MODELLING SPEAKING RATE USING A BETWEEN FRAME DISTANCE METRIC

Andreas Tuerk   *Steve Young*

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
at233@eng.cam.ac.uk,sjy@eng.cam.ac.uk

ABSTRACT

It is well known [5] that variations in speaking rate can account for a significant percentage of errors in practical speech recognition tasks. This is the result of the dynamic nature of speech which is not modelled properly by the standard HMM structure. This paper proposes an extension to the standard HMM that takes advantage of the information about the rate of speech that is contained in inter-frame transitions. The new model can be seen as a combination of Moore and Mealy type HMM's that has output probabilities attached to the transitions between states in addition to the conventional output probabilities attached to states. In this model fast and slow transitions are associated with additional hidden parameters. The output probabilities of the transitions are modelled with gamma distributions.

1. Introduction

This paper tries to overcome two limitations of the standard HMM structure. Firstly, the assumption that successive frames are independent and secondly that a sequence of states which represent part-of- phoneme events account for all the hidden variables in speech. The two main assumptions in this paper are that speaking rate is also an observable that should be modelled by a hidden variable and that there is a relationship between speaking rate and the dependency of successive frames. This paper provides evidence for these assumptions and shows how they can be modelled within the HMM frame-work. In section 2 a straight forward measure of the dependency of successive frames is proposed and its use in deriving information about speaking rate is shown. In section 3 this relationship is included into a standard HMM structure and the properties of this combined Moore-Mealy type HMM are studied.

2. Speaking Rate and the Dependency of Successive Frames

The basic idea of this section is to show that the dependency of successive frames carries information about speaking rate. This is not surprising because the faster the speech the faster the change in the measurements that are applied at each time frame. Therefore one would expect that for fast speaking rates the feature vectors are less dependent, whereas they are closely correlated for slow speaking rates.

2.1. Measuring Speaking Rate

Several measures of speaking rate have been proposed [3, 4] that are defined on various macroscopic levels. They include measurements of phones, words, utterances or whole speakers. The definition of speaking rate that is used throughout this paper is applied to phones. A phone is considered to belong to a certain speaking rate bin if its duration lies between the two boundaries that define this bin. In the case of two speaking rates there is only one threshold which divides the occurrences of a phoneme into fast and slow. These boundaries were derived for each phone separately depending on the statistics of its duration. In the cross-word triphone context that these experiments were carried out in this means that all the triphones that share the same central phoneme have the same speaking rate bins.

2.2. Measuring the Dependency of Successive Frames

If one thinks of a phone as being located in a certain bounded volume in feature space, a possible measure of dependency is the Euclidean distance between successive frames. This measure is useful because the average distance between a feature vector and its nearest neighbour decreases as the number of feature vectors in the volume increases. This corresponds to a decrease in speaking rate. Since the movement of feature vectors through the feature space is not random but follows certain trajectories one would also expect that for an increasing number of vectors in the given volume not only the distance to the nearest neighbour but also the distance between successive frames decreases. For these reasons the measure of dependency between successive frames in this paper is chosen to be the Euclidean distance.

2.3. The Correlation between the Dependency of Successive Frames and Speaking Rate

The following experiments were carried out on a subset of the Broadcast News 1997 training data. The feature vectors were PLP coefficients with an energy component but without first or second order derivatives. The speaking rate bins were derived as described in section 2.1.

The Correlation between Means and Variances of the Euclidean Distance and Speaking Rate

In table 1 and the following tables number one is assigned to the slowest speaking rate and the labels increase with increasing speaking rate. As can be seen from table 1 speaking rate and Euclidean distance between successive frames are monotonically correlated. The same holds true for the variances. This is not very surprising because one would expect that with increasing speaking rate the location of the next feature vector becomes less predictable. Given this relationship one could think of speaking rate as a time parameter in a stochastic process, where as time passes the paths of the process tend to move away from the initial state and become more spread out.

speaking rate	1	2	3	4
mean distance	6.726	7.353	8.411	9.046
variance	11.499	13.432	14.055	14.364
nr. frames	30926	46522	54121	6705
nr. distances	29799	43653	47925	5449

Table 1: Statistics for phoneme **ey**

The Distributions of the Euclidean Distance

In addition to the correlation between Euclidean distance and speaking rate, the distributions of the distance itself reveal an interesting structure. Figure 1 shows an original distribution (continuous line) as compared to an approximation by a gamma distribution (dotted line). As can be seen, the distribution of distance is very close to the gamma distribution.

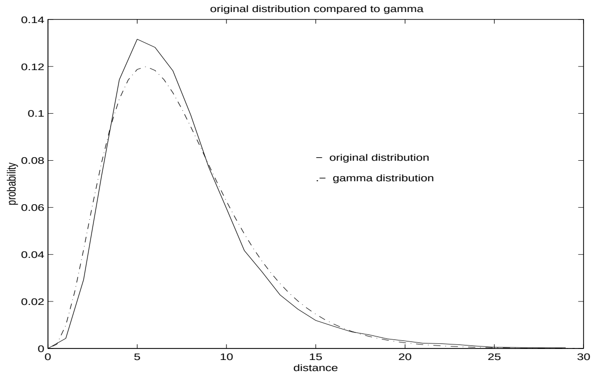

Figure 1: Original and approximating Gamma distributions for phoneme ao at a certain speaking rate

Using the Euclidean Distance as a Classifier for Speaking Rate

For the following experiments MAP estimators were constructed that classified pairs of frames or whole phones as either fast or slow depending on the Euclidean distance between their successive feature vectors. As can be seen from table 2 the discrimination between a class of fast and a class of slow transitions allows for the construction of a classifier with reasonable performance. The first two columns of table 2 show the number of pairs of frames that were classified as either belonging to speaking rate one or two. The last two columns in this table give the number of whole phones that were assigned to one of the two classes. For these experiments a whole phone was classified depending on the predominant class that was assigned to its individual transitions. The rows show the true label of a transition or phone, respectively. From this table one can see that 57.0% of transitions between successive feature vectors were classified correctly, whereas 60.4% of phones were assigned to the correct class.

true \ est.	1	2	1	2
1	204310	158725	12888	7484
2	145941	199560	16635	23933

Table 2: Estimation for phoneme **ey** and two speaking rates based on the length of the phoneme

2.4. An alternative Definition of Speaking Rate

If one uses the Euclidean distance between successive frames directly as a measure of speaking rate, then it is possible to classify a phoneme as being fast if the mean distance between successive frames within the phoneme exceeds a certain threshold, otherwise the phoneme is classified as slow. This was done for the following experiments. As shown in figure 2 the distance distributions become more discriminative for this definition. In this figure the distributions of the Euclidean distance for two speaking rates under the two definitions of speaking rate are shown. The continuous line shows the distributions for the definition of speaking rate from the previous section. The dotted line gives the distributions for the definition that is discussed in this section. In both cases the distribution with the higher mean represents the fast speaking rate. As a result of this increased discrimination, the performance of the MAP estimators increases as well. This can be seen from table 3 which shows that 66.6% of transitions between successive frames were classified correctly and 86.1% of all the occurrences of phoneme **ey**.

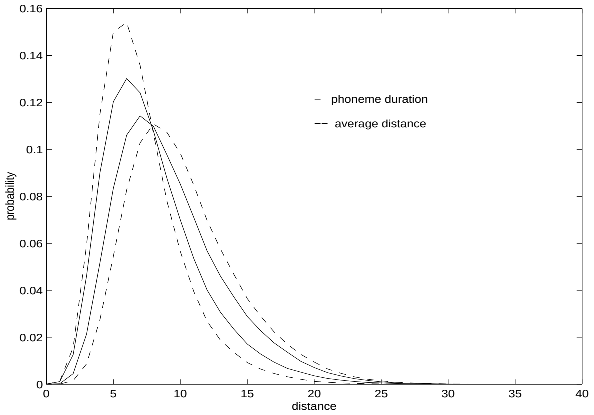

Figure 2: Distance distributions for splitting phoneme axr into two speaking rates by phoneme duration and average distance between frames within one phoneme

true \ est.	1	2	1	2
1	277072	77205	23820	2772
2	159184	195075	5678	28670

Table 3: Estimation for phoneme **ey** and two speaking rates based on the average distance between frames within the phoneme

3. Including Information about Speaking Rate into a standard HMM

The last section showed that there is information about the duration of phones contained within inter-frame transitions. It would therefore be desirable to modify the structure of the standard HMM in such a way as to include this additional information.

3.1. Rescoring

A straight forward approach to make use of the durational information in inter-frame transitions is to rescore a given hypothesised state sequence $\vec{s}$ by multiplying the likelihood of the observation sequence O given the state sequence $\vec{s}$ by the likelihood of observing the sequence of distances D given the sequence of states and therefore the sequence of speaking rates $\vec{v}$. Therefore the likelihood of observing the sequence of feature vectors O and the sequence of distances D becomes

$$L(O, D, \vec{s}) = L(O, \vec{s}) L(D|\vec{s}) \tag{1}$$

where $L(D|\vec{s}) = L(D|\vec{v})$ and $\vec{v}$ is the sequence of speaking rates that is uniquely determined by the sequence $\vec{s}$.
Although one could use the EM algorithm to reestimate the parameters of this model it has the disadvantage that the training procedure cannot learn the notion of speaking rate that is most advantageous for the recognition process. This is because the notion of speaking rate is completely determined by the initialisation.

3.2. Creating a new HMM Structure

If the estimators that were constructed in the last section were perfect maximum likelihood estimators, the following would hold true.

$$L(D|\vec{v}) = \max_{\vec{v}} L(D, \vec{v}|\vec{s}) \tag{2}$$

In this case one can maximise the likelihood of the observation sequences O, D by maximising with respect to $\vec{s}$ and $\vec{v}$ independently, i.e.

$$\max_{\vec{s}} L(O, D, \vec{s}) = \max_{\vec{s}, \vec{v}} L(O, \vec{s}) L(D, \vec{v}|\vec{s}) \tag{3}$$

Although (2) holds only to a certain extent, in this paper the sequences $\vec{s}$ and $\vec{v}$ are treated as being independent. The likelihood of the combined sequences O, D therefore becomes

$$L(O, D, \vec{s}, \vec{v}) = L(O, \vec{s}) L(D, \vec{v}|\vec{s}) \tag{4}$$

Here speaking rate is associated with a new hidden variable v. This model has the advantage that it is less susceptible to the initial definition of speaking rate. Furthermore, as will be shown later the parameters of this model can be reestimated efficiently in the EM frame-work.

3.3. Some Aspects of the new HMM Structure

The likelihood of the sequence D
The likelihood of the sequence of distances D is defined by

$$L(D, \vec{v}|\vec{s}) = \prod_t p(d_t|v_t, s_t, s_{t-1}) p(v_t|v_{t-1}, s_t, s_{t-1}) \tag{5}$$

In this equation $p(d_t|v_t, s_t, s_{t-1})$ stands for the output probability of the transition at time t, which moves from state s_{t-1} to state s_t with speaking rate v_t. This output probability is defined on the distances between successive frames. As was shown in the last section these probabilities can be modelled by gamma distributions and therefore

$$p(d|n, i, j) = \frac{{\eta_{n,i,j}}^{\nu_{n,i,j}}}{\Gamma(\nu_{n,i,j})} d^{\nu_{n,i,j}-1} e^{-\eta_{n,i,j} d} \tag{6}$$

where (n,i,j) stands for the transition from state i to state j at speaking rate n.
The expression $p(v_t|v_{t-1}, s_t, s_{t-1})$ in (5) is the probability for observing speaking rate v_t, given that speaking rate v_{t-1} was observed at the previous transition and the current transition moves from state s_{t-1} to state s_t. Since the notions of speaking rate that were used so far are defined on a per-phone basis these transition probabilities are assumed to be zero for transitions between different speaking rates within a phoneme, whereas between phoneme transitions can be learned from the training data. This is a restriction which is not essential to the structure of the HMM in (4) but which was adopted in this work due to the macroscopic nature of the concept of speaking rate.

The Topology of the new Model
Figure 3 shows the topology of the new model for two hidden variables that correspond to a fast and a slow speaking rate. The HMM structure in this example is the usual left-to-right structure. As can be seen from this figure the restriction that the speaking rate cannot be changed within the model results in two separate paths that can be taken through the model. Each path has its distinctive output probabilities that are attached to the transitions between the states. In principle the output probabilities of the transitions are different for each pair of states. The states in the two paths are actually the same and store the same state output distributions.

3.4. EM Reestimation

The reestimation formulae for the existing HMM parameters stay essentially the same. For example, the following equation shows how the means of the Gaussians are reestimated

$$\bar{\mu}_j = \frac{\sum_t L_\lambda(O, D, s_t = j) o_t}{\sum_t L_\lambda(O, D, s_t = j)}$$

where

$$L_\lambda(O, D, s_t = j) = \sum_{\vec{s}: s_t = j} L_\lambda(O, \vec{s}) L_\lambda(D|\vec{s})$$

It is more interesting to look at the reestimation formulae for the parameters of the gamma distributions. These

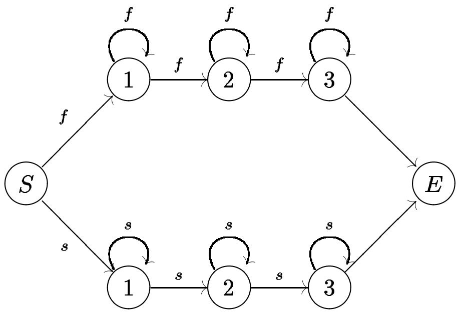

Figure 3: The topology of the new HMM for two speaking rates

are given by

$$\begin{aligned}\frac{\bar{\nu}_{n,j,i}}{\bar{\eta}_{n,j,i}} &= \frac{\sum_t L_\lambda(O,D,n_t,j_t,i_{t-1})d_t}{\sum_t L_\lambda(O,D,n_t,j_t,i_{t-1})} \\ \frac{\Gamma'(\bar{\nu}_{n,j,i})}{\Gamma(\bar{\nu}_{n,j,i})} - \log\bar{\eta}_{n,j,i} &= \frac{\sum_t L_\lambda(O,D,n_t,j_t,i_{t-1})\log d_t}{\sum_t L_\lambda(O,D,n_t,j_t,i_{t-1})}\end{aligned} \tag{7}$$

Here the first equation in (7) is the result of calculating the derivative with respect to $\bar{\eta}_{n,j,i}$ and the second equation results from calculating the derivative with respect to $\bar{\nu}_{n,j,i}$. Note that these formulae differ from their equivalents in [2]. There the reestimation of η, η_{new} was performed by using ν_{old} and similarly η_{old} was used to estimate ν_{new}. In (7) ν_{new} and η_{new} are estimated dependent on each other. The second equation in (7) can be directly used to reestimate the parameters since the behaviour of the function Γ'/Γ, which is known as the ψ-function, is well understood (see [1]). Denoting the right hand side of the first equation in (7) as $m_{n,j,i}$ and the right hand side of the second equation in (7) as $l_{n,j,i}$ and substituting the first into the second equation one can rewrite the second equation in (7) as follows.

$$\psi(\bar{\nu}_{n,j,i}) - \log\bar{\nu}_{n,j,i} = l_{n,j,i} - \log m_{n,j,i} \tag{8}$$

Applying Newton's algorithm one can find estimates for $\bar{\nu}_{n,j,i}$ iteratively by using the following formula

$$\bar{\nu}^{(k+1)}_{n,j,i} = \bar{\nu}^{(k)}_{n,j,i} - \frac{\psi(\bar{\nu}^{(k)}_{n,j,i}) - \log\bar{\nu}^{(k)}_{n,j,i} + \log m_{n,j,i} - l_{n,j,i}}{\psi'(\bar{\nu}^{(k)}_{n,j,i}) - \frac{1}{\bar{\nu}^{(k)}_{n,j,i}}}$$

Given a proper value for $\bar{\nu}_{n,j,i}$ the parameter $\bar{\eta}_{n,j,i}$ can now be calculated from (7).

Interpretation of the Reestimation Formulae

The first equation in (7) is intuitively plausible since $\bar{\nu}_{n,j,i}/\bar{\eta}_{n,j,i}$ is the mean value of the reestimated gamma distribution. However, for the second reestimation equation one might have guessed

$$\frac{\bar{\nu}_{n,j,i}}{\bar{\eta}^2_{n,j,i}} = \frac{\sum_t L_\lambda(O,D,n_t,j_t,i_{t-1})(d_t - m_{n,j,i})^2}{\sum_t L_\lambda(O,D,n_t,j_t,i_t)} \tag{9}$$

because the left hand side of this equation is the variance of the gamma distribution. However, equation (8) can be interpreted in a similar way, where the concept of variance has to be substituted by a KL distance. To see this suppose there is a probability distribution $p(o)$ over the possible outcomes o of an experiment. This describes the situation in the previous section. There the possible outcomes are the d_t's and their probabilities are given by the appropriate ratios in (7). Now, in this frame-work the right hand side of (8) can be written as follows.

$$\begin{aligned}\sum_o p(o)\log o - \log\left(\sum_o p(o)o\right) &= \sum_o p(o)\log\frac{o}{\sum_o p(o)o} \\ &= \sum_o p(o)\log\frac{\frac{p(o)o}{\sum_o p(o)o}}{p(o)} \\ &= -D(p\|q) \qquad (10)\end{aligned}$$

Here $D(p\|q)$ is the Kullback-Leibler distance between the probability distributions p and q where one has to set

$$q(o) = \frac{p(o)o}{\sum_o p(o)o} \tag{11}$$

The value $-D(p\|q)$ can be interpreted as a variance measure, because it depends on the distribution of the values of the possible outcomes around their mean. The closer each individual outcome is to the average the closer the distribution q is to p. Interestingly, the left hand side of (8) can be seen to be just the continuous equivalent of this variance measure for the gamma distribution.

4. Conclusion

It has been shown in this paper that experimental evidence suggests that the transitions between states can be modelled individually depending on the speaking rate. A possible alternative to the standard HMM based on this assumption was proposed. The EM reestimation formulae for this model were derived and shown to be intuitively understandable. Work is now in progress to evaluate this model on the 1997 Broadcast News test set which has a statistically significant portion of spontaneous speech.

5. REFERENCES

1. M. Abromovitz and J.A. Stegun, editors. *Handbook of Mathematical Functions*. New York: Dover Publications, Inc., 1965.

2. S.E. Levinson. Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1:29 – 45, 1986.

3. N. Mirghafori, E. Fosler, and N. Morgan. Making Automatic Speech Recognition More Robust to Fast Speech. Technical report, International Computer Science Institute, December 1995.

4. N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proc.Eurospeech*, volume 4, pages 2079 – 2082, 1997.

5. D. Pallett, J. Fiscus, J. Garofolo, B. Lund, and M. Przybocki. 1993 benchmark tests for the ARPA Spoken Language Program. In *Proc.ARPA Workshop on Spoken Language Technology*, pages 15–40, 1994.