Time-Offset Interaction with a Holocaust Survivor

Ron Artstein[1] David Traum[1] Oleg Alexander[1] Anton Leuski[1] Andrew Jones[1] Kallirroi Georgila[1]
Paul Debevec[1] William Swartout[1] Heather Maio[2] Stephen Smith[3]
[1]USC Institute for Creative Technologies, 12015 Waterfront Drive, Playa Vista CA 90094-2536, USA
[2]Conscience Display, 1023 Fifth Street, Coronado CA 92118, USA
[3]USC Shoah Foundation, 650 West 35th Street, Suite 114, Los Angeles CA 90089-2571, USA
{artstein|traum|oalexander|leuski|jones|kgeorgila|debevec|swartout}@ict.usc.edu
consciencedisplay@gmail.com smithsd@dornsife.usc.edu

ABSTRACT

Time-offset interaction is a new technology that allows for two-way communication with a person who is not available for conversation in real time: a large set of statements are prepared in advance, and users access these statements through natural conversation that mimics face-to-face interaction. Conversational reactions to user questions are retrieved through a statistical classifier, using technology that is similar to previous interactive systems with synthetic characters; however, all of the retrieved utterances are genuine statements by a real person. Recordings of answers, listening and idle behaviors, and blending techniques are used to create a persistent visual image of the person throughout the interaction. A proof-of-concept has been implemented using the likeness of Pinchas Gutter, a Holocaust survivor, enabling short conversations about his family, his religious views, and resistance. This proof-of-concept has been shown to dozens of people, from school children to Holocaust scholars, with many commenting on the impact of the experience and potential for this kind of interface.

Author Keywords

Agents and intelligent systems; Computer-mediated communication; E-Learning and education; Multi-modal interfaces; Dialogue systems; Holocaust testimony preservation

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces – *Natural language*; H.5.1 Information Interfaces and Presentation: Multimedia Information Systems – *Artificial, augmented, and virtual realities*; I.3.3 Computer Graphics: Picture/Image Generation

IUI'14, February 24–27, 2014, Haifa, Israel.

ACM 978-1-4503-2184-6/14/02..$15.00.
http://dx.doi.org/10.1145/2557500.2557540

INTRODUCTION

Recently, a number of computer interfaces (e.g., skype, facetime, hangouts) have been created to allow people in different locations to engage in multi-modal conversational interaction. However all of these require that all conversation participants be available at the same time. If participants are available at different times, current interfaces generally only allow a single message exchange, before waiting for another participant to sign on and reply.

We introduce a new interface for allowing limited interactive conversations with participants who are not currently available for interaction, but who have been able to record relevant material previously. Existing dialogue system technology is used to organize the recorded material such that new partners can engage in conversational interaction with the participant. This "time-offset interaction" allows the intimacy, relevance and interactiveness of conversation for material and participants who were once only accessible via unidirectional communication. This concept has appeared in a number of science fiction and fantasy movies, including *Superman* (1978, and several subsequent versions), in which Superman is able to ask questions and receive instructions from his long dead father; *I, Robot*, in which the Police Detective Del Spooner (played by Will Smith) is able to question a holographic representation of the recently deceased Dr. Alfred Lanning (played by James Cromwell); and *Harry Potter*, in which people in pictures are able to hold conversations with viewers of the picture.

Technology borrowed from Virtual Humans [4] provides many of the basic elements needed to make this idea a reality. We adapt elements from the publicly available Virtual Human Toolkit [5][1] as well as a new video player and video-blending techniques to create a time-offset interaction interface.

One area where time-offset interaction can make a real difference is in introducing new learners to the Holocaust though direct conversational interaction with survivors, who relay their personal experiences as related to specific concerns of the learners. Today this happens frequently, as survivors visit school classrooms, museums, and public lectures. Given the advancing age of the survivors, physical interaction will no

[1] http://vhtoolkit.ict.usc.edu

longer be possible in the near future, but time-offset interaction can preserve much of the interaction style and impact of these encounters.

In this paper, we describe a first proof of concept of a time-offset interaction interface. This proof of concept is the first stage in the *New Dimensions in Testimony* project [7], which will allow people to verbally ask questions to a persistent representation of Pinchas Gutter, a Holocaust survivor, and receive his answers to questions on a range of topics. The current proof of concept is limited in scope – appropriate for a demonstration of the concept, delivered by someone who knows approximately what to ask. Future work involves improving both the fidelity of the display, as well as the breadth of available interactions and accessibility for a broad user group.

In the next section, we describe the user experience, including displays for both audience and demonstrator, and a sample dialogue. The following sections describe the system architecture, the recording and video processing used to create a persistent conversational presence, and a small evaluation of the language processing performance. Finally, we sketch the next steps, that will lead to a more generally usable system for a broad range of naive interactors.

USER EXPERIENCE

The proof-of-concept system is designed to be shown to an audience by a trained demonstrator. It is typically run from a laptop computer, projecting the survivor onto an external display while keeping all the controls on the laptop's built-in display, invisible to the audience (we have tested on both Windows and Mac OS, and on a variety of laptops, including the MacBook Air). The external display is typically a projector screen or large television, but can be any computer display depending on availability; for natural interaction it is usually preferable to show the survivor close to life size, and this can be achieved by rotating a large-screen television to portrait orientation, when feasible.

The system is composed of several components running in separate windows (see the system architecture section below), but a demonstrator typically sets it up with only minimal interaction with three components.

1. The Launcher window (a component of the Virtual Human Toolkit [5]) is used to start the demo.

2. The survivor display (an instance of a video player) is positioned on the external display, and maximized to occupy the full display (Figure 1).

3. The audio acquisition client (also a component of the Virtual Human Toolkit) is brought into focus, and the mouse cursor is moved to the push-to-talk region (Figure 2).

After the initial setup, no further interaction is required with the window displays, and the remainder of the demo consists of natural conversation between the demonstrator and the survivor. The demonstrator talks to the survivor through a microphone, using a mouse or similar pointing device for push-to-talk (push while talking, release when done). The

Figure 1. Pinchas Gutter, a holocaust survivor, as he appears on the screen talking to an audience ©University of Southern California

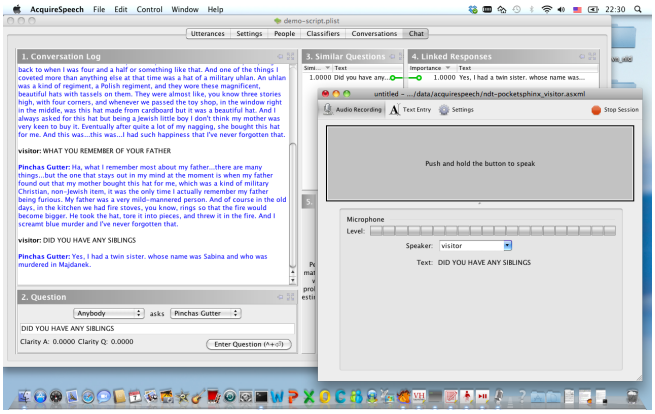

Figure 2. A typical demonstrator screen: the acquisition client is in focus, with NPCEditor in the background ©University of Southern California

2	Opening sequence	2	Survival and coping
3	Childhood	2	Resistance
3	Sister	2	Closing sequence
2	Religion	3	Off-topic

Table 1. Survivor utterances in the proof-of-concept, by topic

microphone is typically a head-mounted close-talking device, though for informal demos the built-in laptop microphone can be used as well. For a large audience, a wireless microphone and mouse can be used, eliminating all physical contact between the demonstrator and the computer, and allowing the demonstrator to assume an ideal position while talking to the survivor and to the audience.

A typical demo will start by greeting the survivor, and proceed with a series of questions depending on the allotted time and the interests of the audience. The present proof-of-concept has 19 statements by the survivor, any of which can be solicited at any time by a question asked by the demonstrator. The statements fall into natural short blocks that form coherent mini-conversations (see Table 1), allowing the demonstrator and survivor to explore a topic in some depth. Three of the survivor's statements are of special status: they are "off-topic" responses which are used when the survivor fails to understand the demonstrator's utterance, in order to deflect the question when the survivor does not have an appropriate answer. An example short demo is provided in the video accompanying the submission, and transcribed in Figure 3. Following the interaction, if the audience is technically inclined, the demonstrator can also drag some of the component windows to the external display to show how the system works.

SYSTEM ARCHITECTURE

Figure 4 shows the system architecture, which makes use of several components from the Virtual Human Toolkit [5], including a module launcher and the VHMsg messaging system, built on top of ActiveMQ.[2] The system consists of automatic speech recognition modules (ASR) – the client and the server; a natural language understanding and response selection module (NLU); and a rendering module (Video Player) that displays the response for the survivor. The individual modules are separate applications that communicate with each other using the ActiveMQ messaging platform. This design allows us to reconfigure the system efficiently and reuse the modules in different projects. The Launcher allows the user to start and stop the individual modules if necessary.

We deliver the system as a single application. When the user double clicks to launch the application, the application starts the ActiveMQ process, starts the Launcher, and the other four components. Termination is done in reverse order: when the user quits the main application, it shuts down individual modules and then terminates the ActiveMQ process. The Video Player module was specifically created for this project; we adapted and customized the other modules from the Virtual Human Toolkit. As the system is intended to be used by people who are not computer experts, the main user interface

[2] `http://activemq.apache.org`

Demonstrator	Hello Pinchas.
Survivor	Hello.
Demonstrator	Can I ask you some questions?
Survivor	You can ask me any questions you would like.
Demonstrator	Growing up, did you have any siblings?
Survivor	Yes, I had a twin sister, whose name was Sabina, and who was murdered in Majdanek.
Demonstrator	Do you remember Sabina?
Survivor	I only remember the plait that she had, she was blonde, blue eyed, just like we all were, my mother were. And when we arrived in Majdanek she was separated from my mother and was put amongst the children. I was together with my father. My father told me to say I was eighteen years old and she must have spotted my mother standing amongst the adult women who were also separated and I saw her too and I saw my sister suddenly running towards my mother. She embraced her. And from that time on, whenever I think of my sister, the only thing that I can see, the only thing that I can visualize, is just that braid, blonde, long, plait and nothing else.
Demonstrator	Pinchas, why do you tell your story?
Survivor	I tell my story for the purpose of improving humanity, if that's possible. Even drop by drop by drop. You know, like a drop of water falls on a stone and it erodes a little bit and smooths it out. So hopefully by me telling my story over and over again, I will achieve, hopefully achieve the purpose of making humanity or the world a better place to live in.
Demonstrator	Thank you, Pinchas, for sharing your story with us.
Survivor	Thank you very much for listening to me.

Figure 3. A short demo (transcription of the accompanying video)

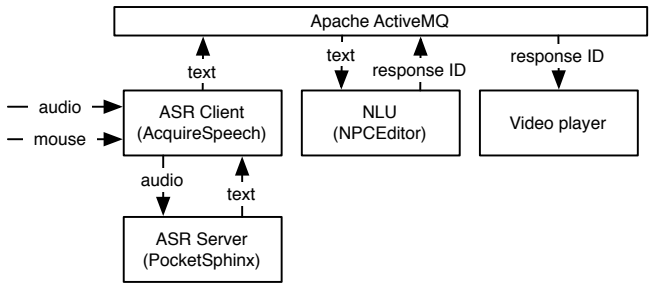

Figure 4. System architecture. We omitted the Launcher and the control events among modules for clarity. ©University of Southern California

challenge was to hide the system complexity from the user by integrating the modules together and making the process of launching and terminating the application as seamless as possible. In the rest of the section we summarize the design of the individual modules.

ASR Client

The speech recognition part of the system consists of two modules: a client and a server. The client, called AcquireSpeech, presents a GUI front end to the user. The ASR currently operates in push-to-talk mode: the user presses a button on the AcquireSpeech window before she starts speaking and releases the button once she is done. AcquireSpeech collects the audio data that arrives from the microphone between the mouse press and release and forwards the data to the server at 0.25 second intervals. In our experiments we observed that users often start speaking before they press the push-to-talk button and release the button before they finish their sentence. This observation led us to modify the AcquireSpeech to extend the recorded audio interval by adding a short segment of recording before the button is pressed and after the button is released. The client listens to the audio constantly and keeps track of the most recent 0.25 seconds of audio in a buffer. When the user presses the button, the application sends the data from the buffer to the server. When the user releases the button, AcquireSpeech keeps recording and sending for another 0.25 seconds. AcquireSpeech is a highly modular and customizable application that allows us to adapt its interface to projects (e.g., [10]) that can use multiple recording sources, recognition engines, and control buttons. The system can be customized to automatically log the audio in files.

ASR Server

Many speech recognition systems are compatible with the VH Toolkit, and could easily be used in this type of application [8]. For the proof of concept, we use CMU PocketSphinx speech recognizer.[3] Using the PocketSphinx libraries we created a server application that accepts a digital audio stream over a TCP connection and sends back the recognition results. The language model (back-off 3-gram) was trained on a set of 81 questions written by the system developers using the CMU Statistical Language Modeling toolkit [2]. Each of the 81 questions corresponded to one of the 19 survivor responses noted in Table 1. For US English speakers we used the Wall Street Journal (WSJ) acoustic models distributed by CMU. For British English speakers we used acoustic models trained with the WSJCAM0 corpus [11]. The CMU pronouncing dictionary v0.7a [12] was used as the main dictionary with the addition of domain-dependent words, such as names.

NLU

For language understanding and response selection we use an NLU system called NPCEditor [6]. The system is based on a statistical language classification approach, using ideas from cross-lingual information retrieval. A designer specifies a list of system responses and links them to some training set utterances. The system uses this data to make an implicit dictionary. When the system receives a new utterance, it applies the dictionary to construct a representation of the most likely response. It compares that response representation to each response in its database and returns a list of matching responses. NPCEditor also knows when it does not know the answer: when the score of the best matching response is below a specified threshold, the classifier returns an empty list. This approach has been shown to be both robust to the errors in ASR output and very effective for implementing simple conversational agents [6].

In addition to the language classifier, NPCEditor includes a dialogue manager that processes the classification results and picks the actual response that is presented to the user. For example, when the classification list is empty, NPCEditor returns one of the responses specified as "off-topic" – e.g., "The question that you asked me, I am afraid I won't be able to answer." When the list contains several responses, the dialogue manager attempts to pick a response that the user has not heard yet. NPCEditor sends the selected response identifier to the Video Player via the ActiveMQ connection.

Video Player

The final module of the system is a custom Video Player. It has one video clip for each survivor response and a collection of 1–2 second "idle" clips, which serve to create listening and waiting behaviors when one of the longer response clips is not playing. Creation of these clips is described in the next section. The Video Player is based on the libraries from the VLC project[4] and is capable of playing MP4-encoded files. When the system is idle, the player plays a sequence of randomly chosen idle clips. Once it receives a response ID from NPCEditor, it selects, schedules, and plays back the appropriate response video clip.

RECORDING PROCESS AND VIDEO PROCESSING

Capture

We captured a long form interview with the survivor illuminated by the Light Stage 6 device [1]. The Light Stage 6 device is a geodesic dome approximately 26 feet in diameter. There are approximately 660 individually controllable light sources mounted on every vertex of the dome. For this shoot we set all the lights to the same intensity, resulting in an even illumination of the survivor. We shot with a RED Epic camera mounted in portrait orientation. The footage was then edited and converted to a set of individual video clips, one response per clip.

Transitions

To create seamless transitions between clips, we asked the survivor to always start and end a response in the same "neutral pose". We morphed the start and end frames of each video clip to a single neutral pose frame. We also created short "idle" clips, 1–2 seconds each, using the morph transitions as connective tissue. The morph transitions were computed automatically using optical flow. However, we found that sometimes the optical flow failed on frames which were too dissimilar. In addition, even when the morphs worked

[3] http://cmusphinx.sourceforge.net

[4] http://www.videolan.org/vlc/index.html

	#Utterances	Word Error Rate (%)	Robustness (%)
US	39	21	95
UK	65	17	89
All	104	19	91

Table 2. Speech recognition and NLU performance

well, the motion appeared too linear and not natural. This convinced us that a more sophisticated transition system was needed.

EVALUATION

Speech recognition

For speech recognition our main evaluation metric was word error rate (WER). The WER is calculated by comparing the speech recognition output with what the speaker actually said (transcription), and can be formulated as follows:

$$\text{WER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Length of transcription string}}$$

Table 2 shows speech recognition results for two demonstrators of the system, one from US and one from UK, as well as the average WER. The lower the WER the better. This evaluation is based on 104 utterances, consisting of several extended demos from each speaker.

Natural language understanding

We measure the NLU's ASR robustness by passing through the NPCEditor both the speech recognition output and the transcription and counting the number of matches between the two outputs. Intuitively this is a measure of how much ASR errors will impact NLU results. NLU ASR robustness can be formulated as follows:

$$\text{NLU ASR Robustness} = \frac{\text{Number of matches}}{\text{Number of utterances}}$$

The same utterances that were used for calculating the WER were also used for calculating the NLU ASR Robustness. Table 2 shows the Robustness for the two demonstrators as well as the average Robustness; the higher the Robustness the better. Since the NLU ASR Robustness is much higher than ASR accuracy ($1 - \text{WER}$), we can see that the NLU can recover from a high percentage of ASR errors without impacting the overall performance. For a demo system, demonstrators can be trained to say things that the NLU would be able to understand, so ASR NLU robustness is the dominant factor in performance. For less experienced users, domain coverage would also be very important, as users might not know how to formulate questions in a way that could be understood (whether or not the ASR performance is good).

The above results show that the system performs well and allows for successful system-user interactions. In the future we intend to experiment with different language and acoustic models and to perform a rigorous evaluation with more users.

User reactions

Probably the most important evaluation of the proof of concept is its impact on viewers of the system. While we have not undergone a formal study of viewer reactions, the system has been demonstrated to hundreds of people, individually, in small groups, and in larger presentations with dozens of audience members. The reaction has been quite positive, both to the concept of time-offset interactions, the specific interface, the content from the survivor, and possible other uses of this kind of interface.

FUTURE WORK

We intend to extend the proof-of-concept into a full prototype, which will provide a richer experience and allow the survivor to interact with people who are not trained as demonstrators. The extensions will include an enhanced graphic display environment, and additional interactive content.

Graphics improvements

We will design and implement an improved video recording process for the survivor testimony which will allow for holographic video projection, 3D stereo display, and multi-view high definition archiving. The goal of the process is to record the appearance of the survivor giving their testimony in as "future-proof" a manner as possible, with the greatest possible fidelity and repurposeability, that reasonable expense will allow.

The survivor will be comfortably seated, attractively illuminated, and recorded by a semicircular array of 50 high-definition video cameras. These will provide video streams of the testimony from all angles from left profile to frontal to right profile. The views along this arc will be sufficiently closely spaced to allow "optical flow" image processing algorithms to realistically synthesize any possible camera position along the arc; the result will be a dataset which can be used by a state-of-the-art life-sized autostereoscopic holographic display system, or viewed interactively on a PC from a user-directed angle, providing a solid sense of three-dimensional presence.

In addition to the semicircular arc of angles, approximately ten of the cameras will be dedicated to recording additional special views of the survivor. These include cameras which will record close-up video of the survivor's face and hands, and cameras which will record additional viewpoints to enable 3D geometric shape information for each frame of the survivor's testimony. Two of the frontal cameras will be placed in optimal positions to record left-eye/right-eye 3D stereo video.

When the subject is available, we will additionally perform a Light Stage 6 full-body geometry and reflectance scan of them wearing the same clothing in which they are interviewed, as well as Light Stage X [3] facial scans of key facial expressions with submillimeter accuracy. These datasets will document the appearance of the survivor with the best technology available today, and could be used in the future to enhance the resolution of the recorded survivor testimony video.

In order to make the transitions between responses appear as seamless as possible, the application will include a transition

system based on cluster analysis of the video frames. The application can run on a 2D monitor or a 3D monitor requiring 3D glasses. The video could also be projected on a 2D screen or a 3D screen (requiring each viewer to wear 3D glasses) at life size.

Finally, additional processing and rendering of the acquired dynamic performance and static reference data will enable, in the not-too-distant-future, for the survivor to be faithfully rendered to match the lighting conditions of any museum or classroom environment, providing the closest possible impression of the survivor being present within the same space and lighting of the participating audience's environment.

Content enhancements

In order to hold coherent conversations with uninitiated people, the survivor will need to record many more statements than are available at present, which would provide answers to the most common questions people want to ask. The full prototype will include additional content on the existing topics of family, religion and resistance, as well as other topics of common interest. But in order to address the common questions (rather than topics), we need to find out what these questions actually are.

Data collection forms a crucial step when transitioning a synthetic character from private demonstration to public interaction, because the language understanding components need to know not only *what* information people want from the character, but also *how* they ask for it in conversation. Preparing a person for interacting with the public will require a similar collection of the actual language people use. To collect the data, we plan to build a mock-up system with only voice recordings of the survivor, to allow expeditious and inexpensive addition of content. To save development time and cost we will use a person to select the survivor's statements in real time rather than trained speech recognition and language understanding ("Wizard of Oz" scenario). The mock-up will be taken to target audiences (e.g., schools, museums), and will be extended iteratively with additional survivor statements until a large set of "frequently asked questions" are identified. Based on these questions we will design the interview for the video recording sessions, so that the recorded statements will be the ones most frequently sought out by the public. Based on experience with museum interactions with synthetic characters [9, 10], we believe that a set of several hundred statements on a small number of select topics can provide sufficient content to allow the survivor to keep sustained interactions with the public well into the future.

ACKNOWLEDGMENTS

Creation of the proof-of-concept was made possible by generous donations from private foundations and individuals. We are extremely grateful to the Pears Foundation, Alan Shay, Lucy Goldman, and the Wolfson Foundation for their support. Special thanks to Pinchas Gutter for sharing his story, and for his tireless efforts to educate the world about the Holocaust.

REFERENCES

1. Chabert, C.-F., Einarsson, P., Jones, A., Lamond, B., Ma, A., Sylwan, S., Hawkins, T., and Debevec, P. Relighting human locomotion with flowed reflectance fields. In *SIGGRAPH 06: ACM SIGGRAPH 2006 Sketches*, ACM (2006), 76.
2. Clarkson, P., and Rosenfeld, R. Statistical language modeling using the CMU-Cambridge toolkit. In *Proc. of Eurospeech* (Rhodes, Greece, 1997).
3. Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X., and Debevec, P. Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, SA '11, ACM (New York, NY, USA, 2011), 129:1–129:10.
4. Gratch, J., Rickel, J., Andre, E., Cassell, J., Petajan, E., and Badler, N. Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems* (2002), 54–63.
5. Hartholt, A., Traum, D., Marsella, S. C., Shapiro, A., Stratou, G., Leuski, A., Morency, L.-P., and Gratch, J. All together now: Introducing the virtual human toolkit. In *International Conference on Intelligent Virtual Humans* (Edinburgh, UK, Aug. 2013).
6. Leuski, A., and Traum, D. Practical language processing for virtual humans. In *Proceedings of the 22nd Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-10)* (2010).
7. Maio, H., Traum, D., and Debevec, P. New dimensions in testimony. *PastForward*, Summer (2012), 22–26.
8. Morbini, F., Audhkhasi, K., Sagae, K., Artstein, R., Can, D., Georgiou, P., Narayanan, S., Leuski, A., and Traum, D. Which ASR should I choose for my dialogue system? In *Proceedings of the SIGDIAL 2013 Conference* (Metz, France, August 2013), 394–403.
9. Robinson, S., Traum, D., Ittycheriah, M., and Henderer, J. What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)* (Marrakech, Morocco, 2008).
10. Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., Leuski, A., Noren, D., and Swartout, W. Ada and Grace: Direct interaction with museum visitors. In *Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September 12–14, 2012 Proceedings*, Y. Nakano, M. Neff, A. Paiva, and M. Walker, Eds., vol. 7502 of *Lecture Notes in Artificial Intelligence*, Springer (Heidelberg, September 2012), 245–251.
11. Vertanen, K. Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments. Tech. rep., Cavendish Laboratory, University of Cambridge, 2006.
12. Weide, R. The CMU pronouncing dictionary, 2008.