Dialogue Management for Multimodal User Registration

Fei Huang, Jie Yang, Alex Waibel

Interactive System Lab, Carnegie Mellon University

Email: {fhuang, yang+, ahw} @cs.cmu.edu

ABSTRACT

User registration refers to associating certain personal information with a user. It is widely used in hospitals, hotels and conferences. In this paper, we propose an approach to interactive user registration by combining face recognition, speech recognition and speech synthesis technologies together through an efficient dialogue manager. In order to minimize a user's effort, we employ a new dialogue management model based on a finite state automaton (FSA), which uses a Baysian network to fuse the user's information from multiple channels (e.g., face image, speech, records stored in a pre-constructed database) to reliably estimate the confidence about user identity. Instead of fixing weights, the FSA adjusts its weights dynamically by integrating partial information from multiple information sources. This is achieved by maximizing an objective function to determine an optimal action at each succeeding state according to current confidence and information cues. Thus the transition between states can be done along the shortest path from the initial state to the goal state. We have developed a multimodal user registration system to demonstrate the feasibility of the proposed approach.

1. INTRODUCTION

User registration refers to associating certain personal information, such as name, email address, and phone number, with a user. Many applications require user registration. For example, patients have to register before any treatment in a hospital. Participants have to register before a conference. In the Interactive System Lab of Carnegie Mellon University, we are working on the Meeting Room of Future, in which user registration is essential for many systems. First, participants' identities play an important role in a multimedia meeting recorder, e.g., to know who said what. Second, knowing who is in the meeting in advance is helpful for enhancing speech recognition by taking advantage of speaker-dependency. Finally, some personal information, like phone number and email account, is useful for dissemination and follow-up of the meeting, such as further discussion, confirmation and communication among participants. However, user registration is a tedious and time-consuming task, especially when it has to be done again and again in different meetings. It is desirable that a user can register automatically or interactively with minimal efforts.

In this paper, we propose an approach to interactive user registration by combining face recognition, speech recognition and speech synthesis technologies together through an efficient dialogue manager. User registration is to keep updated user's information in a database. The database can be pre-constructed using information retrieval techniques. For example, the information of a user can be obtained from a pre-existing database, a personal home page, and a department directory. The database, however, can be incomplete and/or out-of-date. We need to complete, verify and update user's information in real-time. The task requires that we identify a user, and check if information in the database is complete and updated. In order to minimize a user's effort, we employ a new dialogue management model based on a finite state automaton (FSA), which uses a Baysian network to fuse the user's information from multiple channels (e.g., face image, speech, records stored in a pre-constructed database) to reliably estimate the confidence about user identity. Instead of fixing weights, the FSA adjusts its weights dynamically by integrating partial information from multiple information sources. This is achieved by maximizing an objective function to determine an optimal action at each succeeding state according to current confidence and information cues. Thus the transition between states can be done along the shortest path from the initial state to the goal state. We have developed a multimodal user registration system to demonstrate the feasibility of the proposed approach. The system consists of a face recognition module, a speech recognition module, a dialogue management module, and a speech synthesis module. Once a user appears, the face recognition module tries to identify the user. The speech synthesis module and speech recognition module interact with the user. The dialogue module controls the whole user registration process.

The organization of this paper is as follows: in Section 2, we describe the proposed dialogue manager model; in Section 3, we discuss the system's architecture, giving detailed information about different modules; in Section 4, we present some experiment results, including the system interface and a registration sample. In Section 5, we conclude the paper.

2. FSA-BASED DIALOGUE MANAGER

The dialogue management module plays a key role in this research. The dialogue management strategy leads to achieving the goal at the minimum cost of user effort. Much work has been done in the development of efficient dialogue managers for human computer interaction. Heeman et al. advocated "factoring out the grounding behavior" from structured dialogue model [1]. Denecke and Waibel proposed the generation of clarification questions with domain modeling and underspecified representations in order to arrive at a dialogue goal along an optimal sequence of questions [2]. Papineni et al. proposed a free-flow dialogue management model based on a form, which correspond to a specific task in the domain; the dialogue

manager is mainly responsible for choosing the appropriate "form" which matches user's goal best [3]. Ehrlich structures complex dialogs into sub-dialogs and thus reduces the dialogue's complexity at each state without losing its flexibility [4]. Recently a more promising trend is to formalize the dialogue management as an optimization problem ([5][6][7]). With some assumptions about the state transition probabilities and cost assignment, the dialogue system can be processed as a MDP (Markov Decision Process), and the supervised and reinforcement learning algorithms are applied to learn the optimal strategy.

Since the "optimal strategy" is a mapping from a state to an action, i.e. a policy deciding which action should be taken in every possible state, once it is learned, it is fixed and deterministic. For the task of multimodal information integration, such pre-determined strategy may not be appropriate, because information cues from multiple modalities are dynamically available, and switches between different modalities are quite frequent. When and which modality should be switched must be decided according to current available information sources.

Johnston et al. proposed a multimodal integration theme based on unification over typed feature structure, which determine the consistency of two pieces of partial information, and combine them into a single result if they are consistent [8]. However, for multimodal user registration, the consistency between multiple pieces of partial information is already known (they are all from the same user). We are more concerned with the confidence of user identity given the information.

In order to solve this problem, we designed a dialogue manager based on a finite state automaton. Similar to the MDP model, we define the states, transitions and action set. Unlike a traditional MDP model, weights of the FSA model are not fixed. We use a Bayesian network to determine the confidence about user identity by fusing current information cues from multiple channels (e.g., face image, spoken language input and database). Multimodal information cues are integrated incrementally. The weights are adapted based on an evaluation function, which indicates the confidence score, completeness of available information and human-computer interaction cost at current state. By maximizing this function during each dialogue turn, the optimal strategy is determined online rather than learned in advance, and the shortest path from initial state to goal state can be dynamically determined with a minimum of user effort.

2.1 Definition of Finite State Automaton

In the dialogue model, we define a frame containing 4 slots (First_name: Last_name: Phone_number: Email_account:) as the format of the required information. The registration can be regarded as a slot-filling process. Frames filled with different information represent different states in the FSA, as in Table 1. The structure of the FSA is shown in Figure 1.

In each user registration process, state 1, 3 (for user identity verification), 6 (for user information verification) and 7 are definitely visited; other states may be visited or skipped depending on the confidence of user identity or the user's information pre-stored in the database.

Table1: Definition of different states in FSA.

	First-name	**Last-name**	**Phone-number**	**Email-account**
State 1	Empty	Empty	Empty	Empty
State 2	Filled	Empty	Empty	Empty
State 3	Filled	Filled	Empty	Empty
State 4	Filled	Filled	Filled	Empty
State 5	Filled	Filled	Empty	Filled
State 6	Filled	Filled	Filled	Filled
State 7	Filled & Verified	Filled & Verified	Filled & Verified	Filled & Verified

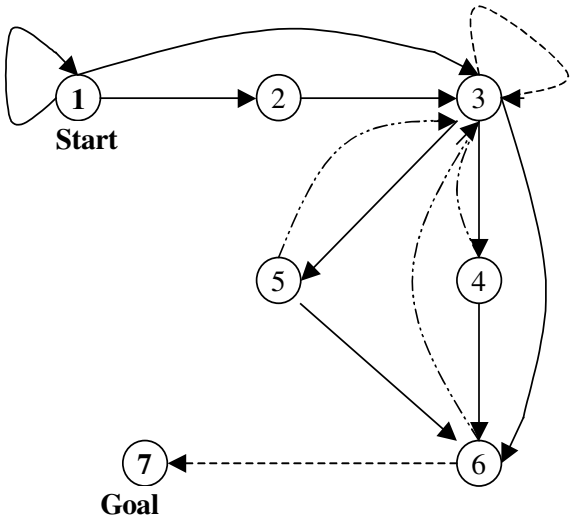


Figure 1: Finite state automaton for the dialogue model.

In the above FSA model, state 1 is the initial state (empty frame), state 7 is the goal state (verified and completed frame).

The transition between states is stimulated by 4 kinds of actions:

- human-computer interaction (e.g., face recognition, dialogue turns);
- user database retrieval;
- change of confidence about user identity by partial information fusion through Bayesian network;
- verification or update of user information.

To be specific, the state transition-action pairs are in Table 2.

Table2: State transitions and corresponding actions.

State transition	Action to be taken
$3 \rightarrow 3$	User identity verification
$3 \rightarrow 7$	User information verification or update
$4 \rightarrow 3$ $5 \rightarrow 3$ $6 \rightarrow 3$	Change of confidence about user identity
All other transitions	Human-computer interaction: • Face recognition • Ask for first name • Ask for last name • Ask for phone number • Ask for email account Database retrieval

2.2 Partial Information Fusion Using Bayesian Network

For user registration, user identity should be reliably determined as soon as possible, so that the user's information pre-stored in database can be fully utilized, and the deviations in a dialogue session can be reduced as much as possible. However, asking for user's name directly via spoken language may not be a good solution. First, many names have similar pronunciations, like "Martin", "Marrin" and "Marvin". Second, many foreign names have their unique foreign pronunciations, and their number is increasing dynamically. As such, for most speech recognizers, it is not easy to reliably identify a user's name. Besides, for some users whose face image is available, user identity can be determined correctly only by face recognition, thus no bother with name recognition via speech.

To reliably estimate the confidence of user identity, a Bayesian network (as in Figure 2) is used to fuse multiple information evidences gained directly from the user (the outermost nodes in the network), and those "features" pre-stored in the database (the nodes in the middle level,):

$$P(id \mid Evidence) = \frac{P(id)P(Evidence \mid id)}{P(Evidence)} = \frac{P(id)\prod_{e \in Evidence} P(e \mid id)}{P(Evidence)}$$
$$= \frac{P(id)\prod_{e \in Evidence,\ f_{id} \in user\ id's\ Featur} P(e \mid f_{id})P(f_{id} \mid id)}{P(Evidenec)} \quad (1)$$

$$because \quad \forall f_{id} \in user\ id's\ Feature, \quad P(f_{id} \mid id) = 1$$
$$\arg\max_{id} P(id \mid Evidence)$$
$$= \arg\max_{id} [P(id) \prod_{e \in Evidence,\ f_{id} \in user\ id's\ Feature} P(e \mid f_{id})] \quad (2)$$

Therefore, the confidence measure can be defined as:

$$Conf = C_0 + \sum_{Feature} C_i P_i \quad (3)$$

where P_i=-log$P(e|f_{id})$ represents the confidence on user's identity given the evidence, C_0=-log$P(id)$ represents the prior probability of user *id*, and C_i is the confidence of the corresponding module.

The ***Evidence*** set consists of currently available information cues, such as a face image taken by computer on the spot, utterance for first name, last name and phone number, etc.; the ***Feature*** set consists of the features corresponding to the evidences ***e***, like the user's actual face image, name and phone number which are pre-stored in the database (some of them may be empty if the records do not exist in database). If f_{id} does not exist in the database, we artificially define $P(e|\ f_{id})$ to be a certain constant.

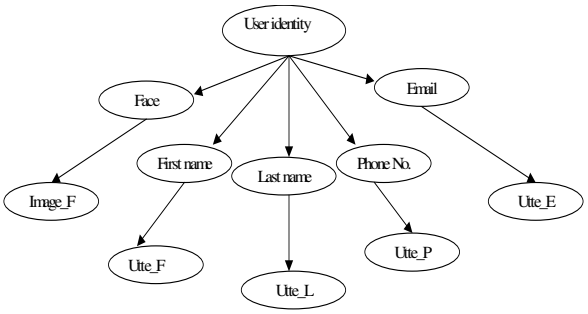


Figure 2: The Bayesian network represents multiple information cues. Image_F represents the face image taken by the camera. Utte_X represents the utterance conveying first name, last name, phone number and email account, respectively.

Furthermore, to minimize the user's effort in registration, it is not necessary to determine user identity only when all information cues are available. As user information cues can be acquired more and more when dialogue transits from state 1(where only face image is available) to state 6 (where all information is available), the fusion can be implemented incrementally as follows:

*Initialization: **Evidence** set is NULL;*

*Take face recognition, and **Evidence** set is increased by face image;*

*Compute P_{conf} = P(id\| **Evidence**)*

While P_{conf} is less than a certain threshold

*1.Ask more information, **Evidence** set is increased by one more information cues;*

*2.Compute P_{conf} = P(id\| **Evidence**)*

User identity is decided as the one who has the largest confidence.

2.3 Online Determination of the Optimal Strategy

Since user information resources are available dynamically, an optimal strategy can only be decided according to current available cues. To make the decision in real-time, we design an evaluation function, which measures the weighted sum of user ***k'***s success at state ***s***, in the way of arriving goal,

$$W(s,k) = W_{conf} + W_{fs} - W_{hci} \quad (4)$$

where

$$W_{conf} = w_1 \bullet p(id \mid evidence);$$
$$W_{fs} = w_2 \bullet m,$$
$$W_{hci} = w_3 \bullet n,$$

m is the number of filled slots at state ***s***, 0 for state 1 and 4 for state 7; ***n*** is the number of human-computer interaction turns; W_{conf} indicates the confidence on user identity based on current evidences at state ***s***, which is 1 after user identity verification; W_{fs} indicates to what extent the dialogue goal is satisfied, i.e. how many slots have been filled (To make sure user identity is confirmed as early as possible, this weight is effective only when the user identity has been confirmed); W_{hci} indicates the cost of interaction, including the cost of potential errors occurred in face and speech recognition.

We define

$$W_{top}(s) = \max_k W(s,k) \quad (5)$$

$$k'(s) = \arg\max_k W(s,k) \quad (6)$$

$W_{top}(s)$ should increase in the state transition process, and reach the maximum value in the goal state, where user ***k'*** is identified as the correct person. Thus the optimal action should be chosen such that $W_{top}(s)$ could be maximized at each succeeding state. The selection of the optimal actions, which can be asking appropriate questions (name, phone number, or verification etc.) or retrieving information from a database, is as following:

*State **s** = initial state, i.e. state 1;*

*While **s** is not goal state (state 7)*

*1. compute **W (s,k)** For each user **k**;*

*2. compute $W_{top}(s)$and top1 candidate user **k'**;*

*3. choose action **a** which can maximize $W_{top}(s)$;*

*4. **s** = **s'**, which is the next state of **s** in FSA, after the action **a** is taken.*

*when **s** = goal state, **k'** is the correct user.*

Because available information will increase as a sequence of actions is taken, *p(id| evidence)* will change for different user ids, and the top candidate ***k'*** will also change accordingly. However, the action ***a*** is chosen to just maximize the current top candidate's ***W*** value, $W_{top}(s)$, it may not be the best action choice for the last ***k'***, the correct user. This is the inherent problem of such a "greedy" algorithm.

3. SYSTEM ARCHITECTURE

We have developed a multimodal user registration system, which has integrated a face recognition module, a speaker-independent large vocabulary speech recognition module, a text-to-speech synthesis module, and a dialogue management module.

The system architecture is shown in Figure 3:

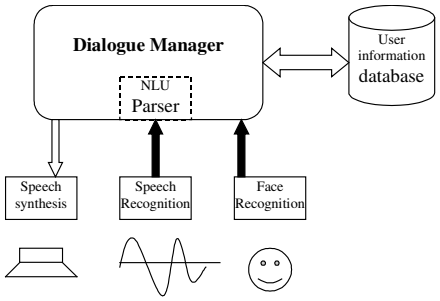


Figure 3: System architecture: communications between modules are through sockets.

In this section, we will discuss each module in more detail.

3.1 Face Recognition Module

The Face recognition module consists of two sub-systems: a real-time face tracker and a face recognition system. Two sub-systems are connected through a socket.

Locating and tracking human faces is a prerequisite for face recognition. By combining the adaptive skin color model with the motion model and the camera model, we have developed a real-time face tracker [9]. The system has achieved a rate of 30+ frames/second on both Unix and PC platforms. The system can track a person's face while the person walks, jumps, sits and rises.

The face recognition system has implemented an "eigenface" algorithm and a dynamic space warping (DSW) algorithm [10]. The techniques based on Principal Components Analysis (PCA), namely "eigenfaces" [11], have demonstrated excellent performance. In the eigenface approach, a face image defines a point in a high dimensional space. Different face images share a number of similarities with each other, so that the points representing these images are not randomly distributed in the image space. They all fall into a lower dimensional subspace. The key idea of the recognition process is to map the face images into an appropriately chosen subspace and perform classification by distance computation. Instead of transforming a face image into one point in the eigenspace, the DSW algorithm breaks down a face image into sub-images using a moving window. When the square window covers the whole image by moving half of the window size each time, we get a sequence of

sub-images. Each sub-image can be transformed to a point in the eigen-space. We then get a set of eigen-points for each face image. During the recognition process, the template set of points is compared to the unknown set of points. The DSW algorithm has better performance than the eigenface algorithm but is slower.

3.2. Speech Recognition Module

The speech recognition module is implemented with the Java Speech API, under which XCalibur, a spoken language R&D Tool Kit developed at Interactive Systems Inc. [12], acts as the core engine.

This Tool Kit supports LVCSR and FSM-based (finite state machine) speech recognition with a very high accuracy rate. Since the sentences most often used in user registration process share some common patterns, we can write some grammars representing such patterns to facilitate the recognition of these sentences. These sentences can be categorized into 4 classes: greeting, update requirement, information presentation and confirmation. Information presentation sentences can be further categorized into: name presentation, phone number presentation and email presentation. We write a grammar for each category, and the accuracy of recognition under such grammars can be over 95%.

3.3. Speech Synthesis Module

We use Festival as the speech synthesis module. Festival is a general multi-lingual speech synthesis system developed at Center for Speech Technology Research, University of Edinburgh. It provides a full TTS (text-to-speech) system with various APIs, and an environment for development and research of speech synthesis techniques. For more detailed information, see [13].

However, synthesizing the pronunciations of many foreign names, which have different pronunciation rules, is a rather tough task. An alternative solution is to record the user's pronunciation when he/she answers the question "What is your first/last name?". This is a part of our future work.

3.4. Dialogue Manager Module

The dialogue manager module gets the user identity hypothesis from the face recognition module, then looks up related information in the pre-constructed information database, makes judgment about the user identity based on the confidence of different information cues (e.g., face, name, etc.), acquires and/or confirms user's personal information via speech recognition module and text-to-speech synthesis module, and finally updates the database if some information is added, deleted or changed.

Since the speech module can achieve very high accuracy for recognizing spoken sentences containing fixed patterns, providing that the grammars representing these patterns are given, the language-understanding task is quite easy. The NLU (natural language understanding) module is embedded in the dialogue manager, based on keyword spotting and parsing. The recognized speech is parsed so that user's information can be extracted with correct format.

3.5. Database Construction

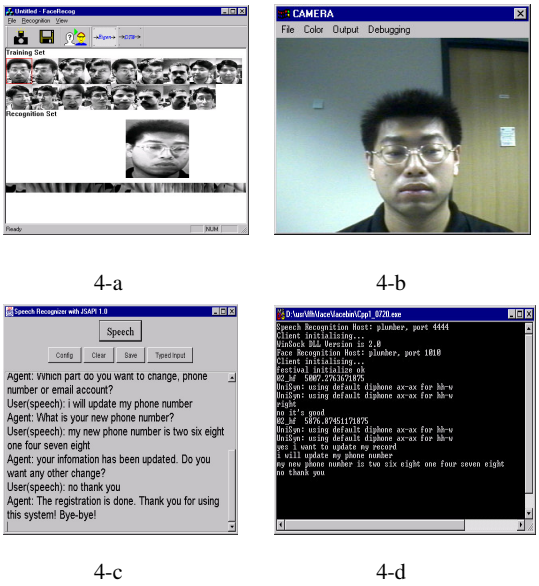
The user information database is automatically constructed by using information retrieval technique to search the web pages, People directory of the School of Computer Science in the Carnegie Mellon University, which contains 1744 users' names, phone numbers and email accounts (The latter two cues may be missed for different users).

4. EXPERIMENTAL RESULTS

4.1. User Interface for the System

Communications between different modules are through Socket. The interfaces for each module are as in Figure 4(a~d).

Figure 4-a is for face recognition module. The upper images are training images (pre-stored in the database, here we just show a few of them.) The lower one is the user's face image taken on the spot, which is to be recognized. Training face images are sorted according to their similarities to the test, the most upper left person has the highest similarity, and is identified as the right person. Figure 4-b is for face tracking system, which locate a face by its eyes and nostrils. Figure 4-c is for speech recognition. Text for recognized and synthesized speech is displayed, with the speaker (user/agent), to facilitate the communication. Figure 4-d is for dialogue manager, which displays the user identity hypothesis and the confidence score.



4-a 4-b

4-c 4-d

Figure 4: user interfaces for each module.

4.2. A Sample Dialogue

Below is a sample dialogue:

Agent: Welcome to the Multimodal User Registration System!

(System takes Face Recognition)

Agent: Hi, you are Fei Huang, right?

User: Yes, I am.

(System retrieves database)

Agent: Hi, Fei, your phone number is two six eight five two nine nine, your email account is fhuang@cs.cmu.edu. Do you want to update your personal information record?

User: Yes.

Agent: Which part do you want to change, phone number or email account?

User: I will update my phone number.

Agent: What is your new phone number?

User: my new phone number is two six eight one four seven eight.

Agent: You information has been updated. Do you want any other change?

User: No. Thank you.

Agent: The registration is done, Thank you for using this system. Bye-bye!

5. CONCLUSION

We proposed an approach to interactive user registration by combining multimodal information cues through an efficient dialogue manager. The dialogue manager employs a FSA-based dialogue model, uses a Bayesian network to incrementally fuse information from multiple channels and changes weights dynamically by maximizing an objective function in order to find the shortest path from an initial state to a goal state and minimize user's effort in registration. A preliminary multimodal user registration system has been developed to demonstrate the feasibility of the proposed approach.

6. ACKNOWLEDGEMENT

The authors would like to thank Robert Malkin, William Kunz, Matthias Denecke and other members in the Interactive Systems Labs for their inspiring discussions and support. This research is partly supported by Defense Advanced Research Projects Agency under contract number DAAD17-99-C-0061.

7. REFERENCES

[1] P. A. Heeman, M. Johnston, J. Denney and E. Kaiser, "Beyond Structured Dialogues: Factoring Out Grounding", In Proceedings of the International Conference on Spoken Language Processing (ICSLP-98), pp933-936, Sydney, Australia, November 1998.

[2] M. Denecke and A. Waibel, "Dialogue Strategies Guiding Users To Their Communicative Goals", In Proceedings of EUROSPEECH97, Vol. 3, pp1339-1342, Rhodes, Greece, September 1997.

[3] K. A. Papineni, S. Roukos, and R. T. Ward. "Free-flow Dialog Management Using Forms", In Proceedings of EUROSPEECH99, Vol. 3, pp1411-1414, Budapest, Hungary, September 1999.

[4] U. Ehrlich, "Task Hierarchies Representing Sub-Dialogs in Speech Dialog Systems", In Proceedings of EUROSPEECH99, Vol. 3, pp1387-1390, Budapest, Hungary, September 1999.

[5] E. Levin, R. Pieraccini, and W. Eckert, "A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies", In IEEE Trans. on Speech and Audio Processing, Vol. 8, No. 1, pp11-23, January, 2000.

[6] E. Levin, R. Pieraccini, and W. Eckert, "Using Markov Decision Process for Learning Dialogue Strategies", In Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP-98), Vol. 1, pp201-204, Seattle, U.S., May 1998.

[7] S. Singh, M. S. Kearns, D. J. Litman, and M. A. Walker, "Reinforcement Learning for Spoken Dialogue Systems", In Proceedings of Neural Information Processing System (NIPS-99), Denver, U.S., November 1999.

[8] M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith, "Unification-based multimodal integration", In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997, Association for Computational Linguistics Press: March, 1997.

[9] J. Yang and A. Waibel, "A Real-time Face Tracker", In Proceedings of Third IEEE Workshop on Applications of Computer Vision (WACV-96), pages 142-147, Sarasota, Florida, USA, December 1996.

[10] R. Gross, J. Yang, and A. Waibel, "Face Recognition in a Meeting Room", Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'2000), Grenoble, France, March, 2000.

[11] M.A. Turk and A. Pentland, "Face Recognition Using Eigenfaces", In Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pages 586-591, Hawaii, USA, 1991.

[12] M. Finke, J. Fritsch, D. Koll and A. Waibel, "Modeling and Efficient Decoding of Large Vocabulary Conversational Speech", In Proceedings of the EUROSPEECH99, Vol. 1, pp467-470, Budapest, Hungary, September 1999.

[13] http://www.cstr.ed.ac.uk/projects/festival/.