Single Frame Selection for Phoneme Classification

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

Our former study [1] has shown that maximum likelihood (ML) based frame selection, which selects reliable frames from a high resolution along the time axis, helps to improve the discrimination between phonemes. In this paper, we present our recent research on single frame selection for a phoneme classification task. A new *single selection*, which only selects one frame for one state in an Hidden Markov Model (HMM), is proposed. The new technique takes likelihoods of frames and their positions in a phoneme segment into account at the same time, and selects very few frames to represent the spectral evolution of the phoneme. Furthermore, we also show that for a low model complexity, a phoneme model trained by selected frames is more discriminative than a model using all frames.

Index Terms: phoneme classification, frame selection.

1. Introduction

While it is commonly agreed that a fixed frame rate, typically 10ms, is not consistent with human perception [2], the fixed frame rate is still used in most of state-of-the-art speech recognizers because of its simplicity and convenience. Thereby, speech frames are often assigned the same importance in pattern classification. However, in the case of continuous speech recognition, observations at the beginning and end of a phoneme are highly influenced by contextual information. Hence, the distributions of these observations that are dominated by co-articulation are broad and the their likelihoods might not be informative. In fact, as claimed in [3], the frames at boundaries may carry more speaker-related information, and are often vague for the speech recognition task. At the same time observations in the steady zone if any - although most likely more reliable - tend to be similar and add redundant information in the decision process.

Some researchers are looking for a substitution for the fixed frame rate. These efforts include variable frame rate (VFR) [4, 5], duration normalization [6], etc. In our former study [1], we proposed a ML-based frame selection technique, which selects reliable frames from a tiny frame shift, to classify phonemes by assuming the boundaries of phonemes are known priorly. The frame selection technique was realized by two methods: *multiple selection* and *single selection*. While the *multiple selection* can be incorporated into a standard Viterbi decoder procedure after which the average frame rate is equal to a pre-defined value, the *single selection* selects only one frame to represent one state in an HMM regardless of the duration of a testing segment. In this paper, we further study factors that influence the accuracy of the *single selection* and investigate the characteristics of the selected frames.

Not every frame contributes equally in a decision process. The *single selection* attempts to select the most distinguish one for the corresponding HMM state. At least two factors affect the validity of selecting the most representative frame to better model the spectral evolution of a phoneme. One is the likelihood of a selected frame, the other is its position in a phoneme segment. In [1] we have shown that the ML-based dynamic frame selection works well for phoneme classification. We will see in this paper that the positions of selected frames are also crucial for the *single selection*. We also find an approach for combining these two factors. Our experimental results on the TIMIT database show that using few frames selected by the combination method is comparable with the traditional fixed frame rate scheme. Moreover, we observe that if a phoneme model is trained by the selected frames instead of using all frames, the effect is dual: in a low model complexity, it shows more discriminative power, while in a high model complexity, due to limited training data, its performance is deteriorated.

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. Single frame selection

The main purpose of the *single frame selection* is to find one typical and representative frame for each HMM state, and the characteristic of a phoneme is only depicted by the selected frames, without taking other discarded frames into account. Some methods are proposed to select the appropriate frames.

2.1. *SEL1*: ML-based single selection

The *single selection* in our former work [1] is based on an ML criterion, which first statistically estimates the expected positions of the selected frames and the mean state durations, then linearly filters the likelihoods of frames which are too far away from the expected positions. In this paper we will first investigate a *single selection* approach, called *SEL1*, which only considers the likelihoods of frames, without the existence of the linear filter.

The criterion can be formulated as follows. Denote the indices of the selected frame for state j ($1 \leq j \leq N$) as t_j ($1 \leq t_{j-1} \leq t_j \leq T$), where N is the number of states for a phoneme HMM, and T is the number of frames in a phoneme segment, the best score S for a certain phoneme model is

$$S = \max_{\forall t_j} \sum_{i=1}^{N} b_i(\mathbf{X}_{t_j}), \tag{1}$$

where $b_i(\mathbf{X}_t)$ is the log likelihood of emitting feature vector $\mathbf{X}_t$ generated by state i. The indices of the selected frames can be retrieved as

$$t_j = \text{argmax}(S) \tag{2}$$

2.2. *SEL2*: fixed percentages based single selection

Another approach to select frames, called *SEL2*, is to simply pick frames at fixed percentages of a phoneme segment without taking their likelihoods into account [7]. For example, we could arbitrarily select frames at 20%, 50% and 80% of every phoneme segment for a three-state HMM, each frame representing one state. Many different combinations of fixed percentages are defined priorly and then tested on the training set. Among those combinations, the best one is grafted to the evaluation set. Note that in this case, the selected frame sets are identical for different phoneme classes.
Assume the fixed percentages c_j are pre-defined, t_j is calculated as $t_j = \text{int}(c_j \times T)$. The score for a phoneme model is nothing more than the sum of the likelihoods of the frames at t_j:

$$S = \sum_{i=1}^{N} b_i(\mathbf{X}_{t_i}) \tag{3}$$

2.3. *SEL3*: combination based single selection

There are limitations contained in the *SEL1* and the *SEL2* implicitly. For example, the *SEL1* does not consider the positions where the frames are selected; when a small part of a phoneme segment is very close to an incorrect phoneme model, with the *SEL1*, all selected frames for that phoneme model probably originate from the small part, resulting in a mis-classification. For the *SEL2*, frames at fixed positions can be noisy or less informative by chance.
In fact the advantages and weaknesses of the *SEL1* and *SEL2* are complementary. *SEL3* is an attempt to compromise the *SEL1* and *SEL2*. The main idea of the *SEL3* is that frames lying at the appropriate positions and meanwhile having high likelihood probabilities would have more opportunity to be selected than other frames. In other words, two factors are considered at the same time: selecting too close frames and selecting frames with low likelihood will be penalized and then prohibited. In our experiment, the appropriate positions are decided empirically by the optimal fixed percentages obtained from the *SEL2*.
Suppose the series of the optimal fixed percentages in the training set is c_j, $1 \leq j \leq N$. The best distances between encouraged selected frames then are $d_j = c_j - c_{j-1}$ ($c_0 = 0$). To emphasize the importance of the span between the selected frames, we impose a filter on frames, which is defined as:

$$F(t_1, t_2, ..., t_N) = \prod_{j=1}^{N} \frac{k}{(\frac{t_j - t_{j-1}}{T} - d_j)^2 + 1} \tag{4}$$

where the parameter k is a scale factor, $t_0 = 0$, and $t_1, t_2, ..., t_N$ indicate any series of possible selected frames. Then, the sum of their likelihoods is filtered as:

$$S(t_1, t_2, ..., t_N) = F(t_1, t_2, ..., t_N) \times \sum_{i=1}^{N} b_i(\mathbf{X}_{t_i}) \tag{5}$$

By this way, frames close to the best positions will be assigned a higher weigtht, while selecting frames too far away, or too close will be penalized. The final sum of probabilities for the testing segment is the maximum of $S(t_1, t_2, ..., t_N)$ thoughout all possible t_j combinations.

3. Experiments

3.1. Evaluation corpus

The standard TIMIT acoustic/phonetic database is used to evaluate the performance of the *single selection* approaches. All "sa" sentences are excluded from the training and recognition because they skew the phoneme occurrences, leaving 3696 sentences for training and 1344 for test. We adopt the same phoneme set and the same evaluation methods as described in [8]. Thus, there are 40516 testing phoneme segments not counting silence segments. The upper bound of the 95% confidence interval is $\pm 0.49\%$.

3.2. Speech processing

A 39-dimensional feature vector, consisting of 12th-order mel cepstra plus energy, their velocity and acceleration coefficients, is extracted for each 30ms speech frame. The frame shifts for the training set and the testing set are 10ms and 2ms respectively. The standard topology of the HMM is used: a 3-state left-to-right context-independent HMM model with 16 diagonal-covariance gaussian mixtures per state is trained for each phoneme by the ESAT continuous speech recognizer. Due to the idea that one selected frame represents one state, with this topology, three frames are required to be selected from each phoneme segment. The phoneme boundaries in the testing set are achieved by a forced alignment procedure using the accompanied transcripts and the phoneme HMMs.

3.3. Baseline: the conventional Viterbi decoder

We start by the standard phoneme classification approach. The accuracy of the common phoneme classification can be experimented with the phoneme HMMs, the Viterbi decoder and the testing feature vectors after down-sampling to the 10ms frame rate. Without any frame selection, the recognition rate is 73.98%. Comparing to the result we presented in [1], we conclude that the forced alignment gives more accurate phoneme boundaries than manual ones, although two phoneme sets involved are also slightly different.

3.4. *SEL1*: ML-based single selection

We first focus on the simple ML-based single selection, *SEL1*. As we described in section 2.1, given the boundaries of a testing segment, for a possible phoneme class, three frames whose likelihoods are maximal against corresponding state models are responsible for the final decision. In our experiment, using the *SEL1* we obtain 63.30% accuracy for the phoneme classification task.
This worsened result is understandable. A certain part of a testing segment can be more likely to an incorrect phoneme than its ground truth. Since all competing HMMs have the freedom to select the most likely frames, the incorrect model tends to select frames all from that part, and thus the sum of their likelihoods is more competitive.

3.5. *SEL2*: fixed positions based single selection

The optimal positions for the *SEL2* can be learned from the training set through comparing different combinations of possible positions. The optimal series of positions will be further used in the *SEL3*.

3.5.1. Learning optimal fixed positions

Thoroughly searching for an optimal combination of fixed positions in the training set is really computationally heavy and can

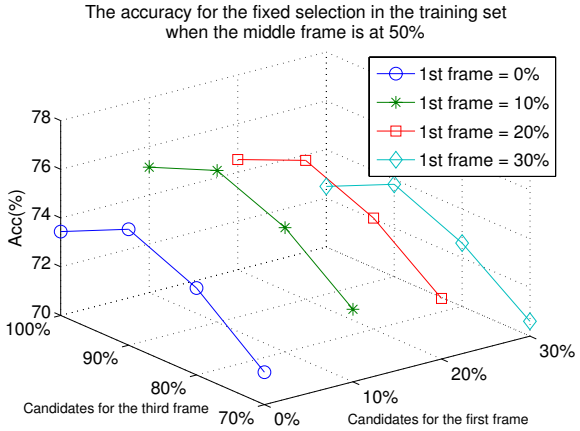

Figure 1: *The accuracies of phoneme classification in the training set when frames are selected at the fixed positions. The middle frame is extracted from 50%.*

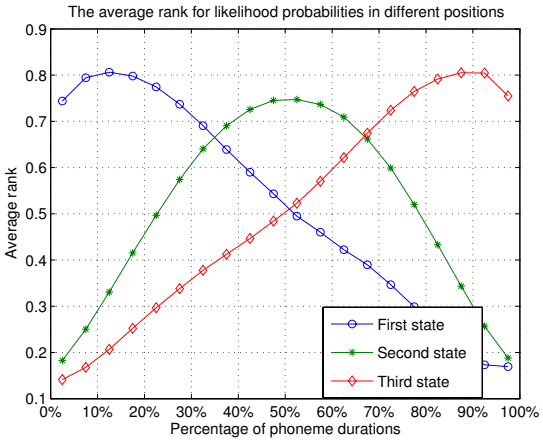

Figure 2: *The average rank for likelihood probabilities with different percentages of phoneme segments*

leard to over-training. Alternatively, we restrict the possibly selected frames to some candidate percentages to search for a quasi-optimal solution. In the experiment, the first frame for a phoneme segment can only be selected from a few pre-defined percentages, say, 0%, 10%, 20% and 30%; the candidate positions for the second and third selected frame are also restrained from 10% to 90% and from 70% to 100% with 10% as a step respectively. 0% indicates the beginning frame of a phoneme segment, while 100% is the end. Concerning the rule of time order, there are 96 possible combinations. To visually plot the results, Fig. 1 shows curves of the phoneme classification accuracies with different combinations of the arbitrary selections for the first and the third frames when the middle frame is fixedly selected from 50%. We do not present the recognition rates with other positions of the middle frame since their trends are similar to the curves in Fig. 1 and 50% for the middle frame is the best choice among its competitors. As can be seen, the best combination of the first and third selected frames does not appear either at phoneme boundaries, or at positions close to phoneme centers; the percentages for the most discriminative frames are 10% and 90%. Although this selection is not optimal over the whole phoneme segments, we still may conclude that frames at boundaries or close to the steady part are not the most representative ones. This experiment confirms the necessarity of a moderate span between selected frames. Some *SEL2* results on the testing set are shown in Tab. 1.

3.5.2. *Learning positions for maximum likelihood probability*

Additionally we also can learn the most possible range where maximum likelihood appears from the training set. The likelihoods of all frames in a phoneme segment are sorted in ascending order and the ranks are normalized by the length of the segment. The average ranks at different percentages for different state models are shown in Fig. 2, which illustrates the highest likelihood for the first state normally shows up at around 10%, around 50% for the middle state and about 90% for the last state.
The position where the maximal likelihood appears is quite consistent with the quasi-optimal fixed position (section 3.5.1). This indicates that the best combination of selected frames may come from somewhere around 10%, 50% and 90%. Frames close to them hold a moderate span and probably reasonable likelihoods, which satisfy our demand for the *SEL3*.

3.6. *SEL3*: ML-based single selection with a span factor

As we have shown in section 3.4, automatically selecting the most discriminative frames is not only relying on their likelihood probabilities, but also depending on their relative distances and their positions in a phoneme segment. Benefiting from the results in section 3.5, we learn that the best selected frames probably stay at around 10%, 50% and 90% of a phoneme segment. The exact positions should be decided by likelihood of their neighboring frames. Thus, we set the optimal fixed positions c_j to $c_j = (10\%, 50\%, 90\%)$, thereby $d_j = (10\%, 40\%, 40\%)$ in Eq. 4 for the *SEL3*. In our experiment, the parameter k is set to 1. The results for the different *single selection* approaches on the testing set, as well as the baseline are summarized in Tab. 1. We observe that the accuracy of *SEL3* is further boosted from *SEL2*, which shows the efficiency of the combination of frame spans and likelihood probabilities. We also notice that the accuracy of *SEL3* is indeed comparable with the baseline, while the number of frames used by the *SEL3* is much less than by the baseline in the decision process.

Table 1: *The comparison for the different single selection methods on the testing set*

Baseline: HMM with 10ms frame rate	73.98%
SEL1: purely ML based	63.30%
SEL2: 0%-50%-100%	70.43%
SEL2: 10%-50%-90%	73.08%
SEL2: 20%-50%-80%	71.17%
SEL3	73.67%

3.7. Global spectral movement in selected frames

A three-frame combination selected by the *SEL3* after discarding most frames shows comparable performance to the traditional

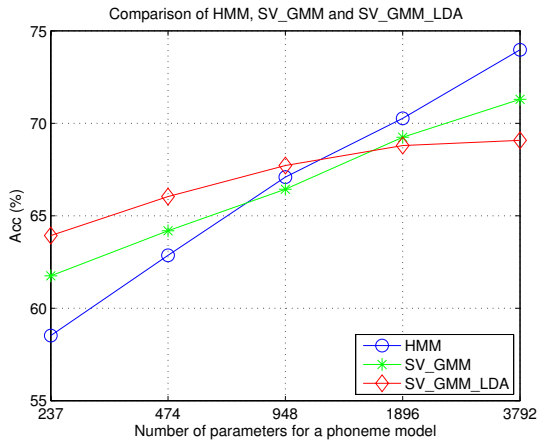

Figure 3: *The comparison of GMM, SV_GMM and SV_GMM_LDA with different numbers of parameters in a phoneme model*

Table 2: *The phoneme classification rates for HMM,* SV_GMM *and* SV_LDA_GMM *with different degrees of freedom. nP: total number of parameters; nS: number of states for a phoneme model; ngps: number of gaussian per state; nD: number of parameters for a gaussian*

Appr. nP in a model $= nS \times ngps \times nD$	HMM (3 states)	*SV_GMM* (1 state)	*SV_LDA_GMM* (1 state)
237	58.52%	**61.75%**	**63.93%**
474	62.86%	**64.19%**	**66.04%**
948	67.09%	66.44%	**67.72%**
1896	70.27%	69.25%	68.80%
3792	73.98%	71.30%	69.08%

HMM. Another interesting experiment is to investigate the global characteristics of the three frames and the evolution of their spectra. For this purpose, the three frames are concatenated to compose a 117-dimensional super vector for each phoneme segment in both the training set and testing set. For each phoneme, a gaussian mixture model (GMM) with a diagonal covariance, like a one-state HMM, is used to model super vectors in the training set. For evaluation, a super vector is classified to a phoneme whose GMM obtains the maximal probability. We denote this implementation as *SV_GMM*. Obviously a weakness of this model is that there is a strong correlation between corresponding cepstral coefficients in different frames. While it is hard to estimate a full covariance matrix for each phoneme GMM due to the high dimension, especially for some infrequent phonemes, the linear discriminant analysis (LDA) technique is adopted to reduce the correlations. An LDA matrix is estimated from the pool of training super vectors, by which each super vector is transformed to a 39-dimensional vector. Afterward, phoneme GMMs, denoted as *SV_LDA_GMM*, are trained using the transformed vectors.

Recognition rates for the traditional HMM, *SV_GMM* and *SV_LDA_GMM* are presented as a function of the number of free parameters in Tab. 2 and Fig. 3. We can see that in low model complexities, both the *SV_GMM* and the *SV_LDA_GMM* surprisingly beat HMM significantly, although in large number of parameters, the HMM outperforms the *SV_GMM* and the *SV_LDA_GMM*. Note that the average number of frames for a phoneme segment is approximately eight, which means after frame selection, only $3/8$ number of frames are used to train *SV_GMM* and *SV_LDA_GMM* models. With the limited training data, the selected frames could more precisely model the global evolution of phoneme variations and show better distinctions. When the number of parameters increases, this advantage is counteracted by lack of sufficient training data. Comparing the *SV_GMM* and the *SV_LDA_GMM* themselves, the LDA transformation indeed reduce the correlations and further enhance model's validity in low complexities.

4. Conclusions

The *single selection* can be seen as an attempt to automatically decide the reliability of a frame. Frames at different positions carry different information. Some of them help to increase the discrimination among phonemes, while some might be useless, or even obscure the difference. The experiments in this paper have shown that the most representative frame for a state does not only rely on its likelihood probability but also depends on its position in the phoneme segment. With three selected frames in the decision process, we still achieve fair performance comparable to the traditional HMM. Furthermore, the phoneme models trained with selected frames are not affected by vague, or less informative frames, thus they are more distinctive with moderate number of parameters.

5. References

[1] T. Y. Wu, D. Van Compernolle, J. Duchateau, and H. Van hamme, "Maximum likelihood based temporal frame selection," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 349–352.

[2] J. H. James and A. 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.

[3] J. Louradour, K. Daoudi, and R. André-Obrecht, "Discriminative power of transient frames in speaker recognition," in *Proc. ICASSP*, Philadelphia, USA, March 2005, pp. 613–616.

[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] J. Epps and E.H.C. Choi, "An energy search approach to variable frame rate front-end processing for robust ASR," in *Proc. EUROSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 2613–2616.

[6] J. Nedel and R. Stern, "Duration normalization for improved recognition of spontaneous and read speech via missing feature methods," in *Proc. ICASSP*, Salt Lake City, USA, May 2001, pp. 313–316.

[7] 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.

[8] K.F. Lee and H.W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on ASSP*, vol. 37, no. 11, pp. 1641–1648, 1989.