

HIDDEN MARKOV MODELS FOR VISUAL SPEECH SYNTHESIS WITH LIMITED DATA

Allan Arb^{*}, Steven Gustafson[†], Timothy Anderson^{††}, Raymond Slyh^{††}

^{*}Air Force Research Laboratory, 3550 Aberdeen SE, Kirtland AFB, NM 87117

[†]Air Force Institute of Technology, 2950 P Street Bldg 640, Wright-Patterson AFB, OH 45433

^{††}Air Force Research Laboratory, 2255 H Street Bldg 248, Wright-Patterson AFB, OH 45433

Email: h.a.arb@ieee.org, steven.gustafson@afit.edu, {tim.anderson, raymond.slyh}@wpafb.af.mil

ABSTRACT

This paper addresses a problem often encountered when estimating control points used in visual speech synthesis. First, Hidden Markov Models (HMMs) are estimated for each viseme present in stored video data. Second, models are generated for each triseme (a viseme in context with the previous and following visemes) in the training set. Next, a decision tree is used to cluster and relate states in the HMMs that are similar in a contextual and statistical sense. The tree is also used to estimate HMMs for any trisemes that are not present in the stored video data when control points for such trisemes are required for synthesizing the lip motion for a sentence. Finally, the HMMs are used to generate sequences of visual speech control points for those trisemes not occurring in the stored data. Comparisons of mouth shapes generated from the artificially generated control points and the control points estimated from video not used to train the HMMs indicate that the process estimated accurate control points for the trisemes tested. This paper thus establishes a useful method for synthesizing realistic audio-synchronized video facial features.

1. INTRODUCTION

One of the major problems facing visual speech synthesis is ensuring that the mouth area moves in a realistic fashion according to the text of the spoken words. It is well known that humans use more than the auditory input to perceive speech. For example, deaf people can be taught to read lips (i.e., visual speech reading). Also, researchers have found that people with normal hearing use a visual mode of perception to enhance aural perception, especially in noisy environments [1,2].

The authors of [3] present a system that synthesizes video of a human lip-synced to a recorded audio track. Bregler and colleagues used 8 minutes of training footage for much of their development work [3]. In many applications the amount of available footage is much less (perhaps less than 3 minutes). Obviously, the probability of encountering a desired triseme in the training footage decreases as the amount of the training footage decreases. Bregler

presents a method of selecting the stored segment to use as a representative triseme based on a distance metric, but errors can and do occur. Incorrect triseme selection could be due to the particular distance metric chosen, or the closest triseme could be simply an unrealistic rendering of the desired mouth motion.

This paper presents the use of Hidden Markov Models (HMMs) for estimating control points for unseen trisemes. First, HMMs are trained for each triseme in the stored footage. Next, using context-based decision-tree state tying, HMMs are estimated for any triseme desired for synthesis that is not in the training footage. Finally, these HMMs are used to estimate control points for use in warping algorithms producing synthetic visual speech.

The next sections present the algorithm for training the HMMs (including audio/visual speech database recording), the synthesis of control points for trisemes not present in the training footage, and experimental results of the synthetic points relative to hand-labeled control points.

2. DATA PREPARATION

Two steps were performed to generate the data required to evaluate the use of HMMs in visual speech synthesis. First, an audio/visual database was recorded. Then, visual speech synthesis control points were estimated and converted to features used to train the HMMs.

2.1. Video Database Recording

Since a visual speech database with full sentences wasn't readily available, we recorded a small database of one speaker uttering 97 prompts. The video was captured at 30 fps with 240 X 320 pixel 24-bit color frames. The monaural audio was sampled at 22 kHz with 16-bit quantization.

The 97 prompts were selected from the DARPA TIMIT speech corpus [4] such that there were a reasonable number of trisemes occurring multiple times. Of the 17,576 (26^3) theoretically possible number of trisemes, 1,453 exist in the 97 recorded



sentences. The entire database of 97 sentences contains 5 minutes 4 seconds of audio and video.

Figure 1: Example of video captured for the database.

Viseme	Phoneme(s)	Phoneme(s)	Viseme
A	/ch/,/jh/,/sh/,zh/	/k/,/g/,/n/,/l/	B
C	/t/,/d/,/s/,/z/	/p/,/b/,/m/	D
E	/f/,/v/	/th/,/dh/	F
G	/w/,/r/	/hh/	H
I	/y/	/ng/	J
K	/eh/	/ey/	L
M	/er/	/uh/	N
O	/aa/	/ao/	P
Q	/aw/	/ay/	R
S	/uw/	/ow/	T
U	/oy/	/iy/	V
W	/ih/	/ae/	X
Y	/ah/	/sil/	Z

Table 1: Phoneme to viseme groupings.

The subject was seated approximately 19 inches from the camera and the entire head and shoulders were captured. Figure 1 illustrates the captured view.

Following the recording, the audio and video were separated for further processing. The audio was automatically phonetically aligned resulting in transcripts containing a list of the phonemes present in the audio and their start and stop times. Once we had the phonetic transcripts, the phoneme labels were converted to the proper viseme class as listed in Table 1.

2.2 Control Point Estimation

The visual speech control points consisted of one each at the outer and inner mouth corners, 7 each on the upper and lower outer lip contours, 5 each on the upper and lower inner mouth contours, and 1 each for the lower edge of the upper teeth and the upper edge of the lower teeth respectively for a total of 30 control points. When the teeth were obscured by the

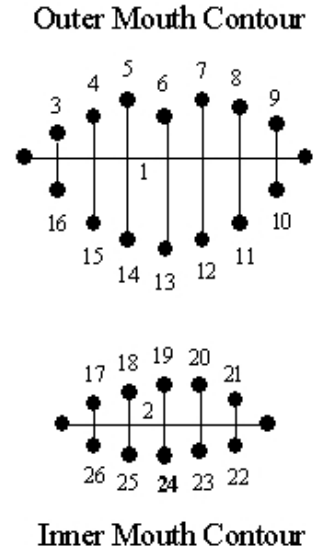


Figure 2: Distances derived from control points.

lips, the edge locations were estimated based on typical positions for the specific viseme.

The Eigenpoints algorithm derived by Covell [5] was used to estimate initial control point locations from models estimated using 200 images with hand-located control points. Each Eigenpoints control point estimate was manually reviewed and corrected when necessary.

2.3 Distance Features

In lieu of using the control point coordinates directly, the work presented here uses distance features derived from those coordinates. Additionally, features capturing the dynamic characteristics of the static features for each frame are calculated. Finally, since there are many occasions where there are not enough examples (frames) in a given sentence to use three-state no-skip HMMs, the 30 fps data is upsampled to 90 fps.

Figure 2 illustrates the 28 static (single frame) distance features derived from the 30 control points. The horizontal distances 1 and 2 are simply the outer and inner mouth widths. The vertical distances are orthogonal distances from the control point to the line connecting the mouth corners. All distances are in pixels.

Let \mathbf{c}_t be the static distance feature vector at time t . The dynamic features at time t are calculated as:

$$\Delta \mathbf{c}_t = \sum_{i=-\Theta}^{\Theta} w(i) \mathbf{c}_{t+i} \quad (1)$$

where Θ is the number of past and future feature vectors to include in the computation and

$$w(i) = \frac{i}{2\sum_{\theta=1}^{\Theta} \theta^2}. \quad (2)$$

Note that Equation 1 implies reliance on features from frames prior to and after the triseme of interest. Care should be taken in selecting Θ as a large Θ could widen the dynamic window to include a previous or future triseme not included in the training data. This situation would require dynamic feature calculation using synthetic data as data would not exist for frames corresponding to those missing trisemes. To ensure this did not happen $\Theta = 1$ for this work.

3. HMM TRAINING AND TRISEME SYNTHESIS

First, three-state left-to-right HMMs are trained for each of the 26 visemes using the static and dynamic features from a set of 49 training sentences (2.75 minutes). Next, these viseme models are cloned to form context-dependent viseme models (or triseme models) for each triseme occurring in the training data. The state transition matrices are tied for all triseme models formed from the same core viseme.

Given the large number of triseme models and the limited amount of training data available for each model, one can use common states across models reducing the number of overall model parameters requiring estimation. That is, states from multiple models can be clustered together into a common state essentially increasing the amount of data available for estimating that state's parameters. While there exist several methods for determining how to cluster the states, this research uses a decision tree to identify which states should be clustered and tied together.

A visemic decision tree is a binary tree where a yes or no question is attached to each node. Figure 3 illustrates the concept for questions related to the type of phonemes the visemes represent. Trees are built using a top-down sequential optimization process. Initially, all models are placed in a single cluster at the root of the tree. Once the increase in log-likelihood resulting from a node split falls below a threshold, splitting stops and states remaining at each new node are grouped into a single state. The log-likelihood threshold used for results presented here was 100. See [6] for analysis of a threshold of 20 and other synthesis factors; in particular threshold settings do affect the resultant quality of the synthetic control points for the trisemes analyzed.

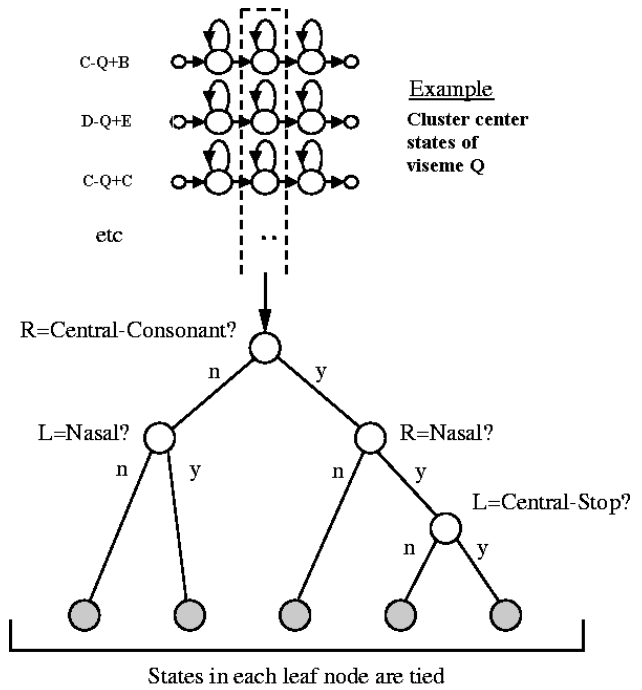


Figure 3: An example of decision tree-based state tying.

The advantage of decision tree-based state tying is that it allows for estimation of models for trisemes not present in the training set [7,8]. State models are chosen for the new trisemes given the decision tree statistics and contextual questions.

In [9], the authors present the use of context-based decision trees for synthesizing control vectors for 3D mesh visual speech synthesis models. While the questions used in [9] were similar to those used here, the trees here do not produce the final feature estimates. Rather, they are used to cluster states from HMMs of trisemes present in the stored footage and estimate new HMMs for trisemes not present in the footage. Additionally, over one hour of stored video footage was used vice the 2.75 minutes used here.

Once HMMs are estimated for each triseme in the training set and for those not in the training set but required for synthesis, new sequences of features \mathbf{c} are estimated using methods derived from [10]. The goal is to find the state sequence, \mathbf{q} , and feature sequence \mathbf{c} that maximizes

$$\begin{aligned} \log P[\mathbf{O} | \mathbf{q}, \lambda] = & -\frac{1}{2}(\mathbf{c} - \boldsymbol{\mu})^T \mathbf{U}^{-1}(\mathbf{c} - \boldsymbol{\mu}) \\ & -\frac{1}{2}(\Delta \mathbf{c} - \Delta \boldsymbol{\mu})^T \Delta \mathbf{U}^{-1}(\Delta \mathbf{c} - \Delta \boldsymbol{\mu}) \\ & -\frac{1}{2} \sum_{t=1}^T \log |\boldsymbol{\Sigma} \mathbf{q}_t| - \sum_{t=1}^T \log |\Delta \boldsymbol{\Sigma} \mathbf{q}_t| \\ & - TM \log 2\pi, \end{aligned} \quad (3)$$

where M is the number of features per vector (28 here), T is the number of frames synthesized, and

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{q}_1}^T & \boldsymbol{\mu}_{\mathbf{q}_2}^T & \cdots & \boldsymbol{\mu}_{\mathbf{q}_T}^T \end{bmatrix}^T \quad (4)$$

$$\mathbf{U} = \text{diag} \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{q}_1} & \boldsymbol{\Sigma}_{\mathbf{q}_2} & \cdots & \boldsymbol{\Sigma}_{\mathbf{q}_T} \end{bmatrix} \quad (5)$$

$$\Delta\boldsymbol{\mu} = \begin{bmatrix} \Delta\boldsymbol{\mu}_{\mathbf{q}_1}^T & \Delta\boldsymbol{\mu}_{\mathbf{q}_2}^T & \cdots & \Delta\boldsymbol{\mu}_{\mathbf{q}_T}^T \end{bmatrix}^T \quad (6)$$

$$\Delta\mathbf{U} = \text{diag} \begin{bmatrix} \Delta\boldsymbol{\Sigma}_{\mathbf{q}_1} & \Delta\boldsymbol{\Sigma}_{\mathbf{q}_2} & \cdots & \Delta\boldsymbol{\Sigma}_{\mathbf{q}_T} \end{bmatrix}. \quad (7)$$

To make Equation (3) independent of $\Delta\mathbf{c}$, we make the substitution

$$\Delta\mathbf{c} = \mathbf{W}\mathbf{c} - \mathbf{b} \quad (8)$$

where

$$\mathbf{W} = \begin{bmatrix} w(0)\mathbf{I}_M & \cdots & w(\Theta)\mathbf{I}_M & \mathbf{0} \\ \vdots & w(0)\mathbf{I}_M & \ddots & \vdots \\ w(-\Theta)\mathbf{I}_M & \ddots & \ddots & w(\Theta)\mathbf{I}_M \\ \mathbf{0} & \ddots & w(-\Theta)\mathbf{I}_M & \cdots & w(0)\mathbf{I}_M \end{bmatrix} \quad (9)$$

with \mathbf{I}_M as the $M \times M$ identity matrix and

$$\mathbf{b} = \begin{bmatrix} -w(-\Theta)\mathbf{c}_{\Theta}^T - w(-\Theta+1)\mathbf{c}_{\Theta+1}^T - \cdots - w(-1)\mathbf{c}_{-1}^T \\ -w(-\Theta)\mathbf{c}_{\Theta+1}^T - w(-\Theta+1)\mathbf{c}_{\Theta+2}^T - \cdots - w(-2)\mathbf{c}_{-1}^T \\ \vdots \\ -w(-\Theta)\mathbf{c}_{-1}^T \\ \mathbf{0}_{M(T-2\Theta) \times 1} \\ -w(\Theta)\mathbf{c}_{T+1}^T \\ \vdots \\ -w(\Theta)\mathbf{c}_{T+\Theta-1}^T - w(\Theta-1)\mathbf{c}_{T+\Theta-2}^T - \cdots - w(2)\mathbf{c}_{T+1}^T \\ -w(\Theta)\mathbf{c}_{T+\Theta}^T - w(\Theta-1)\mathbf{c}_{T+\Theta-1}^T - \cdots - w(1)\mathbf{c}_{T+1}^T \end{bmatrix} \quad (10)$$

To maximize Equation (3), we assume there exists a \mathbf{c}^* such that the partial derivative of Equation (3) with respect to \mathbf{c} evaluated at \mathbf{c}^* yields $\mathbf{0}_{TM \times 1}$. The candidate \mathbf{c}^* for the optimal \mathbf{c} is then

$$\mathbf{c}^* = (\mathbf{U}^{-1} + \mathbf{W}^T \Delta\mathbf{U}^{-1} \mathbf{W})^{-1} (\mathbf{U}^{-1} \boldsymbol{\mu} + \mathbf{W}^T \Delta\mathbf{U}^{-1} (\Delta\boldsymbol{\mu} + \mathbf{b})) \quad (11)$$

Sequences of synthetic distance features are estimated using Equation (11) for every possible state sequence, \mathbf{q} . The state sequence/distance feature sequence pair is selected that produces the largest likelihood result when evaluating Equation (3). The synthetic distance features are then downsampled to 30 fps and converted back to control points for use in the visual speech synthesis system.

4. EXPERIMENTAL RESULTS

In order to investigate the performance of the technique, the following procedure was performed. First, a number of sentences were selected from the portion of the database that was not used for training. Second, trisemes that were not seen in the training data were identified for synthesis from the list of trisemes in the selected sentences. Third, the distance features were synthesized for the unseen trisemes using the procedure discussed in the previous section. Next, the distance features were converted to control points, and the control points were connected to form mouth outlines. Finally, the outlines from the HMM estimated control points were overlaid on the outlines from the original video of the selected sentences.

The next three sections present the outlines for three unseen (not in the data used to initially train the HMMs) trisemes: C-D+X, A-O+G and E-W+B covering a variety of lip positions. The outlines do not include the teeth locations as it is difficult to present the locations in a clear manner in a paper of this length. For illustrations of the teeth location performance, see [6]. Also included in the discussion is the mean squared error between the synthetic features (including teeth locations) and the features calculated using control points from the recorded video. The results presented here are representative of the results produced from the models trained and trisemes investigated in [6].

4.1 Triseme C-D+X

Figure 4 illustrates four examples, each from different sentences, of the triseme C-D+X. The dotted lines in the figures represent the outlines constructed using control points from the original video, while the solid lines represent those outlines from the HMM estimated control points. It is clear that for the models trained, the outlines from synthetic control points are very close to those from the original video for this triseme. The mean squared error (MSE) is also quite small as Table 2 shows. The largest MSE is only 3.01 pixels² and many are less than 0.5 pixels².

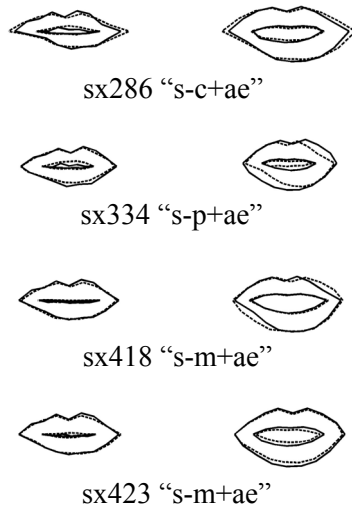


Figure 4: Lip contours for triseme C-D+X (triphones listed with TIMIT sentence ID). The dotted lines are the originally estimated outline and the solid are the synthetic outlines.

TIMIT ID	MSE per Feature (pixels ²)	
	Frame 1	Frame 2
sx286	2.76	2.45
sx334	1.35	3.01
sx418	0.47	0.42
sx423	0.43	1.92

Table 2: MSE per feature between features estimated from HMMs and features from the original video control points for triseme C-D+X.

4.2 Triseme A-O+G

Figure 5 illustrates four examples (again from different sentences) of the triseme A-O+G (triphone ch-aa+r). The mouth is primarily open for all frames of this triseme. Once again, the outlines generated from control points estimated from HMMs are quite close to those from control points estimated from the original video. Table 3 lists the MSE for these examples.

4.3 Triseme E-W+B

Finally, Figure 6 shows another set of four examples from triseme E-W+B. Again, the outlines formed from control points estimated using HMMs are very similar to those from control points estimated directly from recorded video. Table 4 lists the MSE

TIMIT ID	MSE per Feature (pixels ²)		
	Frame 1	Frame 2	Frame 3
sx286	0.29	0.18	
sx334	0.64		
sx418	0.93	1.12	
sx423	0.46	1.41	1.06

Table 3: MSE per for triseme A-O+G.

TIMIT ID	MSE per Feature (pixels ²)		
	Frame 1	Frame 2	Frame 3
sx286	1.19	1.90	4.41
sx334	0.30	0.78	
sx418	0.66	0.83	0.38
sx423	0.44		

Table 4: MSE per feature for triseme E-W+B.

per feature for these examples. Most are less than 1 pixel².

5. CONCLUSIONS

In audio-driven speech synthesis systems using images from stored video for synthesis, a problem arises when the required visual speech class is not available in the stored data. One solution would be to pick a speech class “close” to the required class using some “closeness metric”. This paper presented an alternative solution using context-based decision trees to estimate HMMs for the missing visual speech class. These HMMs were then used to generate control points for visual speech synthesis.

The results presented in Section 4 illustrate that the method worked well for the video database recorded for this work, the data used to train the models, and the trisemes presented.

Future research will include using the control points to warp new images, investigations into different question sets for the decision tree-based state tying, comparisons to other systems and studies on varying the visual perspective relative to the subject (*e.g.*, head tilt or rotation), additional facial locations (jaw and eye movement, etc), and alternative feature sets (in lieu of the distance features used here).

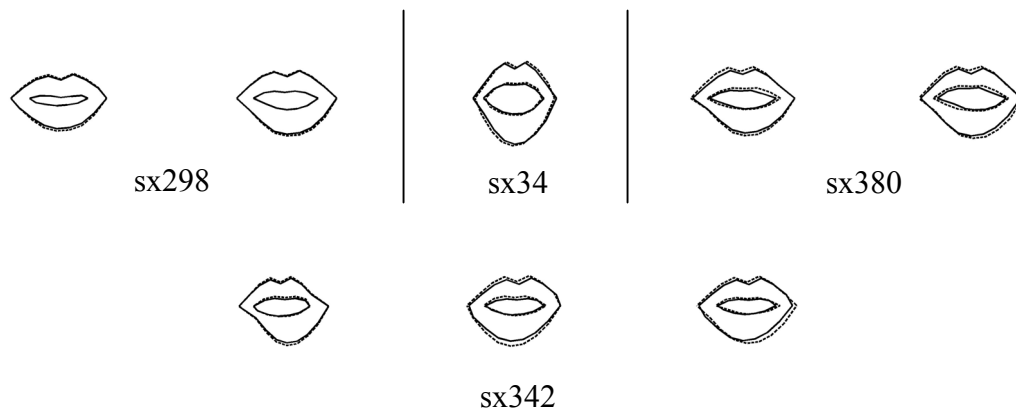


Figure 5: Lip contours for triseme A-O+G (triphone ch-aa+r).

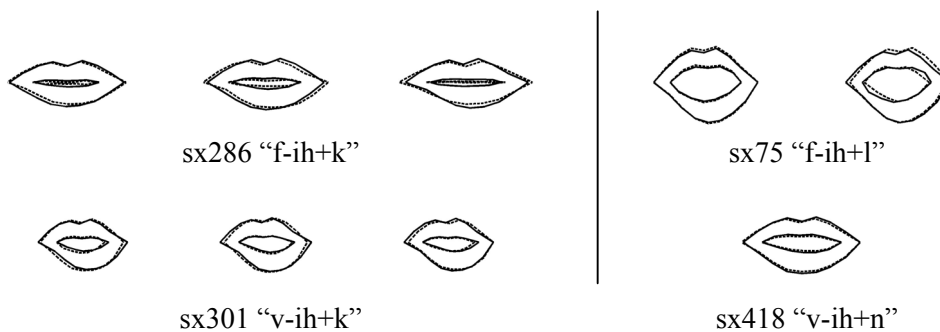


Figure 6: Lip contours for triseme E-W+B.

6. REFERENCES

1. McGurk, H. and J. MacDonald. "Hearing Lips and Seeing Voices," *Nature*, 264: 746-748 (December 1976).
2. Massaro, D., M. Cohen and P. Smeele. "Perception of asynchronous and conflicting visual and auditory speech," *Journal of the Acoustical Society of America*, 100(3):1777-1786 (September 1996).
3. Bregler, C., M. Covell and M. Slaney. "Video Rewrite: Driving Visual Speech with Audio." *SIGGRAPH 97 Computer Graphics Proceedings*. Annual Conference Series. 353-360. 1997.
4. Lamel, L.F., R.H. Kessel and S. Seneff. "Speech database development: Design and analysis of the acoustic-phonetic corpus." *Proceedings of the Speech Recognition Workshop (DARPA)*. 100-109. 1986.
5. Covell, M. "Eigenpoints: Control-point Location using Principal Component Analysis." *Proceedings of the second International Conference on Automatic Face and Gesture Recognition*. Killington, VT: IEEE, October 1996.
6. Arb, H.A. *Hidden Markov Models for Visual Speech Synthesis in Limited Data Environments*. PhD Dissertation, Air Force Institute of Technology, 2001.
7. Young, S.J., J.J. Odell and P.C. Woodland. "Tree-Based State Tying for High Accuracy Acoustic Modeling." *Proceedings of the ARPA Workshop on Human Language Technology*. 307-312. March 1994.
8. Odell, J.J., P.C. Woodland and S.J. Young. "Tree-Based State Clustering for Large Vocabulary Speech Recognition." *Proceedings of the 1994 International Symposium on Speech, Image Processing and Neural Networks*. 690-693. April 1994.
9. Galanes, F.M., and others. "Generation of Lip-Synched Synthetic Faces From Phonetically Clustered Face Movement Data." *Proceedings of the International Conference on Auditory-Visual Speech Processing*. 191-194. 1998.
10. Tokuda, K., T. Kobayashi and S. Imai. "Speech Parameter Generation from HMM Using Dynamic Features." *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing*. 660-663. May 1995.