Speech Decoding using Markov model: search for a prior criterion of quality

S. Soudoplatoff

ABSTRACT

Among all possibilities for Markov models applied to speech recognition, a separation exists between systems that deal with continuous parameters, i.e. lying in $\mathbb{R}^n$, such as spectrum coefficients, LPC, etc..., using parametric density functions, and systems that maps these values into a finite vocabulary, whose size is very small (a few hundreds elements), usually done by a procedure known as vector quantization, or labelling. This paper presents a set of experiments that were run both to compare various types of labelling, and to search for a prior criterion of the quality of a labelling, thus avoiding to go through a complete experiment (training of the parameters, and decoding) anytime the input parameters, or the labels are changed. The standard labelling, which is the reference here, consists in clustering a set of vectors to obtain prototypes, using K-mean type algorithms, then assigning each vector to the nearest class center, according to a Euclidian distance. Other labels are obtained, by changing one, or both elements, of the metric space (acoustic vectors, metric). The experiments were run on parts of a 200000 words, isolated syllables, speech dictation system for French, which is under research at the Paris Scientific Center. They consist of a phonetic decoding, using a sub-optimal strategy. Criteria are obtained either by considering the labels as output of a channel, or from contingency tables. For each set of labels, two different hypothesis were made, depending whether one considers that the output are labels, or strings of labels. Criteria related to information theory, such as mutual information, or related to data analysis, such as Phi square, or Jordan, are computed. Actual results show that lower error rates than one obtained with the reference labels can be achieved, and that the results are globally consistent with the criteria.

INTRODUCTION

Automatic dictation using Hidden Markov Models (HMM) is a 3 steps process, namely acoustic processing, acoustic modeling, and linguistic modeling (Ref 1). They correspond, in the fundamental equation of speech recognition:

$$\hat{W} = \text{Argmax}_{W} p(A|W)p(W)$$

(1)

to the three terms $A$, $p(A|W)$ and $p(W)$. The last term, $p(W)$ is the linguistic model, or the probability of a sequence of words $W$, and is independent of the other two. A Markov model is an approximation for this sequence, which is commonly used to model the emission of a source. The second one, $p(A|W)$, or acoustic model, is the probability of an acoustic observation given a sequence of words, and is approximated using HMM. The first one, or acoustic processing, can be viewed as a pre-process whose aim is to compress the input speech, to discard part of its great redundancy.

The HMM itself, which can be used for other applications than dictation, is basically made of two elements (Ref 7):

- a first-order Markov process to model an unseen sequence of states,
- a zero-th order Markov process to model the emission of the acoustic given a sequence of states.

Variations can be found on the use of HMM, but we shall mention here only the two classes for the probabilities of the acoustic given the state sequence, namely parametric and non-parametric, leading to two different acoustic processing. In the first case, the compression of speech stops after the computation of a highly dimensional space vector, whose probability is modelized usually using one, or many Gaussians (Ref 3). In the second one, speech is further compressed, using vector quantization, and the probability is to be considered only over a finite set. The training of these different probabilities is made via a procedure known as Baum-Welsh algorithm, which has been extensively described in the literature (Ref 2).

Decoding, like training, is the same whatever assumption is made concerning the parametrization of speech, and depends only upon the type of speech processing one is doing (phonetic recognition, word recognition, dictation, etc...).

Our concern is the following: each time a new set of labels is obtained, that is either a new vector quantizer, or a new signal processing is designed, how can we avoid to go through the entire process of training and decoding for testing it. This requirement is due to the heavy computer work necessary to train the data, and to decode for comparing the results with other experiments. The problem is then to compare different labellings, using good criteria, i.e. that are as much as possible consistent (i.e. decreasing) with the error rate.

---

1 Centre Scientifique IBM, 36 Av Raymond Poincaré, 75116 Paris, France.
THE GROUND TRUTH

Inspired by techniques used in the field of remote sensing, we introduced the notion of "ground truth", which is the following: let us assume that we have a script for which we are provided with both the speech signal corresponding to its pronunciation, and a segmentation in phonetic machines (which we shall call phonemes in this paper), for instance obtained by hand; then, for a given labelling, we derive a correspondence between labels and phonemes, under the form of a contingency table. The elements of this table are, with standard notation \( n_{uv} \), and represent the number of labels \( u \) that were associated with the phoneme \( v \). One has to find if there is a relation between a value that would measure the "quality" of the labelling, and the decoding results that one would obtain using it. Although such a relation is not strictly straightforward, its bounds are easy to determine:

- the worst decoding is obtained when there is independence between the labels and the phonemes, because all entries are equi-probables. Actually, it would even not be possible to train those labels.
- The best decoding is obtained when there is complete association between the labels and the phonemes, i.e. a partition of the label space in phonemes, for in that case the vector quantizer itself is doing a perfect decoding.

THE CRITERIA

The next step is to compute criteria on those tables. That is, we need a function \( L \rightarrow \mathbb{R} \), which is minimum in the worst case, and maxima in the best one. There exist many criteria in the literature, that act on contingency tables, and obey this constraint (Ref 5).

With the following notations: \( n_{uv} \) = number of elements in the table for label \( u \) against phoneme \( v \), \( n_u = \sum_v n_{uv} \), \( n_v = \sum_u n_{uv} \), \( N = \sum_{uv} n_{uv} \), we present here two of the most well-known criteria.

- Phi square, given by \( \Phi^2 = \sum_{uv} \frac{n_{uv}^2}{n_u n_v} - 1 \).
- Jordan, given by \( J = \sum_{uv} \frac{n_{uv}}{N} \left( \frac{n_{uv}}{n_u n_v} - 1 \right) \).

Both have the property that they are minima (and equal to zero) in the case of independence, and maxima in the case of complete association, which was our requirement.

However, we adopted another viewpoint for our problem, more inspired from communication theory. We associate to a given vector quantizer the set of probability values \( p(l, \phi) \), \( l \) being a label and \( \phi \) a phoneme, from which we can deduce the margins \( p(l) \) and \( p(\phi) \). It is important to notice that the latter probability is not the probability of emission of a phoneme in the French language, but rather looks like this value multiplied (and normalized) by the average number of labels for a phoneme. We can now look at the situation as the following channel, were labels are the input, phonemes the output, and the vector quantizer is acting in the following way: "the vision of the phoneme \( \phi \) changes the a priori \( p(l) \) to the a posteriori \( p(l \mid \phi) \)." Then, other types of criteria can be used, among which the mutual information is the most common:

\[
I = \sum_{l, \phi} p(l, \phi) \log \left( \frac{p(l \mid \phi)}{p(l)} \right).
\]

It has the required properties, i.e. is minimum when there is independence \( p(l \mid \phi) = p(l) \), \( \forall l, \forall \phi \), and maximum in the case of complete association (there exist a bijection between the set of phonemes and a partition of the set of labels, \( P_1 \), such as \( p(l \mid \phi) = 0 \), \( \forall \phi \notin P_1 \)).

But this vision of the problem allows us now to introduce the last criterion. Let us suppose that we make a decoding which would be based on the output of the vector quantizer. Then, we can estimate the error that we would make, by counting the number of times the most probable phoneme for a given label is not the one that was pronounced, which is, if we introduce \( \lambda(l) = \arg \max_{\phi} p(\phi \mid l) \), the number of pairs \( l, \phi \) for which \( \phi \neq \lambda(l) \).

We resume the criteria that we are dealing with, by giving their formal values.

- Phi square \( \Phi^2 = \sum_{l, \phi} \frac{p(l, \phi)^2}{p(l) p(\phi)} - 1 \).
- Jordan \( J = \sum_{l, \phi} p(l, \phi) \left( p(l, \phi) - p(l) p(\phi) \right) \).
- Mutual information \( I = \sum_{l, \phi} p(l, \phi) \log \left( \frac{p(l \mid \phi)}{p(l) p(\phi)} \right) \).
- Expected error \( e = \sum_{l, \phi} p(l, \phi) (1 - \delta_{\lambda(l) \phi}) \), with \( \delta_{ij} \) the kronecker symbol.
However, one should keep in mind that the values $n_{ij}$ are only estimator for the probabilities $p(l, q)$, and that the criterion we shall so compute are biased, with a bias quite difficult to estimate for certain. Fortunately, they are asymptotically unbiased.

**SEQUENCES OF LABELS**

The main problem, with this formulation, is the fact that the decoding is not performed on a label by label basis, but rather on sequences of labels. If we take this into account, still using a channel, we must change our space, put the phoneme in input, and the space of all label sequences as output. Although the segmentation allows us to obtain relations between sequences of labels and phoneme in the same manner, it is impossible to make any usable contingency tables on that, since we have very few of the occurrences.

But, we can notice that all our criteria can be rewritten in term of expectation over the probabilities of all possible couples $(L, \varphi)$ of label sequences and phonemes.

- $\phi^2 = E \left( \frac{p(L, \varphi)}{p(L)} - 1 \right)$.
- $J = E \left( p(\varphi) \left( p(L, \varphi) - p(L) \right) \right)$.
- $I = E \left( \log \left( \frac{p(L \varphi)}{p(L)} \right) \right)$.
- $e = E \left( 1 - \delta_{L, \lambda_{\varphi}} \right)$, with $\lambda(L) = \arg\max_{\varphi} p(L, \varphi)$.

To estimate those values, we can first use the law of large number to compute the expectations. Then, $p(\varphi)$, which is now the a priori probability of a phoneme, is estimated by counts; thus remains only the problem of $p(L, \varphi)$, since $p(L)$ is easily derived from those values. We choose an approximation which is close to the Markovian case,

$$\tilde{p}(L, \varphi) = p(\nu(L) = n) \prod_{r=1}^{n} p(l_r, \varphi).$$

(3)

with $\nu(L)$ the number of labels in $L$, and $p(L, \varphi)$ estimated by counts, as in the first case. The probability of the length is a little bit more delicate, but can be treated by considering that it is independent of a particular labelling, and so using a HMM already available allows to estimate those probabilities,

$$p(\nu(L) = n) = \sum_{1 \leq i_1 < \ldots < i_n} p(s_1, \ldots, s_n).$$

(4)

The main problem with the estimator of the criterion is of course that they are asymptotically biased, for the expectation is taken on the "true" probability distribution $p'(L, \varphi)$. However, as in the first case, we shall use those figures only for comparison.

**THE LABELLINGS**

We present now the experiments that were made on different labellings. They were run on parts of a 200K words, syllable based recognition system under research at the Paris Scientific Center (Ref 6). They all consists of:

- design of a new set of labels on both a training and a decoding corpus,
- computation of the criteria on part of the training corpus, and a manual segmentation,
- estimation of the parameters of the HMM using Baum-Welsh algorithm,
- phonetic decoding of a test corpus, using a sub-optimal strategy, by computing the probability of a sentence over the most probable path of states (Ref 4).

We worked on the phonemes, since the result we obtained were in accordance with the result we obtained on syllable decoding. When necessary, we used the first experiment as a basis for estimating values, as for instance the length probability, used by the others. The experiments are here shortly described. The signal is digitized at 10kHz, differentiated unless otherwise stated. The last two experiments were intentionally designed to give bad labels for one, by discarding all the information contained in the energy, and perfect labels for the other, by creating a partition of the 200 labels into the phonemes (complete association).

1. FFT on 128 points hamming windows, K-Mean and labels using Euclidian distance.
2. FFT on 256 points overlapping hamming windows, K-Mean and labels using Euclidian distance.
3. FFT on 256 points window null outside a pitch period, K-Mean and Labels using Euclidian distance.
4. FFT on 256 points overlapping hamming windows, K-Mean and labels using Manhattan distance except for silence in K-Mean, using Euclidian.
5. non-differentiated signal, FFT on 256 points overlapping hamming windows, K-Mean and labels using Euclidian distance.
6. FFT on 256 points overlapping hamming windows, K-Mean and labels using Manhattan distance.
7. FFT on 256 points overlapping hamming windows, spectrum equalization, K-Mean and labels using Euclidian distance.
8. Perfect labels.

The results presented are the recognition rates on phonemes:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>82.1</td>
<td>82.4</td>
<td>83.1</td>
<td>84.5</td>
<td>82.4</td>
<td>84.2</td>
<td>75.4</td>
<td>100.0</td>
</tr>
</tbody>
</table>

They show that a manhattan distance $\sum|x_i - y_i|$ gives better results than Euclidian distance.

The next figures present the 4 criterion computed, left on labels, and right on sequences of labels, opposed to the error rates.

**Figure 1:** The four figures on the left are the criteria for the first set of experiences (on labels), on the right for the second set of experiences (on sequences of labels).

They show that the phi square, the mutual information, and the error expectation behave the same for the first experiment. This seems reasonable, at least for the first two criteria, because they measure the same phenomena and are the same at first order. We notice that they are globally decreasing with the error rate. They are not perfectly decreasing, due to errors which occur at three levels: error in the manual segmentation, error in the estimation of the criteria, and imprecision in the decoding error rate which is also estimated on a single test corpus.

As a conclusion, criterion such as mutual information, phi square, or expected error can be computed on labels versus phonemes to estimate the quality of a vector quantizer. Although they are not strictly decreasing with the error rate, a trend exists. The next step would be the estimation of the error made on the computation of the criteria, to allow for statistical tests.

**BIBLIOGRAPHY**