**Frame selection:**

Most speech recognition systems use a fixed frame rate, typically 100Hz, to decompose speech into a series of frames mainly for its convenience. In such a scheme all speech frames are assigned the same importance from a pattern classification viewpoint. But some limitations appear with this arbitrary fixed frame rate. For instance, a noisy frame may dominate the recognition process; the same importance assigned to each extracted frame is inconsistent with human perception (Hant and Alwan, 2001); pitch asynchronous representation (Zilca et al., 2003a,b) caused by the fixed frame rate leads to pitch mismatch due to the presence of pitch-related harmonics in the power spectrum; the well-known implicit weakness of HMM in duration modeling aggravates this concern (see section 1.3.1). Furthermore, in the case of continuous speech recognition, observations at the beginning and the end of a phoneme segment are highly influenced by contextual information. Hence, the distributions of these observations that are dominated by co-articulation are broad and their likelihoods might not be informative. Indeed, Louradour (Louradour et al., 2005) discovered that the frames at boundaries may carry more speaker-related information than speech-related information; hence using frames for speech recognition at the boundaries may involve a risk. Moreover, frames at the boundaries usually carry information from both sides, which damages the HMM assumption that all feature vectors on the same state are identically distributed. On the contrary, observations in the steady state 1 , although more reliable, tend to be similar and redundant in the decision process. While the recognition decision is made by comparing the sums of the likelihoods of all hypotheses, keeping the most discriminative frames and throwing away redundant ones would not greatly affect the maximization operation. Because of those limitations, researchers are looking for a better frame representation for an utterance and a way to reconstruct the spectra more efficiently. These efforts include variable frame rate (You et al., 2004), segment normalization (Nedel and Stern, 2001, 2003), speaking rate estimation (Nanjo and Kawahara, 2004), etc.

Among these efforts, one of the significant studies was done by Hillenbrand and his colleagues (Hillenbrand et al., 1995). The purpose of the study was to replicate and extend the classic study of vowel acoustics by Peterson and Barney (Peterson and Barney, 1952), including formant contours for F1-F4 and vowel durations in context- fixed (in a /hVd/ structure) vowel discrimination. The study showed that the confusions among adjacent vowel categories in the F1-F2 space are much greater than Peterson-Barney’s study suggested, and many of the vowels are in different locations. The authors suggested that “...the vowels of American English are more appropriately viewed not as points in phonetic space but rather as trajectories through phonetic space” (Hillenbrand et al., 1995). In the experiments, they showed that using F1-F3 as features and a quadratic discrimination analysis technique (Johson and Winchern, 1982), the 10-vowel classification accuracy (after omitting two center diphthongs /e/ and /o/) was 81.0% for only using one frame at steady state, 91.6% for two frames taken at 20% and 80% of vowel duration and 91.8% for three frames taken at 20%, 50% and 80%. The authors claimed that “this would seem to suggest that only a very coarse representation of the spectral change pattern is needed for classification” (Hillenbrand et al., 1995). We observe that this simple frame selection method could discard those context-influenced frames at the phoneme boundaries and simultaneously reduce the redundancy possibly occuring in the middle part. On the other hand, this method could be a way to overcome the shortcomings of the fixed frame rate. However, this study was limited to a context-fixed experiment and has not been extensively investigated in more complex phonetic environments than the /hVd/ utterances examined. Furthermore, the concept is not directly applicable to continuous speech recognition as it needs to know phoneme boundaries.

The frame selection technique proposed in this chapter is an extension of the simple frame selection done by Hillenbrand, and an alternative solution for overcoming the problems resulting from the fixed frame rate. Frame selection is a technique to select a subset of speech frames as a representative for the whole speech signal to distinguish the characteristics of speech units. The original set of speech frames is usually generated using a high fixed frame rate, on which a frame selection technique is employed to pick reliable and informative frames. In this chapter, we develop a series of simple-to-complex schemes to demonstrate the advantage of frame selection and to investigate the scope within which frame selection is applicable. This chapter is organized as follows. In section 2.2 we give the Bayesian explanation for frame selection in the context of speech recognition. In section 2.3 and 2.4 the idea of frame selection is presented. Two methods of frame selection, fixed frame selection and variable frame selection are then developed. A significant difference between these two selection methods is that with fixed frame selection, the number of selected frames is pre-defined for each phoneme, independent of the length of the phoneme segments, which is consistent with Hillenbrand’s experiments; while with variable frame selection, neither the length of phoneme segments nor the number of selected frames is known before selection, thus it could be incorporated into a classical speech decoding procedure. We first reproduce a part of Hillenbrand’s experiments and investigate the motivation from another point of view, then we examine the validity of frame selection on a phoneme classification task using the TIMIT phonetic database. Afterwards we extend the framework to phoneme recognition. We analyze the difficulties and feasibility that the frame selection meets in different tasks and propose some possible ways to further implement it. The summary is given in the last section.

**2.1.1 Bayesian explanation for frame selection in speech recognition:**

Given acoustic data and word acoustic HMMs Φw. , the task of speech recognition is to look for a word or state sequence which maximizes the posterior probability. According to the Bayesian formula, this turns to look for the maximum of .

The application of frame selection is to select a subset of frames from the full set to replace the full frame set. I approximate the likelihood as expectation of . The item indicates the likelihood of selection subset given . Further, instead of taking the expectation for all possible . I simplify the expectation operation to the maximum using maximum function.

The first item on the right side represents the log-likelihood of to the acoustic model of W; it can be estimated by a standard likelihood calculation. The second item is the language model related factor; in a phoneme classification task, for example, this item is equal to all possible sequences. The third item stands for the probability that is generated. The last item represents the probability that the subset is selected from the full set . This item depends on how the subset is generated. The following image illustrate the Bayesian surface interference of sound signal for frame length of 0, 20, 40,60.

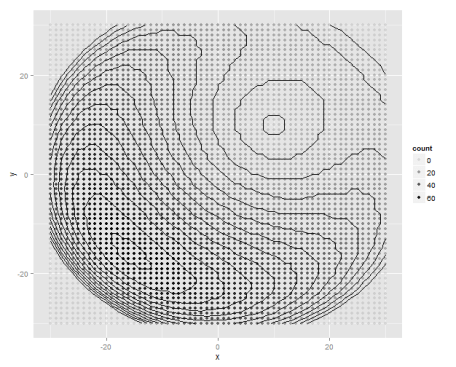


Figure 2.1.1. Bayesian model for multiple frame

**2.1.2 Equal prior distribution:**

In some applications, such as phoneme classification, the prior distribution of each phoneme is presumed equal. Thus the second item log can be ignored.

**2.1.3 Fixed number of selected frame:**

For some frame selection methods, the number of selected frames is determined a priori, independent of the length of the phoneme segment. In this case is equal for all possible. For instance, if we decide to select 1 frame from a phoneme segment with the length T, then for all. Therefore, Equation 2.6 is shortened as:

2.1.4 Fixed Selected Frames

If the selected frames are fixed before recognition, the selected frames are independent of a word sequence, then the items and are identical for all possible hypotheses. As a result, the recognized word sequence is searched by:

**2.1.5 Selected frames dependent on hypothesis:**

In some cases, the selected frame set is dependent on a word sentence hypothesis. In this situation, has to be denoted as W . To abbreviate our notations, we use W to denote W as the frame set selected by hypothesis. Then the recognized word sequence is found as follows:

The third item on the right side is the log probability that is generated. Note that unlike Equation 1.6, cannot be omitted as it is not equal for all hypotheses. In our implementation, a sink model constructed by HMM models for all hypotheses is used to estimate the observation probability of W , as an HMM state is modelled by weighted Gaussian mixtures:

**2.2 Frame Selection in phoneme classification**

Fixed frame selection is a class of methods in which a small and fixed number of frames is selected per phoneme. The number of selected frames is independent of the number of frames a phoneme segment holds. Unlike the method used by Hillenbrand (Hillenbrand et al., 1995), which took frames only at 20%, 50% or 80%, or their combination, the proposed fixed frame selection is to compress the number of selected frames to an extreme case and to investigate at which position(s) frames are the most representative.

In the first study, I attempt to select the most discriminative frame regardless of the duration of a test segment; then we increase the number of desired frames to two, and see how the performance of phoneme classification is enhanced and what their corresponding positions are. Finally, we select one frame to represent each state in a multi-state HMM. In this case the number of selected frames for a test phoneme segment depends on the topology of the HMM. By incrementally increasing the number of selected frames, we are capable of studying factors that influence the accuracy of frame selection and investigate the characteristics of the selected frames. In the selection procedure, two frame properties will be used. One is the likelihood of the frame against all possible classes; the other is the position the frame holds in the phoneme segment. We will show that although both properties are important for accurately spotting the representative frames, the positions of the selected frames are more crucial; but combining these two properties is better than using either property alone

**2.2.1 Position Based fixed frame selection:**

A first family of approaches for selecting frames, called Position based fixed frame selection, is to simply pick frames at fixed relative positions of the duration in a phoneme segment without taking their likelihoods into account (Wu et al., 2004, 2006b). For example, we could arbitrarily select a frame at 50% of the duration as the representative one if the desired number of selected frames is one. We may also pick frames at 20%, 50% and 80% of the duration for a three-state HMM, each frame representing one state. Position based fixed frame selection is an extension of the method used by Hillenbrand. However, in Hillenbrand’s study (Hillenbrand et al., 1995), the context was fixed as all vowels were embedded into a /hVd/ structure; the fixed relative positions he used (20%, 50% and 80%) are probably suitable neither for all phonemes, nor in continuous circumstances. Therefore we need to look for more discriminative positions in our case. We will look for frames positioned at percentages of the segment duration. In practice each segment is normalized to 11 frames, yielding normalized positions 0% (first frame), 10%, 20% ... 100% (last frame). For segments with more than 11 frame originally, this normalization is a down-sampling process; for segments shorter than 11 frames, some frames are duplicated. Practically, because the frame shift we use for frame selection is 2msec, only 3.95% of phoneme segments are shorter than 22msec, thus need to duplicate some frames. This normalization is very similar to the approach adopted in (Nedel and Stern, 2001), where the authors proposed segment normalization to replace the fixed frame rate using down-sampling for longer segments but using the Missing Data Technique (MDT) to reconstruct a phoneme segment shorter than a desired duration. The reason that we use the duplication strategy instead of the MDT is that in our study we are more concerned about the speech event happening at a certain position, while in (Nedel and Stern, 2001) the authors focused on accurate speech recovery. Assume the fixed relative positions E are pre-defined, is calculated as where indicate the is rounded to the nearest integer.

Note that if no knowledge of prior distribution is known, this approach meets the requirement of Equation 2.7 and 2.9: the number of selected frames is fixed and the selected frames are independent of the phoneme hypotheses, thus, a speech segment is classified to a phoneme \* which obtains the maximum likelihood:

where c is a possible phoneme class

**FPOS 1: One frame per phoneme**

In FPOS 1, we look for only one frame at a fixed position to represent the characteristics of the phoneme segment. It is commonly believed that a frame in the steady state is the least influenced by its context, thus is more representative than frames close to boundaries. If the frame at is picked, the log likelihood of the whole segment for a hypothesis class is then defined as, where the maximum operation is used to look for the state which gives the largest likelihood among all possible states for a phoneme class. If we denote the likelihood of the segment to a specific class as , then a phoneme segment is identified to the class whose likelihood is maximal among all possible classes:

**FPOS 2: Two frames per phoneme**

With FPOS 2, two frames are selected at two pre-defined positions. Selecting two frames does not mean to simply add one candidate frame to the frame selected by the FPOS 1 method: although the frame selected from FPOS 1 is the most discriminative one, it does not guarantee that the best performance could be reached by incorporating another distinguished but probably correlated frame. FPOS 2 is a way to investigate at which positions frames could be combined best.

**FPOS N: One frame per HMM state**

A well-known assumption of the HMM is that feature vectors in the same state follow the same distribution. Therefore in frame selection, it is natural to investigate that to what extent the combination of one frame picked from one state could represent the whole segment, as theoretically in the traditional HMM the likelihood of the whole test segment is merely the linear sum of the likelihoods of observations in different states (if the much less influential factor – state transition probability is not considered). Before the investigation, we need to know the best relative position of a state from the training set. To do this, all combinations of three fixed relative positions are tested in the training set and the frame combination which obtains the best performance is retrieved as the best positions. Assuming that the best positions , for all states are found in the training, the score for a phoneme model is nothing more than the sum of the likelihoods of the frames at t :

**2.3.3. ML based fixed frame selection**

Another approach to select frames is based on the likelihoods of frames. Likelihood describes how similar a test sample to a model is, or how close in the feature space a test sample to a model is. The motivation for using likelihood scores as a criterion to select frames is coming from the decision process of an ML framework in speech recognition: a decision is made by comparing the sum of likelihoods of all frames in a test segment to all potential acoustic class models; the one which is the most likely to generate the observation is identified as the test class. The distances, or likelihoods of frames in one speech segment to a hypothesis can be different: some of them may be closer than others; we conjecture that the frames which are the closest to the hypothesis could be a good indicator for the segment itself. In this approach, we have already defined the desired number of selected frames, but the selected frames are dependent on the hypotheses.

**2.3.4 Combination based fixed frame Selection**

There are limitations contained in the position based frame selection and the ML based frame selection implicitly. For example, the ML based frame selection does not enforce the selected positions to span over the full segment. When a small part of a phoneme segment is very close to an incorrect phoneme model, all selected frames could originate from the small part, resulting in a mis-classification. For the position based frame selection, frames at some fixed positions can be noisy or less informative by chance for unknown reasons, and thus result in low likelihoods for the correct phoneme class. We find that in fact the advantages and weaknesses of these two methods are complementary, hence it is possible to combine them: the main idea of the combination is that frames lying at the appropriate positions and at the same time having high likelihoods would have higher opportunity to be selected than other frames. In other words, two factors are evaluated: selecting too closely spaced frames and selecting frames with low likelihoods would be penalized and thus prohibited. Therefore we propose two methods to combine two properties, namely likelihood prior combination and position prior combination. In the combination based fixed frame selection, we only investigate the situation of one frame for one HMM state. In this case, the decision rule degenerates to:

The last item represents to what extent the selected frames are representative for the the full set

Figure 2- The linear penalty for the Combination based method