2010 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing

# Speech and Auditory Interfaces for Ubiquitous, Immersive and Personalized Applications

Lei Xie, Wenhuai Zhao, Xiangzeng Zhou, Xiaohai Tian, Bingfeng Li, Naicai Sun, Yali Zhao, Yanning Zhang
*Shaanxi Provincial Key Laboratory of Speech and Image Information Processing*
*School of Computer Science, Northwestern Polytechnical University, Xi'an, China*
*Email: lxie@nwpu.edu.cn*

***Abstract*—In this demonstration, we introduce our recent progress on speech and auditory technologies for potential ubiquitous, immersive and personalized applications. The first demo shows an intelligent spoken question answering system, which enables users to interact with a talking avatar via natural speech dialogues. The prototype system demonstrates our latest development on automatic speech recognition, keyword spotting, personalized text-to-speech synthesis and visual speech synthesis. The second demo exhibits a virtual concert with immersive audio effects. Through our virtual auditory technology, wearing simple earphones, listeners are able to experience immersive concert audio effects from an ordinary music file. We believe the technologies shown in the two demos can be easily deployed in many significant applications.**

***Keywords*-spoken dialogue system, question answering, speech recognition, keyword spotting, speech synthesis, visual speech synthesis, talking face, human-computer interaction, virtual auditory, head related transfer functions**

## I. Introduction

Speech is the primary and most convenient way of communications between individuals. Therefore, voice-based human-machine interaction is an indispensable part of an intelligent and ubiquitous system. Since the pioneer artificial intelligence research in 1950s, achieving a conversational machine has been an abiding goal of researchers. However, only recently, spoken interaction with a computer has become a practical possibility due to significant advances in speech and language technology as well as the emergence of powerful computers. The first demo shows our latest effort to a prototype of spoken question answering (Q/A) system towards natural human-computer interaction. In this system, users can talk to an avatar using free speech queries, and the avatar is able to response using both synthesized natural speech and speech-synchronized facial animation.

Auditory perception is human's ability to identify, interpret, and attach meanings to sound. Roughly 90% of the information we learn comes from auditory and visual perceptions. Compared with vision, auditory perception is omni-directional. 3D sound localization is one of our major auditory perception phenomena. We can locate sound sources in space naturally through binaural listening [1]. In our second demo, we show our recent work on a virtual concert system. Through HRTF-based binaural synthesis technology, wearing simple earphones, listeners are able to experience 3D spatial audio of a concert.

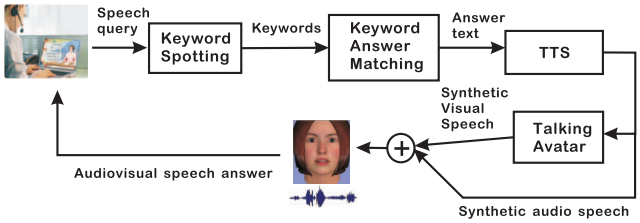


Figure 1. Architecture of the spoken Q/A system.

## II. Spoken Question Answering System

Figure 1 shows the architecture of our Chinese spoken question answering system. The system accepts speech queries form users and answers the queries via synthetic audio-visual speech. The architecture includes four modules: the KWS module detects in-domain keywords from the input speech; the keyword/answer matching module searches for the appropriate answers (in text form) from a database; the TTS module converts the answer text to speech waveform; the talking avatar module realizes visual speech and finally the user gets the response from a talking agent.

### A. Keyword Spotting

Keyword spotting (KWS) is an important branch of speech recognition technology, which is essential to a natural spoken dialogue system. The KWS module for a Q/A system is targeted to: (1) detect a set of pre-set keywords in flow of natural speech questions (that may have disfluency); (2) effectively eliminate out-of-domain questions. We propose a two-step online garbage based KWS approach, which first detects keywords based on competitive subword models and then recalls possible missing keywords in the second step.

### B. Personalized TTS

Statistical parametric speech synthesis, especially Hidden Markov Model (HMM)-based TTS [2], has recently been demonstrated to be very effective in generating high-quality speech. Compared with concatinative or unit selection approaches, HMM-based TTS has several advantanges, such

978-0-7695-4272-0/10 $26.00 © 2010 IEEE
DOI 10.1109/UIC-ATC.2010.59

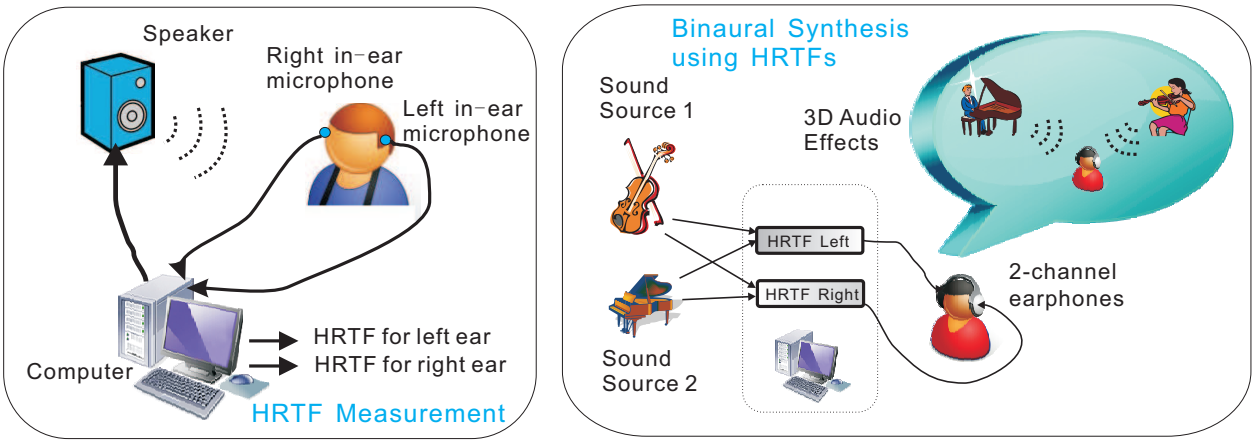


Figure 2. HRTF measurement (left) and binaural synthesis using HRTFs (right).

as flexibility to change voice characteristics and speaking styles, small footprint and robustness. We build an HMM-based Chinese TTS module for the Q/A system. Context-dependent syllable initial-final HMMs are estimated from about 5000 speech sentences uttered by 125 native Chinese speakers, resulting in a *mean voice*. Personalized TTS, e.g., synthesized voice from another speaker, is realized via maximum likelihood linear regression (MLLR) from the mean voice.

### C. Talking Avatar

Speech production and perception are bimodal in nature. We interpret a speaker's meaning, intent and mood from not only the auditory speech but also the movements of lips (lip-read), tongue and other facial muscles, i.e., visual speech. Research shows that the attention and trust of humans towards machines are able to increase by 30% if humans are communicating with talking faces instead of text-only [3], [4]. Therefore, a talking avatar is an indispensable part of natural human-computer interaction. We integrate a text-driven talking face module into the Q/A system. When answering the questions, a lifelike 3D talking avatar is driven by the TTS module and synchronized audio-visual speech is realized.

## III. Virtual Concert

By the two ears, humans can locate sounds in three dimensions: above and below, in front and to the rear, as well as to either side. Therefore, achieving realistic 3D audio effects is an essential factor to many applications pursuing immersive experiences, such as computer games, cinemas and home theaters. Conventional 3D sound technologies (e.g., 5.1 and 7.1 surround sound systems) employ multiple loudspeakers. The complicated and costly configuration hinders them from ubiquitous applications. In other commercial 3D technologies working with two channels (stereo), the sound images that normally extend to the locations of the left and right speakers are widened to extend beyond the speakers (called stereo enhancement). However, these technologies have no ability to position individual sounds around a listener, nor do they have the ability to position sounds behind, above, or below the listener.

HRTF-based virtual auditory technology is able to realize immersive 3D experience through simple headphones or two loudspeakers [5]. Head-related transfer functions describe how a sound from a space point is filtered by the diffraction and reflection properties of the head, pinna, and torso, before the sound reaches the transduction machinery of the eardrum and inner ear. Humans locate sounds through the interaural time difference (ITD), interaural amplitude difference (IAD) and the subtle tonal features raised by pinna and concha of the ear. HRTFs systematically integrate these factors through functions $H_\varepsilon(f, \theta, \phi)$ for the two ears ($\varepsilon = \{left, right\}$) with frequency $f$, the horizontal angle $\theta$ and the vertical angle $\phi$, between the ears and sound source.

HRTFs can be measured by inserting miniature microphones into the ear canals of a human subject or an artificial head (dummy head or manikin) [5], as shown in Figure 2. A measurement sound is played by a speaker (placed with specific $\theta$ and $\phi$ to the head) and collected by the microphones. HRTFs are then derived by a computer from the collected signals. This procedure is repeated for various spatial sound locations. 3D sound synthesis is straightforward, shown in Figure 2, where HRTFs serve as a pair of audio filters to reproduce the sound localization cues at the ears of the listeners.

In fact, there is a latent parameter $a$ in HRTF functions – $H_\varepsilon(f, \theta, \phi, a)$. This parameter describes the personalized feature. That is to say, the HRTF functions depend on the shapes of head and ear, which are different between individuals. However, measuring HRTFs is a complicated and elaborate procedure. Practical 3D audio applications usually use HRTFs measured from a dummy head. Figure 3 show the *BHead* dummy head and a snapshot from our

Figure 3. BHead Dummy head (left) and HRTF measurement in a quite room (right).

HRTF measurement procedure [5].

HRTF-based virtual auditory technology facilities applications in pursuit of ubiquitous and immersive experiences. In this demo, we show a prototype of 3D virtual concert system that provides users with immersive concert audio effects. Through a graphic user interface which simulates a concert scenario, users can drag-and-drop instruments/sound-tracks around the listener. After bianural synthesis, users can easily enjoy a realistic symphony performance through simple earphones, as illustrated in Figure 2.

## IV. Conclusions

Speech and auditory are human's primary capabilities to interact with environments. In this demonstration, we show two typical prototype applications using speech and auditory interfaces. We believe they can be easily integrated into various ubiquitous smart systems and environments. For example, HRTF-based virtual auditory technology [6] has been recently used in enabling humanoid robots with sound localization ability.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China (60802085), the Program for New Century Excellent Talents in University supported by the Ministry of Education (MOE) of China, the Research Fund for the Doctoral Program of Higher Education in China (20070699015), the Natural Science Basic Research Plan of Shaanxi Province (2007F15) and the NPU Foundation for Fundamental Research (W018103).

## References

[1] W. M. Hartmann, "How we localize sound," *Physics Today*, pp. 24–29, November 1999.

[2] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A. W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," in *Proc. APSIPA ASC*, Sapporo, Japan, 2009.

[3] L. Xie and Z.-Q. Liu, "Realistic Mouth-Synching for Speech-Driven Talking Face Using Articulatory Modelling," *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 500–510, 2007.

[4] ——, "A Coupled HMM Approach for Video-Realistic Speech Animation," *Pattern Recognition*, vol. 40, no. 10, pp. 2325–2340, 2007.

[5] X. Tian and L. Xie, "An Experimental Comparison on KEMAR and BHead210 Dummy Heads for HRTF-based Virtual Auditory on Chinese Subjects," in *ICWMMN*, Beijing, China, 2010.

[6] J. Hörnstein, M. Lopes, and J. Santos-Victor, "Sound localization for humanoid robots – building audio-motor maps based on the HRTF," in *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2006, pp. 1170–1176.