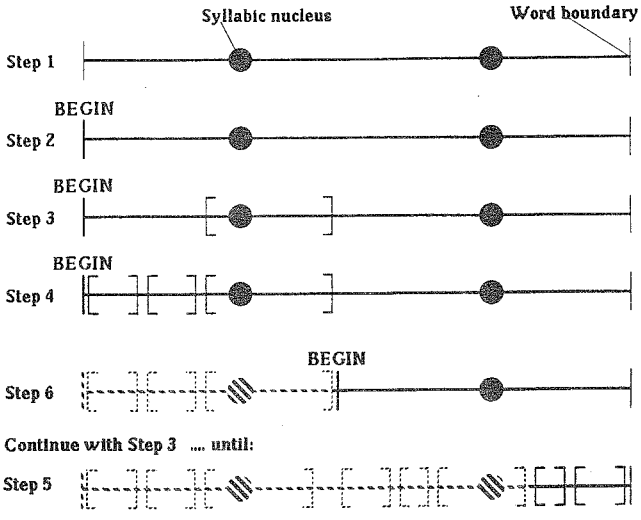

Figure 1: Phonetical segmentation of a word (see text).

Furthermore we have information about the position of the syllabic nuclei by detecting maxima in the loudness function L over time ([Ruske, G., 1988]).

The algorithm for the phonetic segmentation works as described in the following (see fig. 1):

1. Detect all syllabic nuclei v within the word.
2. Set *BEGIN* to the beginning of the word (1st frame).
3. Find the biggest segment around the next syllabic nucleus after *BEGIN* according to the definition above.
4. Find the best segmentation from *BEGIN* to the segment found in 3 by the following rules:
   - Each segment has a minimum length of 2 frames.
   - Segments do not overlap.
   - Start with the biggest possible segment and go on to smaller segments.
5. If there is no further syllabic nucleus detected, find the best segmentation for the rest of the word according to step 4 and stop.
6. Otherwise move *BEGIN* to the right boundary of the segment found in 3 and continue with step 3.

We proofed the outcome of the segmentation by acoustic control of the corresponding speech signal of the detected phonetic segments and found that in most cases the segmentation was correct.

Condensing the phonetic segments to seed models

Because the segmentation is done to all training words the problem arise that there are different ways to segment different utterances of the word and there is no information to condense the 'right' segments into the same state of the seed model. Another problem is the fact that sometimes the syllabic pre-segmentation mentioned in section 3 detects different number of vowels in the training utterances of the same lexical entry. The reason for that can be for instant an elision of a vowel caused by fluent speaking, which is therefore not detected by the syllabic pre-segmentation.

We solved the problem in the following way:

1. Divide the training utterances of a lexical entry into W groups, in which all utterances have the the same number v_w, $w = 1 \ldots W$ of detected syllabic nuclei. Do the following steps for each group of utterances (see fig. 2).
2. For a group of training utterances with v_w syllabic nuclei create a pre-defined seed model with a maximum number of states $n_{w\,max}$

$$n_{w\,max} = 4\,v_w + 3 \tag{4}$$

   The transition probabilities within the seed model are all set equal.
3. The data observed within the phonetical segments around the syllabic nuclei are averaged into the states $4, 8, 12 \ldots$ respectively.
4. The remaining segments are averaged into the 3 states before the first vowel state, after the last vowel state and between the vowel states depending of the number of these segments:
   - *One segment* is averaged to the *first* and *third state.*
   - *Two segments* are averaged into the *first* and the *third state* respectively, and both to the *second state.*
   - *Three segments* are averaged on the *states in right order.*
   - *More than three segments*: the *first segment* is averaged to the *first state*, the *last* to the *third state*, the *remaining* to the *second state.*
5. After condensing the data of all utterances into the pre-defined seed model the states without any data are removed.

By this algorithm we achieve one seed model of variable length for each group of utterances with differing numbers of syllabic nuclei in the training data base.

4. VITERBI TRAINING ON SEED MODELS

The seed models still have no information about the dynamics of the speech signal: the transition probabilities are all equal and there is no information about the non-stationary gaps between the found segments. Therefore all seed models are trained with all training utterances by a Viterbi-like algorithm similar to the segmental-k-means ([Juang, B.H. & Rabiner, L.R., 1990]). Details about this training can be found in [Streicher, M., 1991].

If there exist more than one seed model for one word ($W > 1$ because of different number of detected syllabic nuclei), these W models of different length are tested by Viterbi on all training utterances. The model with the highest emission probabilitiy in average is kept, while the others are removed.

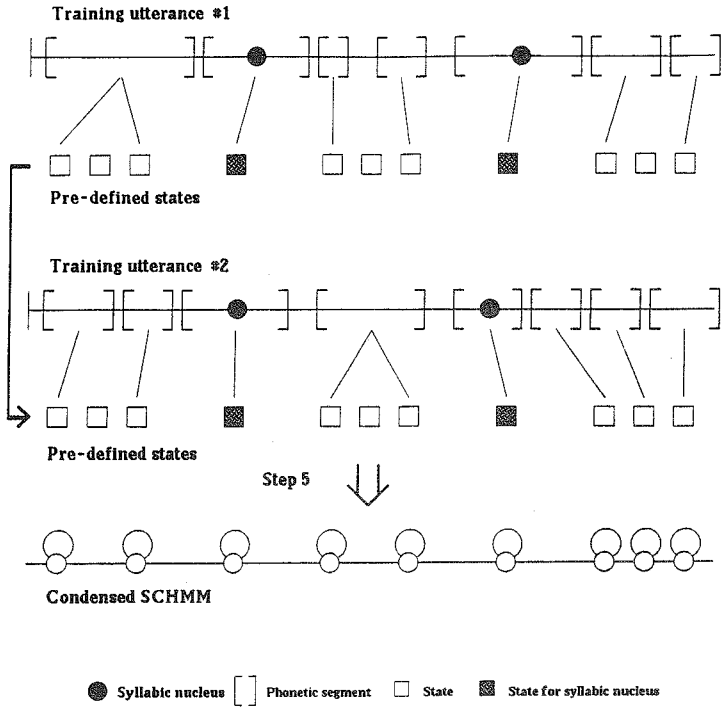

Figure 2: Condensation of phonetic segments to states of a SCHMM: two utterances with 2 syllabic nuclei each are mapped to the $4 \times 2 + 3$ pre-defined states of the seed model. The condensed SCHMM contains only 9 states, because no data were mapped in 2 pre-defined states.

5. TESTING AND RESULTS

The word models were tested with one male speaker of the training corpus (*speaker dependent mode*) and two male speakers that are not in the training corpus (*speaker independent mode*, see section 2). There are 341 words in the lexicon.

At first the number of codebook symbols for the three features S, D and L was varied. We found out that in speaker independent mode a number of 64 symbols is sufficient. Higher numbers let the recognition rates increase in the speaker dependent mode, but decrease in the speaker independent mode. The reason is probably that the models get 'overtrained' for the 6 training speakers.

Then the number of calculated VQ weights R (see section 2) was reduced. We found an optimum at $R = 3$ in *speaker independent mode* which contradicts to the theory that R should be the number of codebook symbols ([Huang, X.D. & Jack, M.A., 1988]) as it is in the training. A reason for this could be the very smooth calculation of the weighting calculated in the VQ (see section 2). Table 1 shows the recognition rates in the speaker independent mode as a function of R. The best result

R	1	2	3	4	8
Recognition Rate	79.35	79.79	80.53	80.24	78.47
Within the best 5	94.34	95.28	94.69	94.25	93.51

Table 1: Recognition rates in speaker independent mode

in the *speaker dependent mode* of 96.46 % was achieved with $R = 1$.

6. CONCLUSION

A new approach for designing semi-continuous HMMs by phonetical seeding was proposed. An automatic segmentation of the training utterances leads to seed models with variable length, which are trained with segmental-k-means. The approach was tested with a lexicon of 341 words cut out from fluent speech and achieved 96.46 % in speaker dependent and 80.53 % in speaker independent mode. There is still a gap between speaker dependent and speaker independent mode. Therefore this approach should be combined with a speaker adaptation of the applied SCHMMs themselves and/or the used codebooks resulting in a better performance in the speaker independent mode ([Schiel, F., 1992]).

REFERENCES

[Huang, X.D. & Jack, M.A., 1988] *Hidden Markov Modelling of Speech Based on Semicontinuous Model*, Electronics Letters, Vol. 24, No. 1, pp. 6 - 7.

[Juang, B.H. & Rabiner, L.R., 1990] *The segmental k-means algorithm for estimating parameters of Hidden Markov Models*, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 38, No. 9, pp. 1639 - 1641.

[Linde, Y. & Buzo, A. & Gray, R.M., (1988)] *An Algorithm for Vector Quantizer Design*, IEEE Trans. on Comm., Vol. COM-28, No. 1, pp. 84 - 95.

[Ruske, G., 1988] *Automatische Spracherkennung*, pp. 117 - 123, (R. Oldenbourg Verlag München Wien).

[Ruske, G. & Beham, M., 1991] *Gehörbezogene automatische Spracherkennung*, in *Sprachliche Mensch-Maschine-Kommunikation*, pp. 33 - 48, (Oldenbourg-Verlag München Wien).

[Schiel, F., 1992] *Rapid Non-Supervised Speaker Adaptation of Semicontinuous Hidden Markov Models*, Proc. of the International Conference on Speech and Language Processing, Banff, Alberta, (in print).

[Streicher, M., 1991] *Entwicklung eines Ganzwort-Erkenners auf der Basis silbenorientierter Startmodelle*, thesis at the Lehrstuhl für Datenverarbeitung, (Technische Unversität München).

[Zwicker, E., 1982] *Psychoakustik*, (Springer Verlag, Berlin Heidelberg New York).