MAXIMUM LIKELIHOOD BASED TEMPORAL FRAME SELECTION

Tingyao Wu, Dirk Van Compernolle, Jacques Duchateau, Hugo Van hamme

Katholieke Universiteit Leuven – Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{Tingyao.Wu, Dirk.VanCompernolle, Jacques.Duchateau, Hugo.Vanhamme}@esat.kuleuven.be

ABSTRACT

In this paper, we propose a maximum likelihood (ML) based frame selection approach. A fixed frame rate adopted in most state-of-the-art speech recognition systems can face some problems, such as accidentally meeting noisy frames, assigning the same importance to each frame, and pitch asynchronous representation. As an attempt to avoid those problems, our approach selects reliable frames from a fine resolution along the time axis. In a phoneme recognition task, we show that significant improvements are achieved with the frame selection approach comparing to a system with a fixed frame rate.

1. INTRODUCTION

Most speech recognition systems use a fixed frame shift, typically 10msec, to decompose speech into a series of frames, basically due to its convenience. But some limitations appear with this arbitrary and fixed frame rate. For instance, a noisy frame may dominate the recognition process; the same importance is assigned to each extracted frame, which has been shown inconsistent with human perception [1]. Besides, pitch asynchronous representation [2, 3] caused by the fixed frame rate leads to pitch mismatch due to the presence of pitch-related harmonics in the power spectrum. Because of those limitations, researchers are looking for better power spectral estimates that are less sensitive to frame position, such as [4, 5]. The frame selection technique proposed in this paper is an alternative solution.

Two methods are proposed to select reliable frames along the temporal axis, which are both based on an ML criterion. The first method, called *multiple selection* selects frames from a tiny frame shift after which the average frame rate matches a pre-defined value. The procedure can be incorporated into a modified Viterbi decoding algorithm. Unlike the traditional Viterbi algorithm, the modified algorithm does not strictly stick to the first and the last frame of a segment, and allows jumping between two non-adjacent frames, which is penalized by a time compensation coefficient. The other method, named *single selection*, selects only one frame for each state of an Hidden Markov Model (HMM). The number of selected frames for a testing unit depends on the topology of the HMM. In [6], we have shown that for certain tasks, three frames arbitrarily picked from the beginning part, the middle part and the ending part of a vowel duration are a better representation than all frames. The *single selection* is an extension of our previous work: instead of selecting frames at fixed positions, it allows to select a single frame for each state dynamically.

The organization of this paper is as follows. The theoretical introduction of *multiple selection* and the description of *single selection* are presented in Section 2. The experimental results will be reported in Section 3 and a summary of the conclusions and future work are given in Section 4.

This work is sponsored by the Fund for Scientific Research Flanders (FWO-project G.0249.03), by Research fund (onderzoeksfonds) K.U.Leuven, project nr. OT/03/32/TBA and by the IWT project SPACE (sbo/040102): SPeech Algorithms for Clinical and Educational applications.

2. METHODOLOGY

2.1. Multiple selection

Starting with a tiny frame shift, k_1, we require that after frame selection, the average frame shift is equal to a pre-defined intended frame shift k_2. This implies that one frame is selected out of $K = \frac{k_2}{k_1}$ frames on average. This selection can be embedded into a modified Viterbi algorithm for speech decoding. Comparing to the traditional Viterbi algorithm, two differences are highlighted. One is that the modified Viterbi algorithm is not a frame-by-frame alignment; jumping between two non-adjacent frames is also allowed. To ensure the average long-term frame rate, the range for searching the best previous frame for the frame at time t is $[t-(K+\lfloor\frac{K}{2}\rfloor), t-(K-\lfloor\frac{K}{2}\rfloor)]$. The other modification to the Viterbi algorithm is that a time compensation factor is used to penalize the time lag between two selected frames. Notations:
N: the number of states
T: the number of frames
π_i: the initial log probability that the Markov chain will start in state i;
$\mathbf{X}_t$: the observation at time t
$b_i(\mathbf{X}_t)$: the log probability of emitting feature vector $\mathbf{X}_t$ when state i is entered
$V_t(i)$: the log probability of the most likely state sequence at time t in state i.
Transition probabilities are ignored. Four extra notations are added for the frame selection procedure.

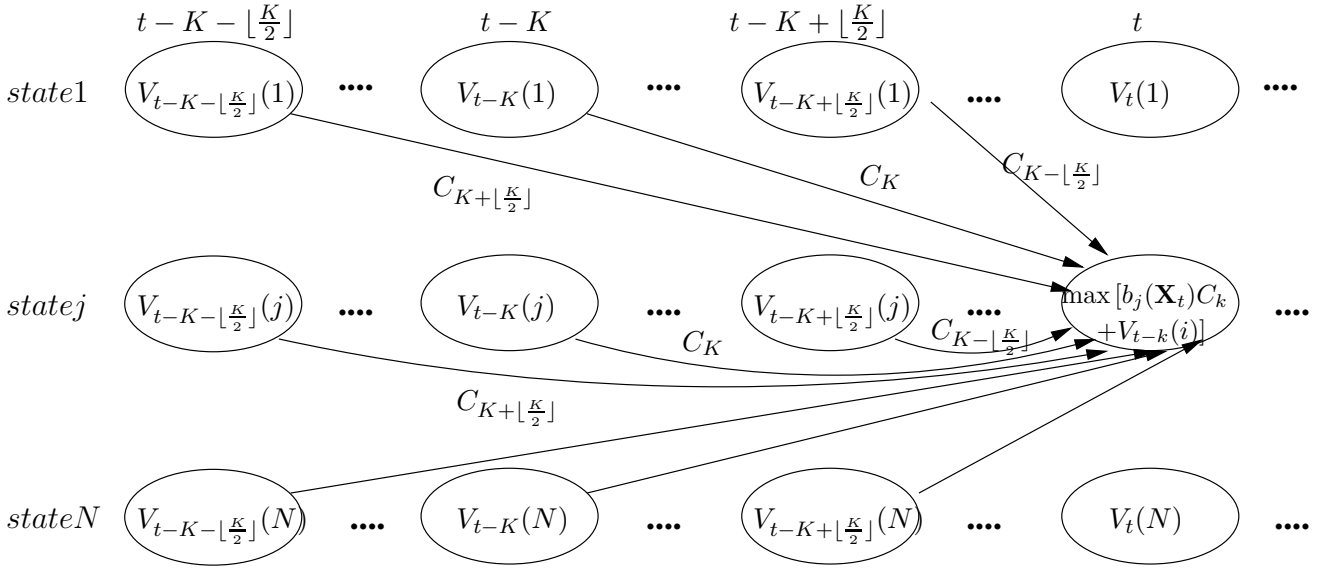

Fig. 1. The Viterbi decoder embedded with frame selection

C_k: the time compensation coefficient for a candidate frame. In this approach, a linear penalty is employed: $C_k = \frac{k}{K}$.
$S_t(j)$: the most likely previous state
$F_t(j)$: the most likely previous selected frame
K: the average number of frames out of which one frame is selected
In the initialization, for $1 \le j \le N$ and $0 \le t < K$,

$$V_t(j) = \pi_j + C_t b_j(\mathbf{X}_t); \tag{1}$$

$$S_t(j) = j; \quad F_t(j) = t. \tag{2}$$

Starting for the initialization, we compute $V_t(j)$, $S_t(j)$ and $F_t(j)$ in a time-synchronous approach:

$$V_t(j) = \max[b_j(\mathbf{X}_t)C_k + V_{t-k}(i)] \quad \text{if } K \le t < T-K, \tag{3}$$

where the maximization is done over $K - \lfloor\frac{K}{2}\rfloor \le k \le K + \lfloor\frac{K}{2}\rfloor$ and $1 \le i \le N$, though limited by the topology of the HMM. $S_t(j)$ and $F_t(j)$ are the tokens of $V_t(j)$:

$$S_t(j) = \underset{1\le i\le N}{\arg\max}[b_j(\mathbf{X}_t)C_k + V_{t-k}(i)], \tag{4}$$

subject to $K - \lfloor\frac{K}{2}\rfloor \le k \le K + \lfloor\frac{K}{2}\rfloor$, and

$$F_t(j) = \underset{K-\lfloor\frac{K}{2}\rfloor\le k\le K+\lfloor\frac{K}{2}\rfloor}{\arg\max}[b_j(\mathbf{X}_t)C_k + V_{t-k}(i)], \tag{5}$$

subject to $1 \le i \le N$.

$V_t(j)$ at the ending frames must also be penalized w.r.t. their time lags to the final frame:

$$\begin{aligned} V_t(j) = \max[&b_j(\mathbf{X}_t)C_k + V_{t-k}(i)] \\ &+ C_{T-1-t}b_j(\mathbf{X}_t) \quad \text{if } T-K \le t \le T-1. \end{aligned} \tag{6}$$

The computations for $S_t(j)$ and $F_t(j)$ at the ending frames are the same as Eq. 4 and Eq. 5.

The best probability for the whole sequence is chosen from the last possible frames:

$$\text{The best score} = \max_{\substack{T-K\le t\le T-1\\ 1\le j\le N}} V_t(j). \tag{7}$$

Then, the sequences of most likely states S^* and most likely selected frames F^* are retrieved in turn:

$$S^*(p) = \begin{cases} \underset{1\le j\le N}{\arg\max}[V_t(j)], & T-K \le t \le T-1 \quad \text{if } p=1 \\ S_{F^*_{p-1}}(S^*_{p-1}) & \text{if } p>1 \end{cases} \tag{8}$$

$$F^*(p) = \begin{cases} \underset{T-K\le t\le T-1}{\arg\max} [V_t(j)], & 1 \le j \le N \quad \text{if } p=1 \\ F_{F^*_{p-1}}(S^*_{p-1}) & \text{if } p>1 \end{cases} \tag{9}$$

The *multiple selection* approach is illustrated in a trellis framework, shown in Figure 1.

2.2. Single selection

The *single selection* approach is an extreme frame selection method, which selects only one frame for one state, regardless of the duration of a testing segment (e.g. a phoneme existing out of 3 states). From here on we suppose we know (or have hypothesized) the segment length.

Furthermore we use following statistics obtained from a training set: s_j is the mean state duration and m_j is the expected position of the selected frame. During training the position of the most likely selected frame $L(j)$ in state j is estimated as:

$$L(j) = \frac{\underset{Ts_j\le t<Ts_{j+1}}{\arg\max} \; b_j(\mathbf{X}_t)}{T}. \tag{10}$$

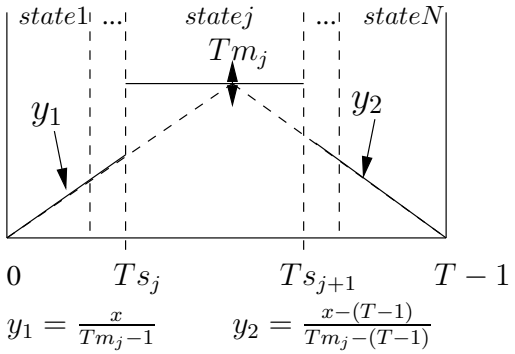

Fig. 2. The linear penalty for the *single selection*

Thus, $m_j = E(L(j))$. For a testing segment, the likelihood probability $b_j(\mathbf{X}_t)$ is linearly filtered by a time penalty proportional to the distance from the expected position for frames outside the expected state duration, as shown in Figure 2 and Eq. 11.

$$\tilde{b}_j(\mathbf{X}_t) = \begin{cases} y_1(t) b_j(\mathbf{X}_t) & t < Ts_j \\ b_j(\mathbf{X}_t) & Ts_j \le t \le Ts_{j+1} \\ y_2(t) b_j(\mathbf{X}_t) & t > Ts_{j+1} \end{cases} \quad (11)$$

The frame whose $\tilde{b}_j(\mathbf{X}_t)$ is maximal is selected as a representation for state j. In case that the time order of the selected frames is inconsistent with HMM states, they will be replaced by the frames at default positions, Tm_j.

2.3. Likelihood probability normalization

The recognition of a sequence of observations $\mathbf{X} = [\mathbf{X}_t]$, $0 \le t < T$, is often estimated by the Bayesian equation:

$$\log(P(h|\mathbf{X})) = \log(P(\mathbf{X}|h)) + \log(P(h)) - \log(P(\mathbf{X})) \quad (12)$$

where $h \in H$, and H is a set of all possible hypothesis. In an isolated word task, the prior probability $P(h)$ can be omitted if every hypothesis has the equal priori probability. The denominator $P(\mathbf{X})$ indicates the probability of the observation sequence. In a fixed frame rate system $P(\mathbf{X})$ can also be ignored without influencing the recognition process due to the maximization operation. However in the frame selection, it plays an important role: all possible hypothesis generate different selected observation sets $\mathbf{X}(h)$, resulting in the unequal estimations of $P(\mathbf{X}(h))$. In our implementation, a sink model constructed by HMM models for all hypothesis is used to estimate the observation probability:

$$P(\mathbf{X}_t) = \sum_{H} \sum_{\forall G} w_G P(\mathbf{X}_t|G) \quad (13)$$

where G is a Gaussian distribution and w_G is its weight.

3. EXPERIMENTS

An isolated phoneme recognition experiment is performed on the TIMIT database, where the boundary of each phoneme is taken from the manual labels. The database contains a total of 6300 sentences, 10 spoken by each of 630 speakers. 73% of the speakers are put into a training set and the rest composes the test set. A phonetic alphabet of 46 symbols is used.

With the 41327 phoneme segments in the test set, the 95% confidence interval is around $\pm 0.47\%$.

3.1. Pre-processing

Initially, speech is decomposed into 30msec-length frames, with 2msec frame shift. For the *multiple selection*, the intended average frame shift after the frame selection is 10msec. According to the analysis in section 2.1, one frame will be selected out of five consecutive frames on average.

Standard 12th-order cepstral coefficients, plus energy, their first and second order derivatives are extracted for each frame. The dynamic features are computed with absolute time differences identical to the ones used in our reference 10msec fixed frame rate analysis.

3.2. Analysis of acoustic models with the frame selection

With the frame selection, the observations of selected frames are more concentrated towards their centers, which can be seen in Figure 3, where solid ellipses indicate the dispersion of observations in the training set after the *multiple selection* and dashed ellipses are for the dispersion of observations with the 10msec fixed frame rate. We can see that the mean parameters of phonemes are not shifted, but the variances are shrunk a little, thus shrinking the "within class" variability, while maintaining the "between class" variability. This property should improve phoneme recognition rates.

Given the stability of the centroids and weak influence of the different variances, we decide to use models trained from 10msec fixed frame rate training data for all experiments.

3.3. Multiple selection

Phonemes are modeled by a three-state left-to-right context-independent HMM, with 16 Gaussians per state. Table 1 shows better recognition rates for the testing segments with the *multiple selection* than with the fixed frame rate. As the same acoustic models are used, we must conclude that many useless, or even harmful frames that occurred in the fixed frame rate, have been replaced by better frames in the neighborhoods during the decoding process.

10msec fixed frame shift	average 10msec after frame selection
61.5%	65.1%

Table 1. Phoneme recognition rates for the system with the fixed frame rate and the system with the *multiple selection*

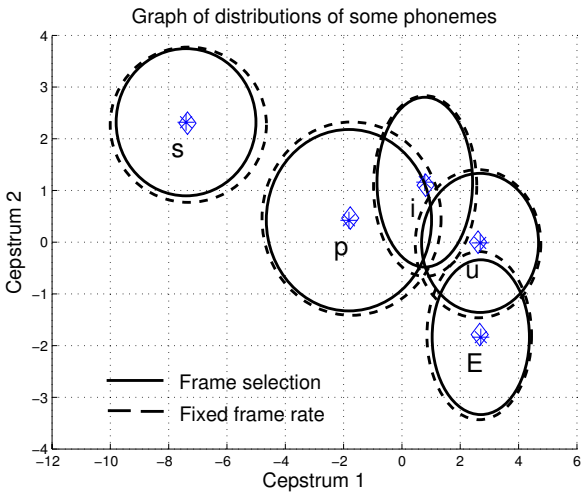

Fig. 3. The distribution of the first and second cepstral coefficients of some phonemes. The star markers and solid ellipses indicate the means and variances of the distribution after the frame selection, while the diamond markers and dashed ellipses are of the fixed frame rate.

3.4. Single selection

Given that the number of states of the phoneme HMMs is three, the *single selection* selects three frames from a testing phoneme to represent its acoustic characteristics with one frame fitting one state. We perform the *single selection* on the test set with 10msec fixed frame shift and the set with 2msec tiny frame shift respectively. Those three-frame segments are decoded by the same phoneme models we used in section 3.3, and the recognition rates are shown in Table 2. From this table, we can see that the *single selection* is capable of selecting more suitable frames from the pool of frames with a finer time resolution, and hence achieves superior performance.

10msec fixed frame shift	2msec tiny frame shift
63.1%	65.6%

Table 2. Phoneme recognition rates when the *single selection* is used

3.5. Discussion

Interestingly, comparing section 3.3 with section 3.4, we find that a segment containing much more frames can be even less discriminative than a segment only represented by three distinguished frames. The reason for this observation can be attributed to the prior knowledge of the phoneme length T which is necessary for the *single selection*, but it may also imply the importance of frame selection: it is quite possible that frames selected by the *multiple selection* are still redundant and useless, or even weakening because of the constraint of the intended average frame rate; some of them can be discarded further. Despite this observation, extending the *multiple selection* to a continuous speech recognition is much easier than the *single selection* as its implementation is very similar to classical time-synchronous Viterbi, while the *single selection* needs the hypothesis of phoneme boundaries.

4. CONCLUSIONS AND FUTURE WORK

Two ML-based frame selection approaches have been presented. From experiments on the TIMIT database, we can conclude that the approaches shrink the variances of observations and thus show promise for improving recognition of speech. The error rate decreases nearly 10% relative for a phoneme recognition task.

The frame selection is a promising technique. Our future research includes three directions. First, we expect the frame selection is robust against noise, since it focuses on frames which are close to the acoustic models, explicitly ignoring contaminated frames. Second, we will investigate the performance when the constraint of the average frame rate is thrown away; in this case the frame selection becomes completely data-driven. At last, we will extend the isolated phoneme recognition to a continuous speech framework.

5. REFERENCES

[1] James J. Hant and Abeer Alwan, "A psychoacoustic-masking model to predict the perception of speech-like stimuli in noise," *Speech Comm.*, vol. 40, no. 3, pp. 291–313, 2003.

[2] R.D. Zilca, J. Navratil, and G.N. Ramaswamy, "Depitch and the role of fundamental frequency in speaker recognition," in *Proc. ICASSP*, Hongkong, Apr. 2003.

[3] R.D. Zilca, J. Navratil, and G.N. Ramaswamy, "SYNCPITCH: A pseudo pitch synchronous algorithm for speaker recognition," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sept. 2003.

[4] H. You, Q. Zhu, and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," in *Proc. ICASSP*, Montreal, Canada, May 2004.

[5] Umit H. Yapanel, S. Dharanipragada, and John H.L. Hansen, "Perceptual MVDR-based cepstral coefficients (PMCCs) for high accuracy speech recognition," in *Proc. EUROSPEECH*, Geneva, Switzerland, Sept. 2003.

[6] T.Y. Wu, D. Van Compernolle, J. Duchateau, Q. Yang, and J-P. Martens, "Spectral change representation and feature selection for accent identification tasks," in *Proceedings of the Workshop on Modelling for the Identification of Languages*, Paris, France, Nov. 2004, pp. 57–61.