On the Use of Higher Frame Rate in the Training Phase of ASR

DARKO PEKAR, NIKSA JAKOVLJEVIC, MARKO JANEV, DRAGISA MISKOVIC, VLADO DELIC
Department of Telecommunications and Signal Processing
University of Novi Sad
Trg Dositeja Obradića 6
SERBIA
darko.pekar@alfanum.co.rs, jakovnik@uns.ac.rs, markojan@uns.ac.rs, dragisa@uns.ac.rs, vdelic@uns.ac.rs

Abstract: - The number of observations which are the basis for parameter estimation plays an important role in the quality of acoustic models. HMM based automatic speech recognition (ASR) systems generally have to cope with an insufficient number of observations for a good estimate. One way of tackling this problem is a well known procedure of state-tying, which is performed in order to gather sufficient information for a reasonable estimate for a large number of models. This procedure introduces an additional bias into the estimates, often leading to poor recognition results. In this paper a simple alternative to that solution is offered. It should be noted that most existing ASR systems use the same frame step size of 10ms in the training of the acoustical models, justifying it with the fact that speech signals exhibit quasi-stationary behavior at shorter durations. We claim that it is fully acceptable to adopt a much smaller frame step size in the acoustical training, thus providing estimators with a significantly higher number of observations compared to the standard 10ms case. This results in better parameter estimates and consequently better recognition results. Beside being justifiable from a phonetical point of view, it is also supported by results of an experimental on a real ASR system.

Key-Words: - ASR, GMM, variable frame rate, parameter estimation, kernel smoothing, KL divergence

1 Introduction
It is well known that estimators in general need a sufficient number of observations in order to provide reasonable estimates [6], [7]. This is also the case when estimating parameters of HMM models (such as the parameters of the Gaussian mixtures) used in continuous ASR systems. The estimators most commonly used for estimation of HMM models are maximum likelihood (ML) estimators (Baum-Welch algorithm, etc.). They are asymptotically unbiased and consistent with their variance reaching the Kramer-Rao lower bound in the case when the number of observations $N \rightarrow \infty$ [7]. Unfortunately, the size of the speech database is always limited, and thus for many acoustically context dependent models (triphones, diphones) an insufficient number of observations is collected. Parameter estimators are sensitive to this problem (especially ML estimators), which leads to poor acoustical estimates and to an increase in the word error rate (WER). A commonly used method to deal with this problem in ASR community is state-tying. It successfully deals with the unseen triphone issue and to some extent with the problem of data sparsity, but it can occasionally lead to undesirably general models and consequently to an increase in WER. Nevertheless, in almost all speech processing systems speech signals are firstly windowed and converted into an array of frames. The frames are typically 20-30ms long, while the frame step size is 10ms. This is especially typical for HMM-based ASR systems. The justification for such segmentation is that speech signals are non-stationary but exhibit rather quasi-stationary behavior at shorter durations. It is known, however, that certain acoustical attributes of speech can manifest themselves at very short durations [1], [2], [9]. Such attributes may be critical for identification and discrimination of speech sounds. A significant number of papers have exploited that fact, but all of them focused on the use of variable frame step size in the actual recognition phase [1], [4], [8]. For example, in [1] the authors presented theoretical and acoustical explanations and experimental results that support the previous claim, as well as the claim that since many consonant-vowel type syllables (for example nasal-/i/) have a very weak formant structure and short transitions, the frame step size of 10ms is too wide to capture the changes. Those conclusions are used to support the idea of using a variable frame rate in the recognition phase, and consequently, when the system detects an increase in the energy weighted Euclidean MFCC distance, the frame step size is decreased to 2ms, and vice versa. However, to our knowledge there were no attempts to use the same idea in order to get more accurate acoustical models in the training phase, and this paper investigates that approach. The idea is to provide a sufficient number of observations to the ML estimators of the HMM model in order to get more accurate estimates and, consequently, more appropriate models. On the other hand, the standard frame step rate is...
retained in the recognition phase, in order to avoid an increase in the computational load.

In Section 2, we present an analysis which supports the given claims from a phonetical aspect. In Section 3, some explanations are given on the method of statistical comparison between acoustical models obtained by using the 10ms frame step size and those obtained by using a much lower frame step size (2ms). Section 4 presents the experimental results, which show that, in some cases, there are significant differences (in terms of KL divergence measure) between the pdf's of the acoustical models obtained by using different frame rates. The experimental results also show that those differences are more significant on the consonant-vowel type of syllables, especially when they include stop or nasal consonants. A very important observation is that the WER of the system that uses the smaller frame step size is lower than the WER of the other system.

2 Phonetic considerations
Changes in spectral characteristics are important cues for discriminating between speech sounds and identifying them. These changes can occur over very short intervals. In [2], the authors explored that fact in order to gain a more representative front-end for ASR focusing on two aspects of audition not included in current representations: the short-term adaptation and the sensitivity to the frequency position of local spectral peaks. Both of these aspects can sometimes manifest themselves over very short durations. In addition, in [3], the authors analyzed human and machine recognition of nasal consonants in the presence of noise. They noticed the strong vowel-context effect with C-/a/ (where C stands for any nasal consonant) being the most robust one in the presence of noise and C-/i/ being the least robust one. The authors concluded that this was the case because the formant transitions in C-/a/ are longer than in the case of C-/i/. They also reported longer formant transitions and greater intensity of the F2 component in a C-/a/ syllable (15-20ms) than in the case of any other syllable. Based on that fact, in [1] the authors suggested the use of variable frame step size in the actual speech recognition process in order to catch all significant formant transitions and thus obtain better recognition results. We will illustrate that fact with an example obtained from the training base that we have used.

In Fig. 1, the first 12 MFCC coefficients for a 100ms long speech segment of the syllable /p/-/i/ are presented, for two different frame rates – the baseline frame step size of 10ms, and the one that we have used (2ms). It can be seen that in the case of the 2ms frame step, the transition region is represented by a larger number of points which could not be obtained by any kind of interpolation from those obtained for a 10ms frame step.

In [1] it was suggested that on syllables with longer formant transitions the system should use a lower frame rate in order to compensate for the additional computational load that resulted from the increase in the frame rate described above. In [4] the authors even proposed the use of an entropy measure in order to detect the situations with shorter formant transitions, where a lower frame step size should be employed, and vice versa.

It is questionable, however, whether the actual detection (and consequently frame rate compensation) could be done efficiently so that the recognition accuracy is improved without increasing the computational load. Our conclusion was that it would be quite reasonable to use a similar idea in the training process as well. This can be carried out in a straightforward way by using a smaller frame step size to obtain a much greater number of observations and thus more accurate estimates of actual HMM parameters. The recognition process itself uses a standard 10ms frame step size, whereby an additional increase in the computational load is avoided. However, recognition accuracy is increased due to the usage of more accurate estimates. As [1], [2], and [3] report, in some cases (C-/i/ for example) formant transitions can be up to 5 times shorter than in other cases. Therefore, by decreasing the frame step size from 10ms to 2ms we obtain five times more observations for critical models without fear that the additional data introduced is statistically redundant.
3 Used statistical methods
In order to support the above discussion with evidence, an experiment has been conducted on a real ASR system. The goal of the experiment was to show that there is a statistical difference between the histograms i.e. underlying densities obtained with the standard baseline frame step size of 10ms and with the proposed step size of 2ms. As expected, it was found that the difference is more significant in the consonant-vowel type of syllables with shorter formant transitions, and that it is less significant within the vowel-vowel syllables where the transitions are longer.

3.1 Kernel smoothing density estimation
In order to obtain density estimates that fit the training data properly, the Gaussian smoothing kernel (GSK) method was chosen. We felt that a non-parametric method should be applied in order to fit the data sufficiently well. In the case of parametric methods, like e.g. Gaussian mixture modeling (GMM), the number of mixtures could be a restrictive factor that can mask the real density divergence when the models are compared. Consequently, for a certain syllable model, the underlying density is estimated using multidimensional GSK technique

\[ f_H(x) = \frac{1}{N} \sum_{i=1}^{N} K_H(x-x_i) \]  

where \( K_H(\cdot) \) is given by:

\[ K_H(x) = |H|^{-1/2} K(H^{-1/2}x) \]  

and \( K(\cdot) \) is the multivariate Gaussian kernel function

\[ K(x) = (2\pi)^{-d/2} e^{-\frac{1}{2} x^T H^{-1} x} \]

and \( H \) is a real symmetric positive definite matrix called the bandwidth matrix. The terms \( x_i \in R^d, i \in \{1, \ldots, N\} \) are the observations that belong to the particular model, on the base on which the density is estimated, while \( x \in R^d \) is an arbitrary point in the feature space. In order to obtain the optimal \( H \), the method proposed in [5] is used: under the assumption that the data are observed from the normal density and the bandwidth matrix is diagonal.

\[ H = \text{diag}(h_1, \ldots, h_d) \]

the optimal bandwidth that minimizes the mean integrated square error (MISE) is obtained [6] as:

\[ h_i = \sigma_i \left( \frac{4}{(d+2)N} \right)^{1/(d+4)} \]

for \( i \in \{1, \ldots, d\} \), where \( \sigma_i \) is the standard deviation of the \( i \)-th variate i.e. \( \sigma_i^2 = \lambda_i \), where \( \lambda_i \) is the \( i \)-th eigenvalue of the particular model data covariance matrix \( \text{Cov}(X) \), \( X = [x_1, \ldots, x_N] \), for the full covariance case.

3.2 Density divergence measure used for model comparison
In order to obtain information on whether any non-redundant data are acquired by increasing the frame rate, the underlying densities of the particular models with the standard and increased frame rates are compared using the Kullback-Leibler (KL) divergence. The densities are estimated based on the GSK technique described in the previous section. The KL divergence between densities \( p \) and \( q \) is defined by:

\[ KL(p,q) = \int_{\mathbb{R}^d} p(x) \ln \frac{p(x)}{q(x)} dx \]  

For particular models \( M_1 \) and \( M_2 \), it is calculated as:

\[ KL(f_1,f_2) = \sum_{\alpha=1}^{M} f_1(x_{\alpha}) \ln \frac{f_1(x_{\alpha})}{f_2(x_{\alpha})} \]

where \( f_1 \) and \( f_2 \) are densities estimated using GSK with appropriate observations corresponding to the models \( M_1 \) and \( M_2 \), respectively, in order to approximate underlying densities. The points \( \{x_{\alpha}\}, \alpha \in \{1,\ldots,M\} \), are random evaluation points uniformly distributed across a significant bounded area \( \Omega \subset \mathbb{R}^d \).

4 Experimental results
In this section the experimental results obtained on real data and real ASR systems are presented. For experimental justification of the proposed idea, two ASR systems are compared, one with the standard 10ms frame step size in the training phase, and the other that uses a shorter (2ms) step. The results clearly show that the system with 2ms frame step size achieves better recognition results in terms of WER.

Also, the collected measurements show that in the case of many consonant-vowel type syllables, KL divergence between the models of the baseline system and the system with the shorter frame step is significant.

4.1 Model comparison results
The obtained experimental results confirm that there is a certain amount of non-redundant data introduced into the training process as a result of the decrease in the frame step size. The estimated densities are significantly different as well. This difference is measured by the KL divergence, as explained in the previous section. In Table 1, KL divergences between the acoustical models obtained for the baseline ASR system and the proposed one are presented.

Both systems use the same mechanism for feature extraction – tree base clustering (TBC) tying strategy [8], and the same training and recognition algorithms. The only difference is the value adopted as the minimal
number of observations needed for a TBC leaf to be used for estimation of emitting HMM densities. The thresholds in the system that uses the 2ms frame step were set to be $5 = 10ms/2ms$ times larger than in the system that uses the 10ms frame step, in order to exclude the differences introduced by model tying.

Table 2: Recognition results for the baseline system and the phonetic considerations given in Section 2.

<table>
<thead>
<tr>
<th>Type of ASR system</th>
<th># states</th>
<th># mixtures</th>
<th>WER[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>REF_sys (10ms fr.)</td>
<td>9113</td>
<td>37249</td>
<td>5.5</td>
</tr>
<tr>
<td>TEST_sys_1 (2ms fr.)</td>
<td>9982</td>
<td>38012</td>
<td>5.1</td>
</tr>
<tr>
<td>TEST_sys_2 (2ms fr.)</td>
<td>9982</td>
<td>46970</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Table 1: KL divergence for various syllables

The results are presented for various syllables that belong to various phonetic groups. Note that KL divergence of the measured syllables is considerably higher in the case of consonant-vowel syllables than in the case of vowel-vowel syllables. The divergence is maximal in the case of stops. In the case of nasals, the formant transitions are shorter than in other cases, confirming the phonetic considerations given in Section 2.

### 4.2 Recognition results

The experiment results also include results obtained on real ASR systems. As previously noted, one system uses the standard frame step of 10ms, and the other one uses the 2ms frame step in the training phase.

### 5 Conclusion

In this paper we argue that it is reasonable to use a higher frame rate in the training phase of the ASR systems than the one that is commonly used in most such systems. This claim is supported by experimental results on actual ASR systems that show that there is significant difference in terms of KL divergence between the model pdf’s obtained for the baseline system that uses a 10ms frame step, and the proposed system that uses a 2ms frame step. In addition, recognition results show that the proposed system achieves a significantly lower WER than the baseline system.

### 6 Acknowledgements

The work described in this paper was partially financed by the Ministry of Science and Technological Development of the Republic of Serbia, within the technological project No. 11001 “Human-Machine Communication”.

References:


