Creating a Speech Enabled Avatar from a Single Photograph

Dmitri Bitouk* Shree K. Nayar†

Columbia University

Abstract

This paper presents a complete framework for creating a speech-enabled avatar from a single image of a person. Our approach uses a generic facial motion model which represents deformations of a prototype face during speech. We have developed an HMM-based facial animation algorithm which takes into account both lexical stress and coarticulation. This algorithm produces realistic animations of the prototype facial surface from either text or speech. The generic facial motion model can be transformed to a novel face geometry using a set of corresponding points between the prototype face surface and the novel face. Given a face photograph, a small number of manually selected features in the photograph are used to deform the prototype face surface. The deformed surface is then used to animate the face in the photograph. We show several examples of avatars that are driven by text and speech inputs.

Index Terms: H.5.2 [Information Interfaces and Presentation]: Multimedia Information Systems—Animations; I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation

1 Introduction

While a substantial amount of work has been done on developing human face avatars, we have yet to see avatars that are highly realistic in terms of animation as well as appearance. The goal of this paper is to create speech-enabled avatars of faces that provide realistic facial motion from text or speech inputs. Such speech-enabled avatars can significantly enhance user experience in a variety of applications including mobile messaging, information kiosks, advertising, news reporting and videoconferencing.

Our approach to facial animation employs the generic facial motion model previously introduced in [3]. The model represents a deformation of the 3D prototype facial surface due to articulation during speech as a linear combination of a small set of basis vector fields. The coefficients of this representation are the time-dependent facial motion parameters. In order to build a speaker-dependent facial motion model for a new subject, we first deform the prototype surface into a novel surface using a small set of feature points on the novel face. Then, the basis vector fields are adapted to the novel facial surface with the help of the deformation obtained in the previous step.

We train a set of Hidden Markov Models (HMMs) using the facial motion parameters obtained from motion capture data of a single speaker. Our facial motion synthesis algorithm utilizes the trained HHMs to generate facial motion parameters from either text or speech input.

We apply the facial motion synthesis algorithm to animate avatars created from a single photograph. Since depth information is not directly available from a single photograph, we flatten (project) both the prototype surface and the basis vectors of the facial motion model to obtain a reduced 2D representation. To create a avatar, we first select a few corresponding points between the prototype surface and the person's face in the photograph. Using these correspondences, we deform the prototype surface and adapt the basis vector fields to obtain the speaker-dependent facial motion model. We have developed real-time rendering software that can produce realistic facial animations of avatars from facial motion parameters generated from either text or speech input. In order to enhance realism, our rendering system also synthesizes eye gaze motion and blinking.

The main technical contributions of our work can be summarized as follows. First, we present an end-to-end system for building an avatar from a single photograph. Second, we have developed a novel, HMM-based facial motion synthesis algorithm. Our algorithm, compared to the existing HMM-based approaches [7, 15], takes into account the effects of lexical stress and co-articulation by learning them from the training data. Finally, to include the effects of eye gaze changes and blinking, we present a texture synthesis based method for obtaining a complete eyeball model from the incomplete view of an eye in a photograph.

In [4], we show how our approach can be used to build volumetric displays featuring speech-enabled 3D avatars. We use a simple method for recovering 3D face geometry and texture from a single mirror-based stereo image. A physical 3D avatar of the person's face is created by engraving the obtained facial surface inside a solid glass block using sub-surface laser engraving technology. The facial motion animation synthesized from text or speech is projected onto the static 3D avatar using a digital projector. Even though the physical shape of the avatar is static, the projection of facial animations onto it results in a compelling experience for the viewer.

2 Related Work

Our work is related to previous works in several fields, including computer graphics and computer vision. Here, we discuss the previous works that are most relevant to ours.

Facial Motion Representation: Existing approaches to facial motion synthesis fall into either image-based or model-based categories. Image-based approaches, such as [8, 11], rely on building statistical models which relate temporal changes in the images at a pixel level to the sequence of phonemes uttered by the speaker. A major disadvantage of image-based models is that they require a large training set of facial images in order to synthesize novel facial animations. In contrast, model-based approaches represent the shape of a speaker's face with either a 2D or 3D mesh. Articulatory facial motion is described as deformation of the mesh and is controlled by a set of parameters. One of the most popular techniques parameterizes mesh deformations with the help of muscle models [19] and uses facial muscle activations to produce facial animation. On the other hand, performance-driven approaches learn facial motion from recorded motions of people. In this paper, we take the model-based approach and use a compact parameterized facial model built from motion capture data presented in [3].

Facial Motion Synthesis: Given a parametric representation of facial motion, the role of speech synthesis algorithms is to generate parameter trajectories from a time-aligned sequence of phonemes. One of the approaches to visual speech synthesis is based on defining a key shape for each of the phonemes and smoothly interpolating between them [16]. Similar to acoustic speech synthesis, visual speech synthesis methods fall either into concatenative or HMM-

*e-mail: bitouk@cs.columbia.edu

†e-mail:nayar@cs.columbia.edu

based categories. Concatenative approaches rely on stitching together pre-recorded motion sequences, which correspond to triphones [8], phonemes [21] or even longer speech units [9]. HMM-based synthesis [7, 15], on the other hand, models the dynamics of visual speech with the help of hidden Markov models. Trajectories of facial motion parameters are generated from HMMs based on the maximum likelihood criteria. Our method trains HMMs that can capture the effects of both coarticulation as well as lexical stress and produce realistic facial motions from either text or speech inputs.

2D Avatars from a Photograph: Although a number of approaches to fitting a deformable model to a photograph have been suggested, generation of speech-enabled avatars from a single image remains an open research problem. For instance, Blantz et al. [5] developed a method to transfer static facial expressions obtained from laser scans to photographs. The main drawback of this work is its high computational cost. A few commercial systems (see [1, 2], for example) introduced recently aim to animate user-supplied facial images, but the facial motions they produce lack realism. We present an end-to-end system for creating speech-enabled avatars from a single photograph which can be animated from text or speech in real-time. We believe the realism of the visual speech produced by our approach is fairly high compared to those of existing systems.

3 Facial Motion Representation

Our approach to synthesizing facial animation from text or speech utilizes the 3D parametric facial motion model previously introduced in [3]. For the sake of completeness, we briefly review this model. First, a generic, speaker-independent facial motion model is estimated. Then, this model is adapted to a novel speaker's face.

The generic facial motion model describes deformations of the prototype face represented by a parametrized surface $\boldsymbol{x}(\boldsymbol{u}), \boldsymbol{x} \in \mathbb{R}^3, \boldsymbol{u} \in \mathbb{R}^2$. The displacement of the deformed face shape $\boldsymbol{x}_t(\boldsymbol{u})$ at the moment of time t during speech is represented as a linear combination of the basis vector fields $\boldsymbol{\psi}_k(\boldsymbol{u})$:

$$\boldsymbol{x}_t(\boldsymbol{u}) = \boldsymbol{x}(\boldsymbol{u}) + \sum_{k=1}^{N} \alpha_{k,t} \boldsymbol{\psi}_k(\boldsymbol{u}). \tag{1}$$

The vector fields $\boldsymbol{\psi}_k(\boldsymbol{u})$ defined on the prototype facial surface $\boldsymbol{x}(\boldsymbol{u})$ describe the principal modes of facial motion and are learned from motion capture data as described in [3]. At each instant of time, the deformation of the prototype facial surface is completely described by a vector of facial motion parameters $\boldsymbol{\alpha}_t = (\alpha_{1,t}, \alpha_{2,t}, ..., \alpha_{N,t})^T$. The dimensionality of the facial motion model is chosen to be $N = 9$.

The above basis vector fields are defined with the respect to the prototype surface and, thus, have to be adjusted to match the geometry of a novel face. We employed the method developed in [3] for facial motion transfer which enables one to map the generic facial motion model using a few corresponding points between the prototype face and the novel face.

3.1 Visual Speech Unit Selection

In large vocabulary speech applications, uttered words are considered to be composed of phones which are acoustic realizations of phonemes. We make use of the CMU phone set, which consists of 39 distinct phones along with a silence unit /SIL/ which describes inter-word intervals. In order to accommodate lexical stress, the most common vowel phonemes are cloned into stressed and unstressed phones (for example, /AA0/ and /AA1/). In particular, we chose to model both stressed and unstressed variants of phones /AA/, /AE/, /AH/, /AO/, /AY/, /EH/, /ER/, /EY/, /IH/, /IY/, /OW/ and /UW/. The rest of the vowels in the CMU set are modeled independent of their lexical stress. Each of the phones, including stressed and unstressed variants, is represented as a 2-state HMM, as shown in Figure 1, while the /SIL/ unit is described using 3-state topology. The HMM states s_1 and s_2 explicitly represent an onset and end of the corresponding phone. The output probability of each HMM state is assumed to be given by a Gaussian distribution over the facial parameters $\boldsymbol{\alpha}_t$, which correspond to the HMM observations.

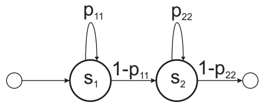


Figure 1: HMM topology for a phone in the CMU set. The allowed transition between the HMM states s_1 and s_2 are shown as arcs with the transition probabilities p_{ij}.

3.2 HMM Training

The goal of the HMM training procedure is to obtain maximum-likelihood estimates of the transition probabilities between HMM states and the sufficient statistics of the output probability densities for each HMM state from a set of observed facial motion parameter trajectories $\boldsymbol{\alpha}_t$, corresponding to the known sequence of words uttered by a speaker. We use facial motion parameter trajectories derived from the motion capture data in [3] as our training set.

In order to account for the dynamic nature of visual speech, we augment the original facial motion parameters $\boldsymbol{\alpha}_t$ with their first and second derivatives. Our implementation of HMM training is based on the Baum-Welch algorithm [18] and similar in spirit to the embedded re-estimation procedure [22]. Overall, the HMM training is realized in three major steps. First, a set of monophone HMMs is trained. Second, in order to capture co-articulation effects, monophone models are cloned into triphone HMMs which explicitly take into account left and right neighboring phones. Finally, we employ decision-tree based clustering of triphone states to improve robustness of the estimated HMM parameters and predict triphones that were not seen in the training set.

The training data consist of facial motion parameter trajectories $\boldsymbol{\alpha}_t$ and the corresponding word-level transcriptions. The dictionary employed in the HMM training process provides two instances of phone-level transcriptions for each of the words – the original transcription and a variant which ends with the silence unit /SIL/. The output probability densities of monophone HMM states are initialized as a Gaussian density with mean and covariance equal to the global mean and covariance of the training data. Subsequently, 6 iterations of the Baum-Welch re-estimation algorithm are performed in order to refine the HMM parameter estimates using transcriptions which contain the silence unit only at the beginning and the end of each utterance. Next, we apply the forced alignment procedure [22] to obtain hypothesized pronunciations of each utterance in the training set. The final monophone HMMs are constructed by performing 2 iterations of the Baum-Welch algorithm.

In order to capture the effects of coarticulation, we refine the obtained monophone HMMs into triphone models, which take into account the preceding and the following phones. The triphone HMMs are initialized by cloning the corresponding monophone models and are consequently refined by performing 2 iterations of the Baum-Welch algorithm. The triphone state models are clustered with the help of a tree-based procedure to reduce the dimensionality of the model and construct models for triphones unseen in the training set. The resulting models are often referred to as tied-state triphone HMMs in which the means and variances are constrained to be the same for triphone states belonging to a given cluster. The final set of tied-state triphone HMMs is obtained by applying another 2 iterations of the Baum-Welch algorithm.

4 Facial Motion Synthesis from Text and Speech

In order to synthesize trajectories of facial motion parameters $\boldsymbol{\alpha}_t$ either from text or acoustic speech signal, we firstly generate a sequence of time-labeled phones. When text is used as input, we employ an acoustic text-to-speech (TTS) engine to generate a waveform and the time-aligned sequence of phonemes. To synthesize facial animation from acoustic speech input, we utilize a speech recognizer and then use the forced alignment procedure [22] to obtain time-labels of the phones in the best hypothesis generated by the speech recognizer.

In the beginning of the synthesis stage, we convert the time-labeled phone sequence to an ordered set of context-dependent HMM states. Vowels are substituted with their lexical stress variants according to the most likely pronunciation chosen from the dictionary with the help of a monogram language model. Next, we create an HMM chain for the whole utterance by concatenating clustered HMMs of each triphone state from the decision tree constructed during the training stage. The resulting sequence consists of triphones and their start and end times. Since each triphone unit is modeled as a two-state HMM, the start and end times of HMM states cannot be directly obtained from phone-level segmentation. However, state-level segmentation can be inferred in the maximum-likelihood sense by utilizing the state transition probabilities estimated during HMM training stage.

The mean durations of the HMM states s_1 and s_2 with transition probabilities, as shown on Figure 1, can be computed as $p_{11}/1-p_{11}$ and $p_{22}/1-p_{22}$. If the duration of a triphone n described by a 2-state HMM in the phone-level segmentation is t_N, the durations $t_n^{(1)}$ and $t_n^{(2)}$ of its HMM states are proportional to their mean durations and are given by

$$t_n^{(1)} = \frac{p_{11} - p_{11}p_{22}}{p_{11}+p_{22}-p_{11}p_{22}} t_n, \quad t_n^{(2)} = \frac{p_{22} - p_{11}p_{22}}{p_{11}+p_{22}-p_{11}p_{22}} t_n. \quad (2)$$

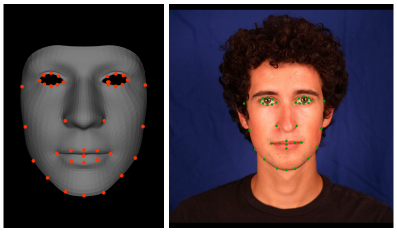
Using equation (2), we obtain the time-labeled sequence of triphone HMM states $s^{(1)}, s^{(2)}, \ldots, s^{(N_S)}$ from the phone-level segmentation. Smooth trajectories of facial motion parameters $\hat{\boldsymbol{\alpha}}_t = \left(\alpha^{(1)}, \ldots, \alpha^{(N_P)}\right)^T$ corresponding to the above sequence of HMM states are generated using the variational spline approach described in [4].

5 Speech-Enabled Avatars

In this section, we use our facial motion synthesis algorithm to build a speech-enabled avatar from a single photograph of a person. We deform the prototype face model to match the shape of a person's face in the photograph and adapt the facial motion using approach presented in Section 3. In order to increase the realism of the appearance, we present a method for automatic synthesis of eye gaze motion as well as blinking.

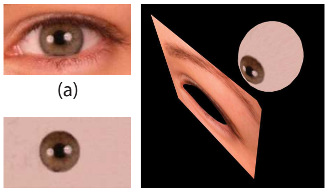
5.1 Fitting the Prototype Face to a Photograph

We start with a photograph of a person looking at the camera with a neutral facial expression. In order to establish correspondence between the generic facial model and subject's face, we manually mark 38 predefined feature points on the photograph, as illustrated in Figure 2. With the help of these correspondences, the prototype face is deformed to fit the geometry of the novel face in the photograph using thin-plate splines [6]. The obtained deformation is subsequently employed to transfer the generic motion model onto the resulting mesh using the approach developed in [3]. The entire procedure of generating a speaker-dependent motion model from feature points takes only a few seconds on a PC.



(a) Prototype face surface (b) A photograph of a subject

Figure 2: A set of manually selected corresponding features on (a) the prototype face model and (b) the novel face photograph.



(a)

(b) (c)

Figure 3: Eye texture synthesis and rendering. (a) Image of the right eye. (b) The complete texture map of the eyeball obtained from (a) by using texture synthesis. (c) The eyeball is placed behind an eyeless image of the face and it is rotated to synthesize eye gaze changes.

5.2 Synthesis of Eye Motion and Blinking

Changes in eye gaze direction can help to make an avatar appear more life-like. Since some regions of the iris and the sclera are obstructed by the eyelids in the input photograph (see Figure 3 (a)), we develop a method for obtaining a complete eyeball model from its partial view in the photograph. We use a sampling-based texture synthesis algorithm [10] to create the missing parts of the cornea and the sclera, as shown in Figure 3 (b). Using the points marked around the eyes, we first extract image regions which contain the eyeballs. Then, the position and shape of the iris are found using generalized Hough transform [13] in order to segment the eye region into the iris and the sclera. Finally, the complete eyeball image is generated by synthesizing the missing texture inside the iris and sclera regions. This eyeball texture is mapped onto a sphere to obtain the complete eyeball model. Each eyeball is placed behind the eyeless face surface, as shown in Figure 3 (c). The eye gaze motion is generated by rotating the eyeballs around their centers. We use a previously proposed stochastic model [14] to generate the eye gaze changes.

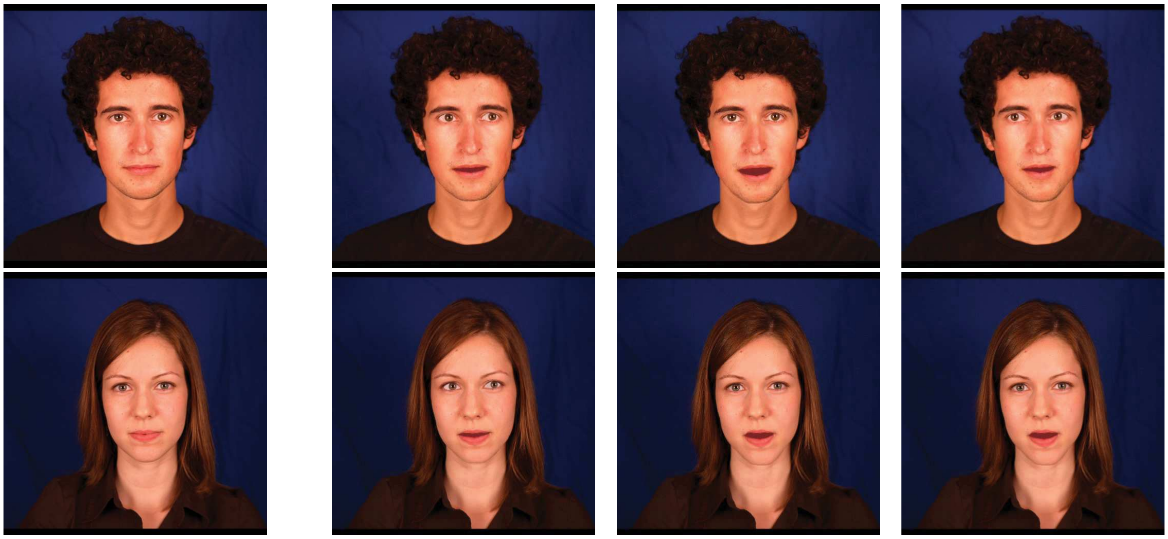
In addition to eye gaze motion, we synthesize eye blinks using the blend shape approach [17]. The eye blink motion of the prototype face model is generated as a linear interpolation between the two key shapes, corresponding to the eyelid in the open and closed positions. The duration of an eye blink was chosen to be 200 ms and the interval between the consecutive blinks is randomly generated with the average interval of 4 seconds. In order to map the eye blink motion to the novel face, the key shapes are deformed using the approach presented in section 5.1.

5.3 Examples of Avatars

We have developed a real-time rendering software which creates face animations of avatars from text input. Our system is compatible with any SAPI 5.1 acoustic text-to-speech synthesis engine. Figure 4 displays a few sample frames from speech-enabled avatars synthesized using the approach presented above.

6 Discussion

In this work, we have developed an end-to-end system for creating speech-enabled avatars from a single photograph. Such avatars are animated from text or speech input with the help of a novel motion synthesis algorithm. We have also developed a method for synthe-



(a) Input photographs

(b) Synthesized images

Figure 4: Photographs of two persons and images of speech-enabled avatars created from these photograph.

sizing eye gaze motion from a photograph. In [4], we demonstrate that our approach can also be used to build volumetric displays that feature speech-enabled 3D avatars. We now discuss the limitation of our work and open problems we plan to address in the future.

The HMM-based facial motion synthesis approach implicitly assumes that the visual and acoustic realizations of phonemes are synchronous. However, there is cognitive evidence that there exists only loose synchronicity between them [12]. For example, facial articulations sometimes precede the sounds they produce. One may expect an improvement in the quality of synthesized facial animations if the visual and acoustic speech are modeled asynchronously by extending the HMM-based approach using, for example, dynamic Bayesian networks.

In order to transform the generic facial motion model and obtain the geometry of a novel face from a photograph, our approach requires a few corresponding points to be established between the prototype face and the novel face. In our current implementation, the corresponding features are marked manually on the input photograph. In the future, we expect that avatars can be created from photographs automatically using face detection algorithms [20].

The realism of the appearance of our speech-enabled avatars is limited by absence of teeth in the facial animations. We plan to address this issue in the future by adding a textured teeth model to an avatar. Another factor that can enhance the perception of speech-enabled avatars is rigid head motion animation. Automatic synthesis of realistic head motions is a challenging problem since head motion is often influenced by prosodic features of speech and will be included in the future implementations of our system.

Acknowledgements

This work was supported by the Columbia University Initiative in Science and Engineering. The authors thank Hila Becker, Blake Shaw and Oliver Cossairt for providing the photographs used to create the avatars shown in the paper.

References

[1] Crazy talk. http://www.reallusion.com/crazytalk.

[2] Motion portrait. http://www.motionportrait.com.

[3] D. Bitouk. *Head-pose and Illumination Invariant 3-D Audio-Visual Speech Recognition*. PhD thesis, The Johns Hopkins University, 2006.

[4] D. Bitouk and S. Nayar. Speech Enabled Avatar from a Single Photograph. Technical report, CUCS-045-07, Department of Computer Science, Columbia University, 2007.

[5] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating Faces in Images and Video. In *Eurographics*, volume 22 (3), pages 641–650, 2003.

[6] F. Bookstein. Principal Warps: Thin Plate Splines and the Decomposition of Deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:567–585, 1989.

[7] M. Brand. Voice Puppetry. In *SIGGRAPH*, pages 21–28, 1999.

[8] C. Bregler, M. Covell, and M. Slaney. Video Rewrite: Driving Visual Speech with Audio. *Computer Graphics*, 31:353–360, 1997.

[9] Y. Cao, P. Faloutsos, E. Kohler, and F. Pighin. Real-time Speech Motion Synthesis from Recorded Motions. In *Eurographics Symposium on Computer Animation*, pages 347–355, 2004.

[10] A. Efros and T. K. Leung. Texture Synthesis by Non-parametric Sampling. In *IEEE International Conference on Computer Vision*, volume 2, pages 1033–1038. 1999.

[11] T. Ezzat, G. Geiger, and T. Poggio. Trainable Videorealistic Speech Animation. *ACM Transactions on Graphics*, 21(3):388–398, 2002.

[12] K. Grant, V. van Wassenhoveb, and D. Poeppelb. Detection of Auditory (Cross–Spectral) and Auditory-Visual (Cross–Modal) Synchrony. *Speech Communication*, 44:43–54, 2004.

[13] C. Kimme, D. Ballard, and J. Slansky. Finding Circles by an Array of Accumulators. *Communications of the ACM*, 18(2):120–122, 1975.

[14] S. P. Lee, J. P. Badler, and N. I. Badler. Eyes Alive. *ACM Transactions on Graphics*, 21(3):637–644, 2002.

[15] T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda. Text-to-Visual Speech Synthesis Based on Parameter Generation from hmm. In *IEEE International Conference of Acoustics, Speech and Signal Processing*, volume 6, pages 3745–3748, 1998.

[16] P. Muller, G. Kalberer, M. Proesmans, and L. V. Gool. Realistic Speech Animation Based on Observed 3D Face Dynamics. *Vision, Image & Signal Processing*, 152:491–500, 2005.

[17] F. I. Parke. Computer Generated Animation of Faces. In *ACM National Conference*, volume 1, pages 451–457, 1972.

[18] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[19] D. Terzopoulos and K. Waters. Analysis and Synthesis of Facial Image Sequences using Physical and Anatomical Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:569–579, 1993.

[20] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Computer Vision and Pattern Recognition*, pages 511–518, 2001.

[21] K. Wampler, D. Sasaki, L. Zhang, and Z. Popovic. Dynamic, Expressive Speech Animation from a Single Mesh. *Eurographics Symposium on Computer Animation*, pages 53–62, 2007.

[22] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book*. Cambridge University Press, 2005.