

MULTIMODAL SPEECH SYNTHESIS

<http://www.research.att.com/projects/tts>

J. Schroeter, J. Ostermann, H.P. Graf, M. Beutnagel, E. Cosatto, A. Syrdal, A. Conkie, and Y. Stylianou

AT&T Labs – Research
Florham Park and Redbank, New Jersey
U.S.A.

ABSTRACT

Multimodal Speech Synthesis (“Talking Heads”) encompasses synthesis of speech from text (“Text-to-Speech”, TTS) plus synthesis of a visual presentation of a face that is lip-synced to the generated audio (“Visual TTS”, VTTS). Talking Heads are now practical because of the ever-increasing computing power and falling prices of computer hardware. This paper highlights recent technological breakthroughs relevant to the two modalities. In addition, it exposes synergies between the audio and visual technology components. Finally, the paper summarizes test results that highlight the impact of Multimodal Speech Synthesis in communications and e-commerce applications.

1. INTRODUCTION

Text-to-speech (TTS) synthesis technology gives machines the ability to convert arbitrary text into audible speech. TTS systems provide voice output for all kinds of information stored in databases, e.g., phone numbers, addresses, car navigation information, or for information services such as restaurant locations and menus, movie guides, etc. TTS may also be used for reading books and for voice access to large information stores such as encyclopedias, reference books, law volumes, etc.

Visual TTS (VTTS) synthesis technology gives TTS systems a “face” with the goal of enhancing human-computer interaction. Usually run as a separate component alongside the (audio) TTS system, a VTTS module is controlled by phonetic information (phoneme identity and timing) plus facial expression data it receives from the TTS engine. Applications such as, for example, virtual operators or customer care/help desks on the Web require realistic “visual agents” that look reasonably human and speak naturally. For these and other types of applications, lip-synchronization of Audio TTS and Visual TTS is essential. Such visual agents can be implemented as cartoon-like characters (avatars) using 3D-models, or they are synthesized photo-realistically using sample-based technologies. Both of these approaches can be driven by an interface complying with the MPEG4-standard.

This paper is organized as follows. Section 2 summarizes the state-of-the-art in audio text-to-speech. Section 3 does the same for visual text-to-speech. Section 4 describes our

way of synchronizing audio and visual subsystems. Section 5 addresses issues with evaluating Multimodal Speech Synthesis systems. Finally, section 6 concludes with a look at where this technology is going over the next few years.

2. AUDIO TEXT-TO-SPEECH

A block diagram of a typical TTS system is shown in Fig. 1. The first block is the message text analysis module that converts the message text to a string of phonetic symbols and prosody targets (i.e., for fundamental frequency, duration, and amplitude). The text analysis module actually consists of a multitude of sub-modules with separate, but in many cases intertwined, functions. Input text is first analyzed and transcribed. For example, in the sentence “Dr. Smith lives on Elm Dr.”, the first “Dr.” is transcribed as “doctor”, while the second “Dr.” is transcribed as “drive”. Next a syntactic parser recognizes the part of speech for each word in the sentence and disambiguates the sentence constituent pieces in order to generate the correct string of phones with the help of a pronunciation dictionary. Thus, for the above sentence, the verb “lives” is disambiguated from the potential noun “lives” (plural of “life”). Finally, with punctuated text, syntactic and phonological information available, a prosody module predicts sentence phrasing and word accents and generates a prosody contour. The second block in Figure 1 assembles the units (traditionally diphones, i.e., transitions from one phone to the next that are cut in stationary portions of the speech sounds) according to the list of

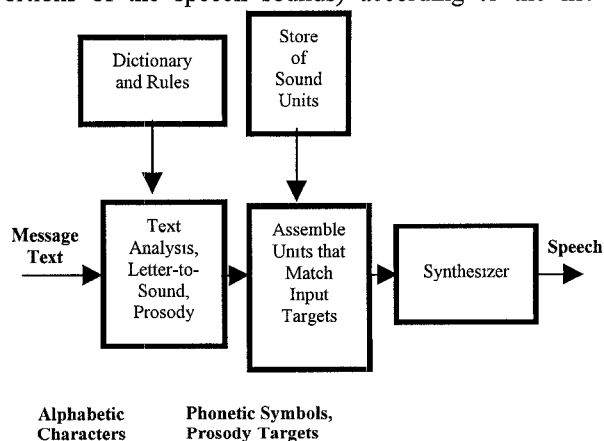


Fig. 1: Simplified Block Diagram of a TTS system.

phonetic targets generated by the front-end. Then the selected units are fed into a speech synthesizer that generates the speech waveform for presentation to the listener. For a more general, introductory overview of TTS technology, see, for example, [1].

Recently, a new method, called “Unit Selection Synthesis” has emerged for obtaining higher quality TTS. Based on earlier work done at ATR in Japan [2], this method employs speech databases recorded using a “natural” (lively) speaking style. In contrast to earlier concatenative synthesizers, unit-selection synthesis automatically picks the optimal synthesis units (on the fly) from an inventory that can contain thousands of examples of a specific diphone, and concatenates them to produce the synthetic speech.

This process is outlined in Fig. 2, which shows how the method finds dynamically the optimal path through the unit-selection network corresponding to the sounds for the word ‘two’. The best choice of units depends on factors such as spectral similarity at unit boundaries and on matching prosodic targets set by the front-end. There are two good reasons why the method of unit-selection synthesis is capable of producing higher quality speech synthesis than the older methods. First, on-line selection of speech segments allows for longer units (whole words, potentially even whole sentences) to be used in the synthesis if they are found in the inventory. This is the reason why unit-selection appears to be well suited for limited-domain applications such as synthesizing telephone numbers to be embedded within a fixed carrier sentence. Even for open-domain applications, such as e-mail reading, advanced unit selection can reduce the rate of unit-to-unit transitions and, consequently, increase the segmental quality of the synthetic output. Second, the use of multiple instantiations of a unit in the inventory, taken from different linguistic and prosodic contexts, reduces the need for prosody modifications that degrade naturalness. More details can be found in [3, 4]. Interactive demos are available on our website (see title page for URL).

3. VISUAL TEXT TO SPEECH

Visual Text-to-Speech systems that provide “Talking Heads” are playing an increasingly important role in computer and communication interfaces. Such VTTS systems are built either upon 3D-models (model-based VTTS) or upon recorded video clips (sample-based VTTS). Using either of the two methods, we may take advantage of the same MPEG-4-based mechanism for interfacing between TTS and VTTS.

Model-based VTTS employs a 3D polygon mesh face model with defined facial actions. The face model in its neutral position is defined as a scene graph consisting of transform nodes for rotation and translation of head and eyes as well as polygon meshes specifying the shape of the

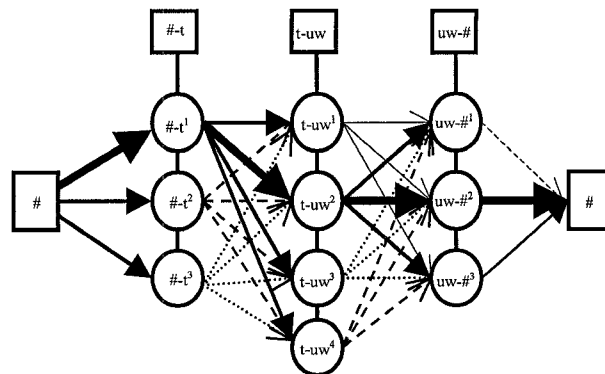


Fig. 2: Diphone Unit Selection for the word “two”.

object surface. The surface can be colored or texture-mapped. In addition to this static model, we define the facial expressions for the model using a facial animation table (FAT) for each expression that we want to animate. Each FAT defines for every vertex of the model its displacement as a function of the amplitude of this expression. Our model allows for facial expressions like smiling, anger etc. [5]. For lip motions, we use the coarticulation model of [6].

A different approach is our sample-based VTTS [7], which employs a set of video snippets of the mouth area that have been extracted from video recordings of a real person talking. Related work using non-synthesized speech has also been published in [8]. In our work, the samples are labeled according to the mouth shape. This database of mouth shapes is used to select at each time step the mouth unit that is optimal for the speech pronounced by the TTS system. Selecting the optimal mouth shape is done in a manner much like the unit-selection Viterbi-search depicted in Fig. 2. This yields a system that is functionally equivalent to our 3D-model VTTS, but the image looks much more natural. Note, however, that in previous versions, the lip movements looked somewhat jerky and therefore less natural due to artifacts in the transition between video units. The Viterbi search for optimal video sequences alleviated the problem.

4. AUDIO-VISUAL INTEGRATION

The MPEG-4 standard [9] anticipates that talking heads will serve an important role in future customer service applications. For example, a customized agent model can be defined for games or web-based customer service applications.

A key issue for integrating TTS and VTTS is the synchronization of the speech stream with the Face Animation Parameter (FAP) stream through phonetic and timing information that is sent by the TTS engine. This process is illustrated in Fig. 3 that shows a simplified block diagram of an MPEG-4-compliant interface between TTS and VTTS engines.

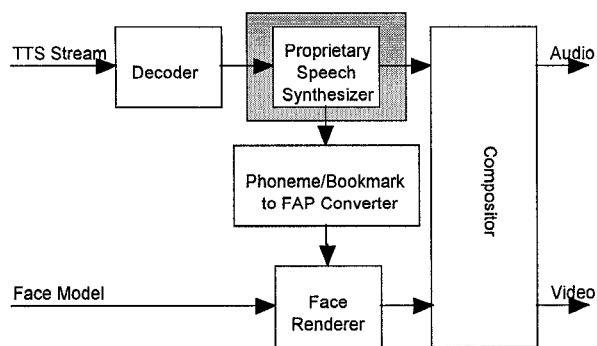


Fig. 3: Block Diagram of an MPEG-4 TTS/VTTS Interface.

Our Multimedia Speech Synthesis System uses input text that may be augmented with prosodic and facial expression information. The MPEG-4 Decoder decodes this information and feeds it to the TTS system. The synthesized speech samples are then handed to the Compositor. The Compositor presents synchronous audio and video to the user, perhaps in a streaming format.

In order to assure complete synchronization between audio and video, the TTS system is implemented in a two-channel client/server architecture with the (audio) TTS acting as the server and the VTTS system acting as the client. One channel of the interface is used for inputting text and for outputting synchronous speech samples; the other is used to input asynchronous commands like “stop”, “pause”, “resume”, etc. In addition, this second channel of the TTS server outputs an asynchronous stream of “bookmarks”, or event notifications, which identify points in the audio timeline. Bookmarks include both user-specified instants defined in the text input (e.g., FAP tags) and notifications generated within TTS, identifying phoneme, word, and sentence boundaries. Bookmarks are reported as soon as the associated times are known accurately, giving the VTTS client as much time as possible for rendering before playback. If the delay is not sufficient (e.g., if the audio queue ever becomes empty because of the VTTS client running too slowly), the Compositor must provide additional buffering to insure synchronization. In our Multimedia Speech Synthesis System, the synchronous and asynchronous channels are part of an S.100 [10] compliant

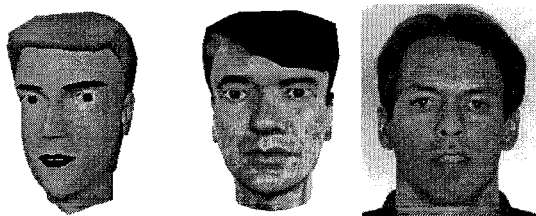


Fig. 4: Standard cartoon-like Talking Head (left), texture-mapped TH (middle), and sample-based TH (right)

implementation of a wireline protocol.

5. EVALUATION RESULTS

Multimedia Speech Synthesis is designed to enhance human computer interactions. Does it do so and, if it does, in what sense? Before answering this important question, we briefly summarize relevant evaluation techniques.

Evaluation of Multimedia Speech Synthesizers needs to employ human subjects, “listeners” and “viewers”. Proper evaluation paradigms using “naïve” subjects are a necessity in any synthesis-related R&D effort because of the danger of researchers and developers being biased towards favoring their own product over other products. The purpose of an evaluation may be diagnostic (what’s wrong with the system), or may be to determine the adequacy of the overall system for a specific application (functional assessment), or may be benchmarking against other systems. TTS and VTTS evaluations are more difficult than, for example, speech coder evaluations, because of the many TTS modules that all have to work together for optimal results: any minor shortcoming of any module will have a negative effect on the quality of the output. The scope of a test may be the “system” as a whole, or may be any specific part of it. Tests of (audio) TTS, for example, may evaluate speech segments (vowels and consonants), prosody (pitch, durations, and amplitudes), voice quality (of the voice donor), or overall quality (functional adequacy, comprehension). VTTS systems may be evaluated in terms of “jerkiness” of lip movements, “natural-looking” head movements, or “skin color and texture naturalness”, among many other aspects. On a higher level, we may be interested in evaluating intelligibility (% correct lip-reading for VTTS) and naturalness (how believably “human” is the synthetic output?). Because of the multi-dimensionality of any synthesis system, these tests cannot just sample a few contexts and then generalize to others. Consequently, we should always test a synthesis system in the context of the intended application with as large a test corpus as possible. This is a tedious process, but when done properly evaluations can drive synthesis technology towards higher quality [11].

In order to answer the question about the benefits of a Multimedia Speech Synthesis System (“Talking Head”), we have conducted several evaluation tests. In 1998, in a test with 190 subjects, we explored whether the system can help users perform certain tasks more accurately or more efficiently. Other aspects of the test evaluated “appeal” to the user and acceptability, as well as intelligibility under noisy conditions. Detailed results of our study are available in [12].

Three experiments were carried out at, and with collaboration of, Princeton University. In the first experiment, we evaluated the performance benefits of using

a Talking Head (TH) in a number intelligibility test in noise. In the second and the third experiment, using an information kiosk application scenario, we explored whether user interest in the task and its appeal increased or not and whether a TH can bridge system-inherent waiting times. In experiment 3, three different versions of THs (Fig. 4) pronounced a simple welcome message. The first two THs were 3D-model based, the last TH was sample-based. The texture-mapped 3D-model based TH was used only in experiment 3.

Results: In experiment 1 we found no significant advantage of using a TH in conveying digits to subjects when no background noise was used. However, when airport babble noise was injected at -2dB SNR, digit presentation without TH (reference condition) resulted in 17% digit errors, the cartoon-style TH improved the error rate to 9%, and the sample-based TH resulted in only 7.5% errors. This improvement had to be “paid for” by a slight increase of task completion time (+10%). A frame rate of at least 18 Hz was needed to achieve these results. In experiment 2 the subjects were asked to use a simple interactive real-time information system on theater shows. THs were used to fill simulated server access and Internet transport delays but did not deliver any actual verbal information. Audio-only and text-only versions of the task served as reference conditions. Subjects were generally more satisfied (shorter perceived waiting times, appearance of faster service) with the audio and/or TH versions than with the text-only version. Once exposed to the non-text versions, the text-only version was judged more “annoying”. The THs were judged as “fairly friendly” and even “marginally useful” (note that the THs had been scripted not to deliver any useful information!). Experiment 3 (3 different THs uttering a welcome message) resulted in our 1998 snapshot of TH ratings: subjects preferred the cartoon-like TH (appeal rating of 5.0) over the sample-based TH (3.3) and the texture-mapped TH (2.7). Note that a previous, non-unit selection, version of the sample-based TH was used.

In a variation of the 1998 evaluation [14] we recently found interesting cross-modality (video to audio, audio to video) influences in perceived quality ratings [also see, e.g., 14]. Most important, however, is our finding that ratings of the user experience in an e-commerce scenario are positively correlated with the likelihood of a purchase and that adding a high-quality TH results in higher ratings of the user experience. In another study [15], we found that a TH interface has a significant effect on building a cooperative environment between machines and users. These results provide important support for the notion of using Multimedia Speech Synthesis in future real-world applications in communications and e-commerce.

6. CONCLUSION

This paper summarized technological advances that enable the introduction of Multimedia Speech Synthesis applications. The quality of synthetic audio and video produced by TTS/VTTS systems is now high enough to start evaluating the usefulness of such systems. To this end, we have established that Multimedia Speech Synthesis Technology (“Talking Heads”) enhance the user experience in customer care and e-commerce. In the future, we will have to improve non-verbal aspects of the visual “agent” interface such as gestures, emotions, and natural-looking head movements.

REFERENCES

- [1] Pickett, J. M., Schroeter, J., Bickley, C., Syrdal, A., and Kewley-Port, D. Speech Technology, in *The Acoustics of Speech Communication*, Ch. 17, J. M. Pickett (Ed.), Allyn and Bacon, Boston, pp. 324-342, 1998.
- [2] Hunt, A., and Black, A. “Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database,” *Proc. ICASSP '96*, pp. 373-376, 1996.
- [3] Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A., “The AT&T Next-Gen TTS System,” *Proc. Joint Meeting of ASA, Forum Acusticum, and DAGA, J. Acoust. Soc. Am.* 105, No. 2, Pt. 2, p.1030, also see Conference CDROM Paper 2ASCA__4, 1999.
- [4] A. Conkie, “Robust Unit Selection for Speech Synthesis,” *Proc. Joint Meeting of ASA, Forum Acusticum, and DAGA, J. Acoust. Soc. Am.* 105, No. 2, Pt. 2, p. 978, also see Conference CDROM Paper 1PSCB_10, 1999.
- [5] Ostermann, J. “Animated Talking Head with Personalized 3D Head Model,” *J. VLSI Signal Processing* 20, pp. 97-105, 1998.
- [6] M. M. Cohen and D. W. Massaro, “Modeling Coarticulation in Synthetic Visual Speech,” In: M. Thalmann & D. Thalmann (Eds.) *Computer Animation '93*. Springer Verlag, Tokyo
- [7] Cosatto, E., and Graf, H. P. “Sample-Based Synthesis of Photo-Realistic Talking-Heads,” *Proc. of Computer Animation*, IEEE Computer Society, pp 103-110, 1998.
- [8] Bregler, C., Covell, M., Slaney, M. “Video Rewrite: Driving Visual Speech with Audio,” *Proc. of ACM SIGGRAPH*, pp. 353-360, 1997.
- [9] Ostermann, J. “Animation of Synthetic Faces in MPEG-4,” *Proc. of Computer Animation*, IEEE Computer Society, pp. 49-55, 1998.
- [10] S.100 Media Services API reference, available from <http://www.ectf.org/ectf/home.htm> : for an overview, try <http://www.ectf.org/ectf/tech/ctspwg.htm> .
- [11] R van Bezooijen, V. van Heuven, “Assessment of synthesis systems,” in D. Gibbon, R. Moore, R. Winski (Eds.) *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin and New York, 1997, Ch. 12.
- [12] I. Pandzic, J. Ostermann, D. Millen, “Synthetic Faces: What are they good for?” *The Visual Computer*, in press, 1999.
- [13] D. Millen, J. Ostermann, I. Pandzic, “An Evaluation of Facial Animation in Multimodal Online Applications,” *submitted to ICME 2000*.
- [14] J. G. Beerends, F. E. de Caluwe, “The Influence of Video Quality on Perceived Audio Quality and Vice Versa,” *J. Audio Eng. Soc.*, Vol. 47, No. 5, 1999, pp. 355-362
- [15] J. Ostermann, and D. Millen, “Talking Heads and Synthetic Speech: An Architecture for Supporting Electronic Commerce,” *submitted to ICME 2000*.