Belief Hidden Markov Model for Speech Recognition

Siwar Jendoubi* Boutheina Ben Yaghlane† Arnaud Martin‡
*University of Tunis, ISG Tunis, LARODEC Laboratory
jendoubi.siouar@laposte.net
†University of Carthage, IHEC Carthage, LARODEC Laboratory
boutheina.yaghlane@ihec.rnu.tn
‡University of Rennes I, IUT de Lannion, UMR 6074 IRISA
arnaud.martin@univ-rennes1.fr

***Abstract*—Speech Recognition searches to predict the spoken words automatically. These systems are known to be very expensive because of using several pre-recorded hours of speech. Hence, building a model that minimizes the cost of the recognizer will be very interesting. In this paper, we present a new approach for recognizing speech based on belief HMMs instead of probabilistic HMMs. Experiments shows that our belief recognizer is insensitive to the lack of the data and it can be trained using only one exemplary of each acoustic unit and it gives a good recognition rates. Consequently, using the belief HMM recognizer can greatly minimize the cost of these systems.**

***Index Terms*—Speech recognition, HMM, Belief functions, Belief HMM.**

I. Introduction

The automatic speech recognition is a domain of science that attracts the attention of the public. Indeed, who never dreamed of talking with a machine or at least control an apparatus or a computer by voice. The speech processing includes two major disciplines which are the speech recognition and the speech synthesis. The automatic speech recognition allows the machine to understand and process oral information provided by a human. It uses matching techniques to compare a sound wave to a set of samples, compounds generally of words or sub-words. On the other hand, the automatic speech synthesis allows the machine to reproduce the speech sounds of a given text. Nowadays, most speech recognition systems are based on the modelling of speech units known as acoustic unit. Indeed, speech is composed of a sequence of elementary sounds. These sounds put together make up words. Then, from these units we seeks to derive a model (one model per unit), which will be used to recognize continuous speech signal. Hidden Markov Models (HMM) are very often used to recognize these units. HMM based recognizer is a widely used technique that allows as to recognize about 80% of a given speech signal, but this recognition rate still not yet satisfying. Also, this method needs many hours of speech for training which makes the automatic speech recognition task very expensive.

Recently, [7], [6] extend the Hidden Markov Model to the theory of belief functions. The belief HMM will avoid disadvantages of probabilistic HMM which are, generally, due to the use of probability theory. Belief functions are used in several domains of research where incertitude and imprecision dominate. They provide many tools for managing and processing the existent pieces of evidence in order to extract knowledge and make better decision. They allow experts to have a more clear vision about their problems, which is helpful for finding better solutions. What's more, belief functions theories present a more flexible ways to model uncertainty and imprecise data than probability functions. Finally, it offers many tools with a higher ability to combine a great number of pieces of evidence.

Belief HMM gives a better classification rate than the ordinary HMM when they are applied in a classification problem. Consequently, we propose to use the belief HMM in the speech recognition process. Finally, we note that this is the first time where belief functions are used in speech processing.

In the next section we talk about the probabilistic hidden Markov model and we define its three famous problems. In Section three we present the probabilistic HMM recognizer, the acoustic model and the recognition process. The transferable belief model is introduced in section four. In section five we will talk about the belief HMM. In section six, we present our belief HMM recognizer, the belief acoustic model and the belief recognition process. Finally, experiments are presented in section seven.

II. Probabilistic HMM

A Hidden Markov Model is a combination of two stochastic processes; the first one is a Markov chain that is characterized by a finite set[1] Ω_t of non observable N states (hidden) and the transition probabilities, $a_{ij} = \mathrm{P}\left(s_j^{t+1} \mid s_i^t\right), 1 \leq i, j \leq N$, between them. The second stochastic process produces the sequence of T observations which depends on the probability density function of the observation model defined as $b_j\left(O_t\right) = \mathrm{P}\left(O_t \mid s_j^t\right), 1 \leq j \leq N, 1 \leq t \leq T$ [4], in this paper we use a mixture of Gaussian densities. The initial state distribution is defined as $\pi_i = \mathrm{P}\left(s_i^1\right), 1 \leq i \leq N$. Hence, an HMM $\lambda\left(A, B, \Pi\right)$ is characterized by the transition matrix $A = \{a_{ij}\}$, the observation model $B = \{b_j\left(O_t\right)\}$ and the initial state distribution $\Pi = \{\pi_i\}$.

[1] t notes the current instant, it is put in exponent of states for simplicity.

There exist three basic problems of HMMs that must be solved in order to be able to use these models in real world applications. The first problem is named the evaluation problem, it searches to compute the probability $P(O/\lambda)$ that the observation sequence O was generated by the model λ. This probability can be obtained using *the forward propagation* [4]. Recursively, it estimates the forward variable:

$$\alpha_t(i) = \mathrm{P}\left(O_1 O_2 \ldots O_t, q_t = s_i \mid \lambda\right) \quad (1)$$

$$\alpha_t(i) = \left(\sum_{i=1}^{N} \alpha_{t-1}\left(i\right) a_{ij}\right) b_j\left(O_t\right) \quad (2)$$

for all states and at all time instant. Then, $P(O/\lambda) = \sum_{i=1}^{N} \alpha_T\left(i\right)$ is obtained by summing the terminal forward variables. Also, *the backward propagation* can be used to resolve this problem. Unlike forward, the backward propagation goes backward. At each instant, it calculates the backward variable:

$$\beta_t(i) = \mathrm{P}\left(O_{t+1} O_{t+1} \ldots O_T \mid q_t = s_i, \lambda\right) \quad (3)$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j\left(O_{t+1}\right) \beta_{t+1}\left(i\right) \quad (4)$$

finally, $\mathrm{P}\left(O \mid \lambda\right) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$ is obtained by combining the forward and backward variable. The second problem is named the decoding problem. It searches to predict the state sequence S that generated O. The Viterbi [4] algorithm solves this problem. It starts from the first instant, $t = 1$, for each moment t, it calculates $\delta_t(i)$ for every state i, then it keeps the state which have the maximum $\delta_t = \max_{q_1, q_2, \ldots, q_{t-1}} \mathrm{P}\left(q_1, q_2, \ldots q_{t-1}, q_t = i, O_1 O_2 \ldots O_{t-1} \mid \lambda\right) = \max_{1 \leq i \leq N} \left(\delta_{t-1}\left(i\right) a_{ij}\right) b_j\left(O_t\right)$. When, the algorithm reaches the last instance $t = T$, it keeps the state which maximize δ_T. Finally, Viterbi algorithm back-track the sequence of states as the pointer in each moment t indicates. The last problem is the learning problem, it seeks to adjust the model parameters in order to maximize $\mathrm{P}\left(O \mid \lambda\right)$. Baum-Welch [4] method is widely used. This algorithm uses the forward and backward variables to re-estimate the model parameters.

III. Prebabilistic HMM based recognizer

A. Acoustic model

The acoustic model attempts to mimic the human auditory system, it is the model used by the HMM-based speech recognizer in order to transform the speech signal into a sequence of acoustic units, this last will be transformed into phoneme sequence and finally the desired text is generated by converting the phoneme sequence into text. Acoustic models are used by speech segmentation and speech recognition systems.

The acoustic model is composed of a set of HMMs [4], each HMM corresponds to an acoustic unit. To have a good acoustic model some choices have to be done:

a) The acoustic unit: the choice of the acoustic unit is very important, in fact, the number of them will influence the complexity of the model (more large the number, more complex the model). If we choose a small unit like the phone we will have an HMM for every possible phone in the language, the problem with this choice is that the phone do not model its context. Such a model is called ***context independent model***. These models are generally used for speech segmentation systems. Other units that take the context into account can be used as acoustic unit as the diphone which model the transition between two phones, the triphone which model the transition between three phones, subwords, words. These models are called ***context dependent models***. According to [5], when the context is greater, the recognition performance improve.

b) The model: for each acoustic unit we associate an HMM, then types of HMM model and the probability density function of the observation must be chosen. Generally, left-right models are used for speech recognition and speech synthesis systems [4]. In fact, Speech signal has the property that it changes over time, then the choice of the left-right model is justified by the fact that there is no back transitions and all transitions goes forward. The number of states is fixed in advance or chosen experimentally. [2], [3] fixed the number of state to three. This choice is justified by the fact that most phoneme acoustic realization is characterized by three sub-segments, hence we have a state for each sub-segment. [1], [12] used an HMM of six states. Finally, we choose the probability density function of the observation. They are represented by a mixture of Gaussian pdf, the number of mixtures is generally chosen experimentally.

The next step, consists on training parameters of each HMM using a speech corpus that contains many exemplary of each acoustic unit. Speech segments are transformed into sequence of acoustic vectors by the mean of a feature extraction method like MFCC, these acoustic vectors are our sequence of observations.

Then, HMMs are concatenated to each other and we obtain the model that will be used to recognize the new speech signal. The recognizer contains three levels; the first one is the ***syntactic level***. It represents all possible word sequences that can be recognized by our model. The second level is the ***lexical level***. It represents the phonetic transcription (the phoneme sequence) of each word. Finally, the third level is the acoustic level. It models the realization of each acoustic unit (in this case the phone).

B. Speech recognition process

The model described above is used for the speech recognition process. Let S be our speech signal to be recognized. Recognizing S consists on finding the most likely path in the syntactic network. The first step, is to transform S into a sequence of acoustic vectors using the same feature extraction method used for training, then we obtain our sequence of observation O. The most likely path is the path that maximizes the probability of observing O such the model $\mathrm{P}\left(O|\lambda\right)$. This probability can be done either by using the forward algorithm, or the Viterbi algorithm.

IV. Transferable Belief Model

The Transferable Belief Model (TBM) [11], [10] is a well used variant of belief functions theories. It is a more general system than the Bayesian model.

Let $\Omega_t = \{\omega_1, \omega_2, ..., \omega_n\}$ be our frame of discernment, The agent belief on Ω_t is represented by the basic belief assignment (BBA) m^{Ω_t} defined from 2^Ω to $[0,1]$. $m^{\Omega_t}(A)$ is the mass value assigned to the proposition $A \subseteq \Omega_t$ and it must respect: $\sum_{A\subseteq\Omega_t} m^{\Omega_t}(A) = 1$. Also, we can define conditional BBA. Then we can have $m^{\Omega_t}\left[S^{t-1}\right](A)$ which is a BBA defined conditionally to $S^{t-1} \subseteq \Omega_{t-1}$. If we have $m^{\Omega_t}(\emptyset) > 0$, our BBA can be normalized by dividing the other masses by $1 - m^{\Omega_t}(\emptyset)$ then the conflict mass id redistributed and $m^{\Omega_t}(\emptyset) = 0$.

Basic belief assignment can be converted into other functions. They represent the same information under other forms. What's more, they are in one to one correspondence and they are defined from 2^Ω to $[0,1]$. We will use belief bel, plausibility pl and commonality q functions:

$$bel^\Omega(A) = \sum_{\emptyset\neq B\subseteq A} m^\Omega(B), \forall A \subseteq \emptyset,\ A \neq \emptyset \quad (5)$$

$$m^\Omega(A) = \sum_{B\subseteq A} (-1)^{|A|-|B|} bel^\Omega(B), \forall A \subseteq \Omega \quad (6)$$

$$pl^\Omega(A) = \sum_{B\cap A=\emptyset} m^\Omega(B), \forall A \subseteq \Omega \quad (7)$$

$$m^\Omega(A) = \sum_{B\subseteq A} (-1)^{|A|-|B|-1} pl^\Omega\left(\bar{B}\right), \forall A \subseteq \Omega \quad (8)$$

$$q^\Omega(A) = \sum_{B\supseteq A} m^\Omega(B), \forall A \subseteq \Omega \quad (9)$$

$$m^\Omega(A) = \sum_{A\subseteq B} (-1)^{|B|-|A|} q^\Omega(B), \forall A \subseteq \Omega \quad (10)$$

Consider two distinct BBA m_1^Ω and m_2^Ω defined on Ω, we can obtain $m_{1\cap2}^\Omega$ through the TBM conjunctive rule (also called conjunctive rule of combination CRC) [9] as:

$$m_{1\cap2}^\Omega(A) = \sum_{B\cap C=A} m_1^\Omega(B)\, m_2^\Omega(C), \forall A \subseteq \Omega \quad (11)$$

Equivalently, we can calculate the CRC via a more simple expression defined with the commonality function:

$$q_{1\cap2}^\Omega(A) = q_1^\Omega(A)\, q_2^\Omega(A), \forall A \subseteq \Omega \quad (12)$$

V. Belief HMM

Belief HMM is an extension of the probabilistic HMM to belief functions [7], [6], [8]. Like probabilistic HMM, the belief HMM is a combination of two stochastic processes. Hence, a belief HMM is characterized by:

- The credal transition matrix $A = \left\{m_a^{\Omega_t}\left[S_i^{t-1}\right]\left(S_j^t\right)\right\}$ a set of BBA functions defined conditionally to all possible subsets of states S_i^{t-1},
- The observation model $B = \left\{m_b^{\Omega_t}\left[O_t\right]\left(S_j^t\right)\right\}$a set of BBA functions defined conditionally to the set of possible observation O_t,
- The initial state distribution $\Pi = \left\{m_\pi^{\Omega_1}\left(S_i^{\Omega_1}\right)\right\}$.

The three basic problem of HMM and their solutions are extended to belief functions. As we know the forward algorithm resolves the evaluation problem in the probabilistic case. [7] introduced the ***credal forward algorithm*** in order to resolve this problem in the evidential case. It needs as inputs $m_a^{\Omega_t}\left[S_i^{t-1}\right]\left(S_j^t\right)$ and $m_b^{\Omega_t}\left[O_t\right]\left(S_j^t\right)$ to calculate the forward commonality:

$$q_\alpha^{\Omega_{t+1}}\left(S_j^{t+1}\right) = \left(\sum_{S_i^t\subseteq\Omega_t} m_\alpha^{\Omega_t}\left(S_i^t\right).q_a^{\Omega_{t+1}}[S_i^t]\left(S_j^{t+1}\right)\right) \cap q_b^{\Omega_{t+1}}\left[O_t\right]\left(S_j^t+1\right) \quad (13)$$

This last is calculated recursively from $t = 1$ to T. [6] exploits the conflict of the forward BBA (obtained by using formula 10) to define an evaluation metric that can be used for classification to choose the model that best fits the observation sequence or it can also be used to evaluate the model. Then, given a model λ and an observation sequence of length T, the conflict metric is defined by:

$$L_c(\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log\left(1 - m_\alpha^{\Omega_{t+1}}[\lambda](\emptyset)\right) \quad (14)$$

$$\lambda_* = \arg\max_\lambda L_c(\lambda) \quad (15)$$

A ***credal backward algorithm*** is also defined, recursively, it calculates the backward commonality from T to $t = 1$. More details can be found in [7], [6]. For the decoding problem, many solutions are proposed to extend the Viterbi algorithm to the TBM [7], [6], [8]. All of them search to maximize the state sequence plausibility. According to the definition given in [8], the plausibility of a sequence of singleton states $S = \left\{s^1, s^2, \ldots, s^T\right\}$, $s^t \in \Omega_t$ is given by:

$$pl_\delta(S) = pl_\pi\left(s^1\right).\prod_{t=2}^{T} pl_a^{\Omega_t}\left[s^{t-1}\right]\left(s^t\right).\prod_{t=1}^{T} pl_b\left(s^t\right) \quad (16)$$

Hence, we can choose the best state sequence by maximizing this plausibility. For the learning problem, [6], [8] have proposed some solutions to estimate model parameters, we will talk about the method used in this paper. The first step consists on estimating the mixture of Gaussian models (GMM) parameters using Expectation-Maximization (EM) algorithm. For each state we estimate one GMM. These models are used to calculate $m_b^{\Omega_t}\left[O_t\right]\left(S_j^t\right)$. [6] proposes to estimate the credal transition matrix independently from the transitions themselves. He uses the observation BBAs as:

$$m_{\overline{a}}^{\Omega_t\times\Omega_{t+1}} \propto \frac{1}{T-1} * \sum_{t=1}^{T}\left(m_b^{\Omega_t}\left[O_t\right]^{\uparrow\Omega_t\times\Omega_{t+1}} \cap m_b^{\Omega_{t+1}}\left[O_{t+1}\right]^{\uparrow\Omega_t\times\Omega_{t+1}}\right) \quad (17)$$

where $m_b^{\Omega_t}\left[O_t\right]^{\uparrow\Omega_t\times\Omega_{t+1}}$ and $m_b^{\Omega_{t+1}}\left[O_{t+1}\right]^{\uparrow\Omega_t\times\Omega_{t+1}}$ are computed using the vacuous extension operator [9] of the BBA $m_b^{\Omega_t}\left[O_t\right]\left(S_j^t\right)$ on the cartesian product space as:

$$m_b^{\Omega_t\uparrow\Omega_t\times\Omega_{t+1}}(A) = \begin{cases} m_b^{\Omega_t}(B) & \text{if } A = B\times\Omega_{t+1} \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

This estimation formula is used by [8] as an initialization for ***ITS (Iterative Transition Specialization)*** algorithm. ITS is an iterative algorithm that uses the credal forward algorithm to improve the estimation results of the credal transition matrix. It stops when the conflict metric (formula 14) converged.

VI. Belief HMM based recognizer

Our goal is to create a speech recognizer using the belief HMM instead of the probabilistic HMM. HMM recognizer uses an acoustic model to recognize the content of the speech signal. Then, we seek to mimic this model in order to create a belief HMM based one. We should note that existent parameter estimation methods presented for the belief HMM cannot be used to estimate model parameters using multiple observation sequences. This fact should be taken into account when we design our belief acoustic model.

A. Belief acoustic model

In the probabilistic case, we use an HMM for each acoustic unit, its parameters are trained using multiple speech realization of the unit [5], [1], [2], [12], [3]. In the credal case, a similar model cannot be used. Hence, we present an alternate method that takes this fact into account.

Let K be the number of the speech realization of a given acoustic unit. These speech realization are transformed into MFCC feature vectors. Hence, we obtain K observation sequences. Our training set will be:$O = \left[O^1, O^2, \ldots, O^K\right]$ where $O^k = \left(O_1^k, O_2^k, \ldots, O_{T_k}^k\right)$ is the k^{th} observation sequence of length T_k. These observations are supposed to be independent to each other. So instead of training one model for all observation set O, we propose to create a belief model for each observation sequence O^k. These K models will be used to represent the given acoustic unit in the recognition process.

Like the acoustic model based on the probabilistic HMM, we have to make some choices in order to have a good belief acoustic model. In the first place, we choose the acoustic unit. The same choices of the probabilistic case can be adopted for the belief case. In the second place, we choose the model. We should note that we cannot choose the topology of the belief HMM, this is due to the estimation process of the credal transition matrix. In other words, the resultant credal observation model is used to estimate the credal transition matrix which does not give as the hand to choose the topology of our resultant model. Consequently, choosing the model in the credal case consists on choosing the number of states and the number of Gaussian mixtures. In our case we fix the number of states to three and we choose the number of Gaussian mixtures experimentally.

B. Speech recognition process

The belief acoustic model is used in the speech recognition process. Now, we explain how the resultant model will be used for recognizing speech signal.

Let S be our speech signal to be recognized. Recognizing S consists on finding the most likely set of models. The first step, is to transform S into a sequence of acoustic vectors using the same feature extraction method used for training, then we obtain our sequence of observation O. This last is used as input for all models. The credal forward algorithm is then applied, each model gives us an output which is the value of the conflict metric. An acoustic unit is presented by a set of models, every model gives a value for the conflict metric. Then we calculate the arithmetic mean of the resultant values. Finally, we choose the set of models that optimizes the average of the conflict metric instead of optimizing the conflict metric, as proposed by [6], using formula 15.

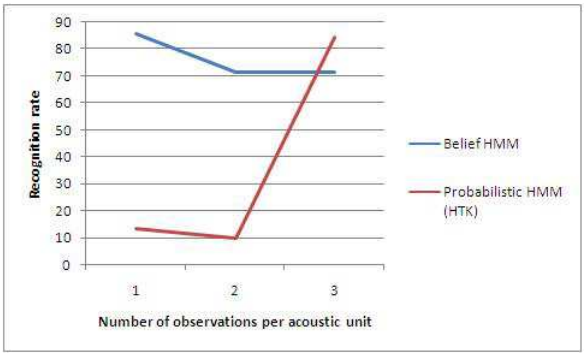

Figure 1. Influence of the number of observations on the recognition rate

VII. Experiments

In this section we present experiments in order to validate our approach. We compare our belief HMM recognizer to a similar one implemented using the probabilistic HMM.

We use MFCC (Mel Frequency Cepstral Coefficient) as feature vectors. Also, we use a three state HMM and two Gaussian mixtures. Finally, to evaluate our models we calculate the percent of correctly recognized acoustic units (number of correctly recognized acoustic unit / total number of acoustic units). We use a speech corpus that contains speech realization of seven different acoustic units and we have fifteen exemplary of each one. Results are shown in figure 1.

The lack of data for training the probabilistic HMM leads to a very poor learning and the resultant acoustic model cannot be efficient. Then using a training set that contains only one exemplary of each acoustic unit leads to have a bad probabilistic recognizer. In this case our belief HMM based recognizer gives a recognition rate equal to 85.71% against 13.79% for the probabilistic HMM which is trained using HTK [13]. This results shows that the belief HMM recognizer is insensitive to the lack of data and we can obtain a good belief acoustic model using only one observation for each unit. In fact, the belief HMM models knowledge by taking into account doubt, imprecision and conflict which leads to a discriminative model in the case of the lack of data.

HTK is a toolkit for HMMs and it is optimized for the HMM speech recognition process. It is known to be powerful under the condition of having many exemplary of each acoustic unit. Hence, it needs to use several hours of speech for training. Having a good speech corpus is very expensive which influence the cost of the recognition system. Then, the speech recognition systems are very expensive. Consequently, using the belief HMM recognizer can greatly minimize the cost of these systems.

VIII. Conclusion

In this paper, we proposed the Belief HMM recognizer. We showed that incorporating belief functions theory in the

speech recognition process is very beneficial, in fact, it reduces considerably the cost of the speech recognition system. Future works will be focuced on the case of the noisy speech signal. Indeed, existent speech recognizer still not yet good if we have a noisy signal to be decoded.

REFERENCES

[1] F. Brugnara, D. Falavigna, and M. Omologo. Automatic segmentation and labeling of speech based on hiddenmarkov models. *Speech Communication*, 12:370–375, 1993.

[2] P. Carvalho, L. C. Oliveira, I. M. Trancoso, and M. C. Viana. Concatenative speech synthesis for european portuguese. In *3rd ESCA/COCOSDA Worshop on Speech Synthesis*, pages 159–163, 1998.

[3] S. Cox, R. Brady, and P. Jackson. Techniques for accurate automatic annotation of speech waveforms. In *Proc. ICASSP*, 1998.

[4] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77:257–286, 1989.

[5] L. Rabiner and B. H. Juang. Fundamentals of speech recognition. *Prentice-Hall, Inc. Upper Saddle River, NJ, USA*, 1993.

[6] E. Ramasso. Contribution of belief functions to HMM with an application to fault diagnosis. *IEEE International Workshop on Machine Learning and Signal Processing, Grenoble, France*, September 2-4 2009.

[7] E. Ramasso, M. Rombaut, and D. Pellerin. Forward-backward-viterbi procedures in thetransferable belief model for state sequence analysis using belief functions. *ECSQARU, Hammamet: Tunisie*, pages 405–417, 2007.

[8] L. Serir, E. Ramasso, and N. Zerhouni. Time-sliced temporal evidential networks: the case of evidential hmm with application to dynamical system analysis. *IEEE International Conference on Prognostics and Health Management. Denver, Colorado, USA*, 2011.

[9] P. Smets. Beliefs functions: The disjunctive rule of combination and the generalized bayesian theorem. *IJAR*, 9:1–35, 1993.

[10] P. Smets. Belief functions and the transferable belief model. *Available on www. sipta.org/documentation/belief/belief.ps*, 2000.

[11] P. Smets and R. Kennes. The transferable belief model. artificial intelligence. 66(2):191–234, 1994.

[12] D. T. Toledano, L. A. H. Gomez, and L. V. Grande. Automatic phonetic segmentation. *IEEE Trans. Speech, Audio Processing*, 11(6):617–625, 2003.

[13] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. The htk book for htk version 3.4. *Microsoft Corporation and Cambridge University Engineering Department*, 2006.