Computer-Assisted Audiovisual Language Learning

Lijuan Wang, Yao Qian, Matthew R. Scott, Gang Chen, and Frank K. Soong, *Microsoft Research Asia*

Advances in speech-processing technology have enabled novel ways to learn a foreign language online. With Engkoo, researchers in China are working to turn any computer into a language learning assistant and make searching a language easier.

A confluence of trends has led to a rapid rise in the demand for learning English as a foreign language in East Asia. The region's economic emergence plays an important role since English is widely considered to be the lingua franca of business.[1] The Internet is also a significant factor because of the growth in e-learning, accelerated by increased mobile Web access, speed, and coverage.

The epicenter of the demand lies in China, where the world's fastest growing economy is also the nation with the largest number of both Internet users and English learners.[2] The massive demand for language tools, coupled with the ease of Web-based deployment-driven research (DDR), has created a unique opportunity for computer scientists to reimagine and rapidly experiment with new computer-assisted language learning at scale.

The DDR approach could lead to some interesting future scenarios. Imagine a child learning from his favorite TV star who appears to be personally teaching him English on his handheld device. Another youngster might use an avatar to tell mystery stories in a foreign language to her classmates. After an international meeting, a Chinese businessperson easily writes English summaries in which unfamiliar words can be input phonetically, that is, based on how the words sound.

The combination of massive-scale Web mining with two emerging speech-processing technologies—*talking head* and *phonetic similarity search*—has the potential to enable such scenarios. Microsoft Research Asia has successfully tested these technologies in Engkoo (www.engkoo.cn), a computer-assisted audiovisual language-learning service used by 10 million English learners in China each month that was the recipient of *The Wall Street Journal*'s 2010 Asian Innovation Readers' Choice Award.[3] Because it continuously crawls English/Chinese bilingual webpages, the system synchronizes with the latest terminology that people use on a daily basis.

SPEECH-PROCESSING TECHNOLOGIES

Talking head generates karaoke-style short synthetic videos demonstrating oral English. The videos consist of a photorealistic person speaking English sentences extracted from the Internet. The technology leverages a computer-generated voice with native-speaker-like quality and synchronized subtitles at the bottom of the video. To increase user engagement, it emulates popular karaoke-style videos specifically designed for a Chinese audience.

Compared to using prerecorded human voice and video in English education tools, talking head not only creates a realistic look and feel, but also greatly reduces the cost of content creation by generating arbitrary content sources synthetically and automatically. There is also opportunity for personalization. For example, users can choose a voice

 Published by the IEEE Computer Society

based on preferred gender, age, speaking rate, or pitch range and dynamics, and the system uses the selected type of voice to adapt a pretrained text-to-speech (TTS) option to customize the synthesized voice.

Phonetic similarity search can suggest similar-sounding word candidates according to the input text query guessed by the user. Inputting a word with its correct spelling is a common challenge for nonfluent English as a second language (ESL) learners when they are not sure about a word's pronunciation. This technology addresses the phonetic distance problem that correlates to a user's limited vocabulary and is customized based on regional language patterns. It quickly searches similar pronunciations in a large dictionary database, and greatly enhances the error correction suggestion capability that spell-checkers typically struggle with.

ENGKOO MOTIVATION

The introduction of the multimedia computer in the early 1990s was a major breakthrough for language teachers because it combined text, images, sound, and video in one device and permitted the integration of the four basic skills of listening, speaking, reading, and writing. Nowadays, as smartphones and tablet computers increasingly dominate the market, multimedia and multimodal language learning can be ubiquitous and more self-paced.

For foreign language users who do not have access to a personal tutor, learning correct pronunciation is an arduous task, primarily because using audio tapes, the most common method for learning pronunciation, is generally not engaging. In addition, it does not offer users complete instruction on how to move their mouths or lips to sound out phonemes (basic speech sound units) that might not exist in their mother tongue. Studies in cognitive informatics confirm that humans process information more efficiently when audio and visual techniques are used together.

Many researchers have successfully used visualized information and talking head technology to facilitate language learning. For example, Dominic Massaro[4] used visual articulation to show the mouth's internal structure, enabling learners to visualize the tongue's position and movement. Pierre Badin and colleagues[5] inferred learners' tongue positions and shapes to provide visual articulatory corrective feedback in second-language learning. Additionally, studies focused on overall pronunciation assessment and segmental/prosodic error detection have found that computer feedback helps language learners improve their pronunciation.[6]

Based on the belief that a lifelike assistant offers a more authoritative metaphor for engaging language learners, particularly among youth, Engkoo generates a photorealistic, lip-synced talking head. The long-term goal is to create a low-cost, multimodal, Web-based technology that can help users anywhere, anytime develop language skills with training that ranges from detailed pronunciation to conversational practice. Such a service is especially important for augmenting human teachers in areas of the world where native, high-quality instructors are scarce.

Figure 1. Screenshots of karaoke-like talking heads on Engkoo.

KARAOKE AS A MODEL

Karaoke, or KTV, is a favorite pastime in China, with numerous KTV clubs in major cities. Engkoo includes a karaoke-like feature that lets English learners practice their pronunciation online by mimicking a photorealistic talking head lip-syncing within a search and discovery ecosystem. This KTV function consists of videos generated from a vast set of sample sentences mined from the Web. Users can easily launch the videos with a single click on the sentence of their choosing. As Figure 1 shows, similar to the karaoke format, the videos display the sentence on the screen, while a model speaker says it aloud, teaching the users how to enunciate the words.

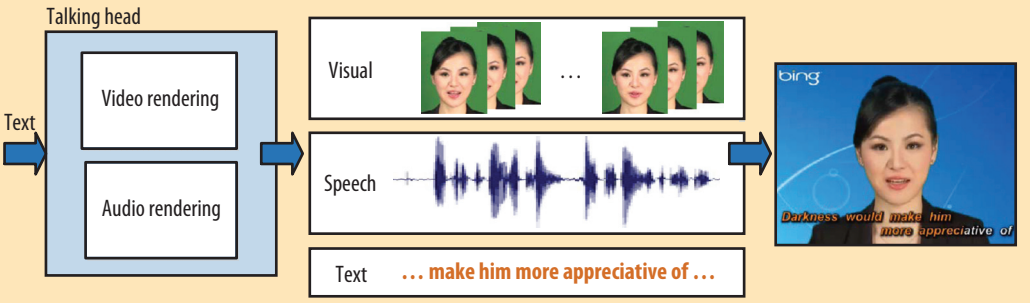

Figure 2. Talking-head synthesis technology in Engkoo.

While karaoke subtitles are useful, pacing is especially valuable when learning a language. The rhythm and prosody (pitch contour) embedded in the KTV function offer users the timing cues required to utter a given sentence properly. Although pacing can be learned by listening to a native speaker, this system uniquely offers the ability to get this content at scale and on demand.

Engkoo's KTV function offers a low-cost method for creating highly engaging, personalizable learning material utilizing state-of-the-art talking-head rendering technology. A key benefit is the generation of lifelike video as opposed to cartoon-based animations. This is important from a pedagogical perspective because it appears closer in nature to working with a human teacher, reducing the perceptive gap from the physical classroom to the virtual learning experience, which is particularly important for younger pupils.

The technology can drastically reduce language-learning video production costs in situations where the material requires a human native speaker. There is no need to repeatedly tape an actor speaking; instead, the system can synthesize the required audio and video content automatically. Teachers can also generate a talking head for students to take home and learn from, further bridging the classroom and e-learning scenarios.

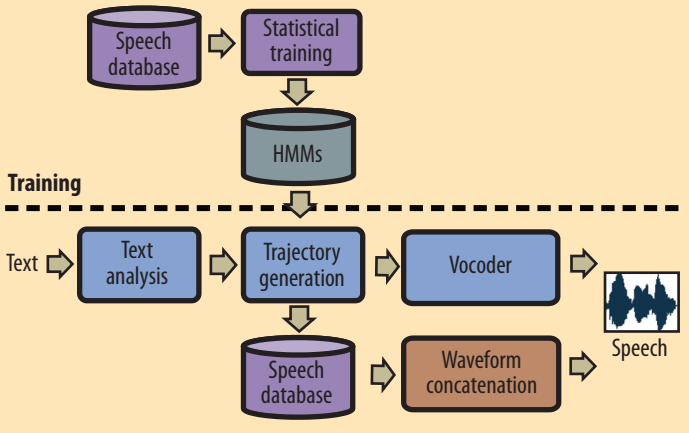

Figure 3. Text-to-speech (TTS) audio synthesis system.

TALKING HEAD

As Figure 2 shows, Engkoo's talking head captures shots of all the different pronunciations articulated by a speaker in roughly 30 minutes, along with simultaneously recorded speech. For any input text sentence, it synthesizes a speech signal using semantic, prosody, phonetic, and timing data. The technology then finds the best match between the lips' shape and what word the programmers want the model to say, creating an accurate lip-sync. Finally, it combines the synthesized audio, visual streams, and synchronized text into a video presentation in which the model speaker mouths the words of the sample sentence while a computerized voice reads it out loud.

Text-to-speech audio synthesis

Figure 3 shows talking head's text-to-speech audio synthesis system.

For a given input sentence, the text analysis module determines the sequence of phonemes and companion prosody by looking them up in a built-in pronunciation dictionary or by statistically predicting them in a "most probable" manner. For words with multiple pronunciations, the text analyzer takes into account the word's contextual information in a sentence—for example, whether "live" is a verb or adjective. Pronunciation of a foreign word, particularly the proper name of a person or place, that is not included in the internal dictionary must be guessed from its spelling through letter-to-sound (LTS) rules. Correct acronym pronunciations can also be challenging. For example, "IEEE" should be pronounced as "I-triple-E" instead of "I-E-E-E."

With the decoded phoneme sequence and the companion prosody pattern, the TTS system uses statistically trained hidden Markov models (HMMs) to generate and predict the sentence's corresponding speech parameter trajectories.

The system then uses the predicted acoustic trajectories to drive a digital-filter-based analysis/synthesis system or searches through the original training speech database for the appropriate

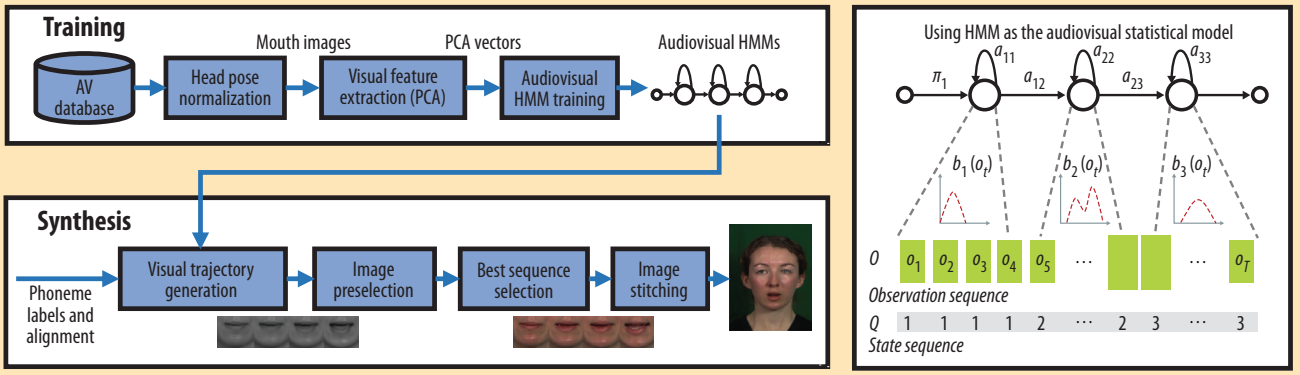

Figure 4. Lip-sync visual synthesis system.

waveform segments to concatenate the final speech output.

TTS is algorithmically language independent, and researchers can use different speech databases to adapt it to regional language patterns.

Lip-sync visual synthesis

Figure 4 shows talking head's lip-sync visual synthesis system.

Using the recorded audiovideo training data, talking head extracts, analyzes, and parameterizes all relevant mouth images, and forms a library of visual lips images and the corresponding audio speech signals. It automatically trains a statistical HMM to characterize how lips move to articulate different speech sounds. In synthesis, the trained statistical model can then predict the trajectory of lip movements for any given text input.

The trajectory serves as a guide to select a lip-synced, smooth mouth sequence from the image library. In each frame, the system selects the closest sample lips images in the library as nearest-neighbor candidates and, from among all candidates, finds an optimal, smooth lips-image sequence through a Viterbi search. It then stitches the mouth sequence to a background head video. The system can also render and stitch natural head motions and facial expressions into the video. The final output is a photorealistic talking head lip-synced with speech.

The entire process is data driven, fully automatic, and statistically model-based.

TRAJECTORY MODELING AND SYNTHESIS

As Figures 3 and 4 show, TTS and lip-sync consist of two parts: training and synthesis. These can be mathematically expressed as:

Training

$$\hat{\lambda} = \arg\max_{\lambda} p\left(O|W,\lambda\right), \text{ and}$$

Synthesis

$$\hat{o} = \arg\max_{o} p\left(o|w,\hat{\lambda}\right),$$

where λ are the model parameters, O the training data, W the transcriptions, o the synthesized speech, and w the input text.

TTS training and synthesis

Training in TTS is similar to that used in speech recognition, such that for the given training data (speech data and the corresponding transcriptions), the system iteratively adjusts HMM parameters to maximize the data likelihood. TTS extracts the speech parameters of short-time spectra—including their dynamic counterparts and excitation—and the fundamental frequency or pitch (F0) and its dynamic counterpart from the speech database. It models these parameters with HMMs depending on the speech's phonetic, linguistic, and prosodic contexts. Each HMM uses state duration models to capture an utterance's temporal structure. Consequently, the system models spectrum, excitation, and duration in a unified HMM framework.[7]

In speech synthesis, TTS first converts a given text to a context-dependent label sequence and concatenates the context-dependent HMMs according to the sequence. It then determines the state durations of the utterance HMM based on the state duration model. Next, the system generates the sequence of spectral and excitation parameters that maximize their output probabilities. Finally, TTS synthesizes speech signals from the generated spectral and excitation parameters using the corresponding speech synthesis filter.

Lip-sync training and synthesis

Talking head likewise uses HMMs to convert speech into visual sequences. As Figure 4 shows, audio and video are jointly modeled in HMMs that can generate visual trajectories.[8,9] The system uses these visual trajectories to create all lips animations. It uses principal component

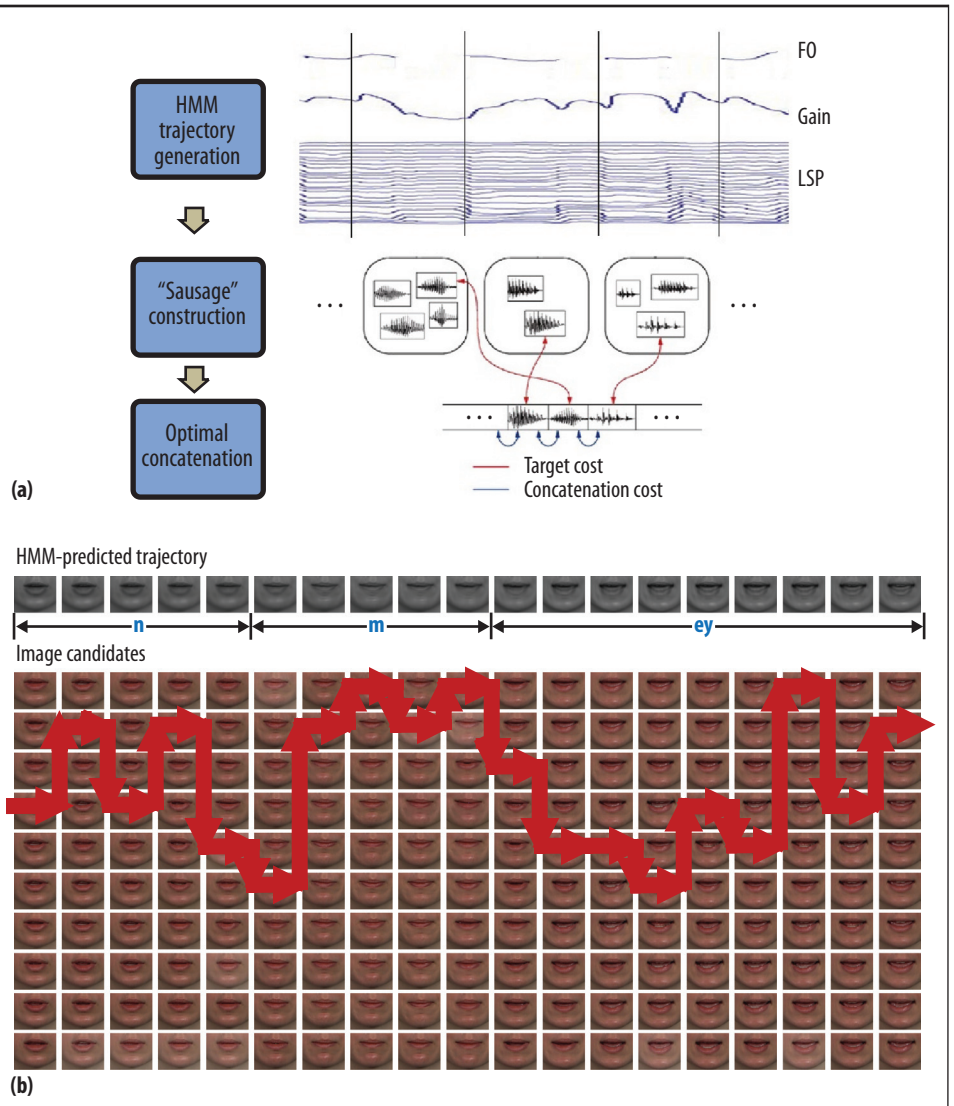

Figure 5. High-quality rendering via trajectory tiling in (a) TTS and (b) lip-sync.

analysis (PCA) to project the lips parameters to a low-dimensional subspace. Given the entire training database containing multiple visual trajectories, along with their contextual phonetic labels, the system trains statistical HMMs to capture the underlying visual patterns associated with the contextual speech labels. It trains context-dependent HMMs and applies tree-based clustering to acoustic and visual feature streams separately to improve the corresponding model's robustness.

In synthesis, lip-sync uses HMMs to predict and generate visual trajectories for any new phoneme sequence in a maximum probability sense. It then uses the predicted visual parameter trajectories to render the lips sequences.

TRAJECTORY TILING

Talking head uses HMM-based trajectory tiling to achieve high-quality TTS and photorealistic lip-sync rendering.

TTS rendering

As Figure 5a shows, in TTS the trajectory tiling algorithm has three stages: trajectory generation, construction, and concatenation.[10,11] First, well-trained statistical HMMs generate high-quality speech trajectories for three speech parameters: pitch (F0), gain (loudness), and short-time spectral envelope parameters of line spectral pairs (LSPs). The second step involves "tiling" the generated trajectories with appropriate speech segment candidates. The resulting candidates, along the time axis, form a "sausage"-like lattice. Third, using a dynamic programming procedure, the algorithm searches the lattice for the best path to find a sequence of tiles that minimizes the accumulated target cost (the tiles are close to the predicted trajectory) and the concatenation cost (the resultant path will induce low cost due to connection).

The final synthesized speech is of high quality in both intelligibility (the selected segment tiles sound like the correct target sounds) and naturalness (there are a minimum number of "glitches" at the concatenating points).

Lip-sync rendering

The predicted visual parameter trajectory by HMM is a compact description of articulator movements in the lower rank "eigen-lips" space. However, the lips image is blurred due to dimensionality reduction in PCA and the averaging of the maximum-likelihood-based model parameter estimation and trajectory generation. The blurring muffles the synthesis results and reduces the corresponding dynamic range for both TTS and synthesized moving lips. To alleviate this blurring effect, a trajectory-guided real sample selection approach searches for the closest real image sample sequence in the library to paste the predicted trajectory as the optimal solution.[12]

Figure 5b illustrates HMM-trajectory-guided sample selection. The top-line image sequence is the HMM-predicted visual trajectory, which is generated with low-dimensional PCA coefficients. The colored images at the bottom are the *K*-nearest real lips image samples in the lips image library closest to the rendered PCA lips image (in the low-rank PCA space). Among all candidates, the algorithm finds a smooth path via a Viterbi search. Thus, it reproduces the articulation movement in the visual trajectory to guide a selection of photorealistic lips sequences. In the final step, the algorithm stitches the best-matched mouth sequence to a background head video to create a natural, lip-synced, and photorealistic animation.

TALKING HEAD EVALUATION

For talking head to be suitable for audiovisual language learning, ESL users must find the synthesized speech natural-sounding and the model speaker's mouth movement in the videos close to that of a human teacher. We have evaluated the technology both objectively, by comparing synthesized results with the original recording (ground truth) using distortion metrics, and subjectively, by obtaining feedback from human subjects.

Blizzard and LIPS challenges

To compare talking head with other similar systems and to identify the upper bound in both naturalness and intelligibility, we entered the TTS and lip-sync technologies in two international contests, the 2009 LIPS Challenge[13] and the 2010 Blizzard Challenge.[11]

The LIPS Challenge was conducted as part of the Auditory-Visual Speech Processing Workshop, in which 20 native speakers with normal hearing and vision subjectively evaluated the rendering results of various contending systems in terms of audiovisual consistency. When each talking head video sequence was played together with the original speech, the viewer was asked to rate the naturalness of visual speech gestures (articulation movements in the lower face) with a five-point mean opinion score (MOS) ranging from 1 (poor) to 5 (excellent).

Our talking head received a MOS score of 4.15, the highest among all participants. The original, high-quality AV studio recording had a score of 4.8, which serves as an upper bound of talking-head rendering performance for this database.

The Blizzard Challenge is an international TTS contest sponsored by the Speech Synthesis Special Interest Group of the International Speech Communication Association. All participants receive common databases and evaluate the speech synthesized by the competing systems based on naturalness and intelligibility. Naturalness is measured by an MOS, while intelligibility testing requires listeners to transcribe semantically insensible sentences. At Blizzard 2010, our talking head achieved high scores in both naturalness and intelligibility.

Improving the ESL user experience

To improve the learning experience, we further refined the KTV function based on DDR feedback. First, we added KTV-style subtitles time-synchronously to the video and accompanying audio, as well as a countdown at the beginning to give users some preparation time. Second, we enhanced the videos with different backgrounds and beginning/ending gestures (smiling, laughing, head turning, and so on). Finally, because HMM-based trajectory modeling enables flexible control of the speaking rate in trajectory generation, we synchronously slowed down both the synthesized audio and video by about 1.5 times to a preferred speaking rate of ESL users.

Mouth Model Speaker Competition

The Engkoo team, along with popular video websites in China, held the English Mouth Model Speaker Competition in 2010 to recruit the first talking head speaker for the Chinese market. The competition went viral, with nearly 1,000 submissions receiving more than 70 million views by 12 million unique users. This traffic benefited our research by producing more feedback for analysis, including both implicit data via usage trends like drop-off and explicit data via direct messaging.

PHONETIC SIMILARITY SEARCH

In a Web search, misspelled queries can lead to poor search results. To improve the search completion rate, many search engines attempt to detect and correct misspelled queries by automatically suggesting likely candidates.

Spelling errors are common, particularly for complex or unusual words, and are due to either mistyped keystrokes (typographic errors) or the user's lack of knowledge of the correct spelling (cognitive errors).[14] Examples of typographical errors are "betwen" (missing keystroke) and "bewteen" (swapped keystrokes) for the word "between." Examples of cognitive errors due to mispronunciation are "sheap" for "ship" and "reed" for "read." Cognitive errors occur more frequently in person and place names, especially in foreign or transliterated foreign words.

Misspelled queries due to cognitive errors are more difficult to correct than typical typographical errors. The latter are generally easier to fix with a conventional spell-checker. People in nonnative-English-speaking regions tend to use the transliterated, romanized spelling of words for technical jargon or proper names that are of English origin.

To detect input errors due to phonetic mispronunciation, Engkoo uses an efficient phonetic candidate generator to produce phonetically similar candidates for a given

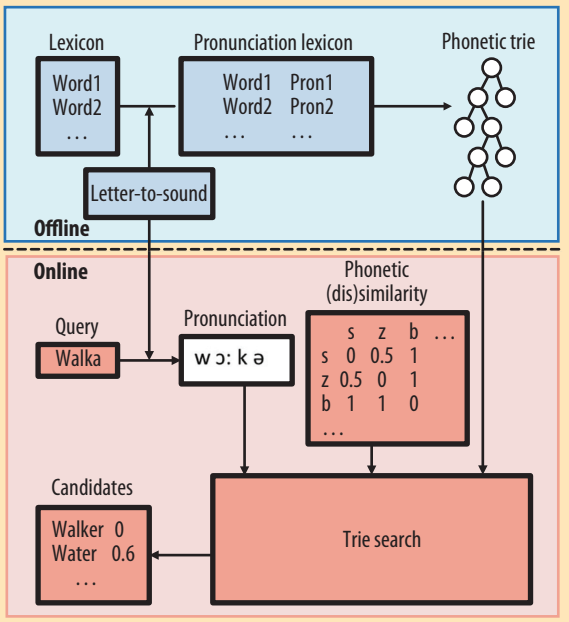

Figure 6. To detect input errors due to phonetic mispronunciation, Engkoo generates phonetically similar candidates for a given query.

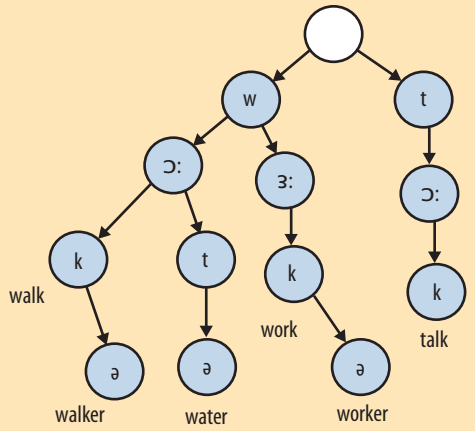

Figure 7. Example of a phonetic trie. A node in the tree stores a phoneme (key), all the descendants of a node have a common prefix of the phoneme string, the root node is associated with the empty string, and the values are associated with terminal leaves.

query.[15] Figure 6 shows the generator's offline and online components.

The system uses a lexicon (dictionary) for simple spell-checking. This consists of a list of word spellings augmented when possible with correct pronunciations; for words without available pronunciations, an LTS module "guesses" the pronunciation. The system compiles the resulting pronunciation lexicon into a phonetic "trie," an ordered prefix tree, and constructs an acoustic-phonetic similarity table with a speech database of many different speakers. It can adapt this table to different user populations to accommodate common phonetic mistakes.

In searching for phonetically similar candidates of a given input query, the system first uses LTS to generate the most probable phoneme sequence. It then computes phonetic similarity scores for each entry in the prestored pronunciation lexicon. Finally, the system outputs candidates that are phonetically most similar to the query.

LTS conversion

A key component in phonetic candidate generation is LTS conversion, which converts a letter sequence into a phoneme sequence.[16] For a language such as Spanish, where a letter string can map onto a corresponding phoneme string rather regularly, LTS can be easily derived using simple rules. However, for English, where such a mapping is less deterministic and simple grammatical rules cannot predict the conversion well, it is necessary to construct (train) the rules statistically with many examples of paired letters and mapped phonemes.

The pronunciation lexicon typically contains a paired letter sequence and phoneme sequence. The length of a letter sequence is, in general, different from its corresponding phoneme sequence. It is then necessary to align letters and phones before feeding them into a machine learning algorithm to train the statistical mapping rules. The letter-to-phoneme alignment is nontrivial and arguably not unique due to its one-to-one, many-to-one, or many-to-many possibilities. However, it is still possible to train such rules optimally in the "most likely" sense.

KLD-based phonetic similarity measurement

The phonetic similarity between two phoneme strings can be defined as an edit distance when one string is converted (edited) into another string with accumulated insertions, deletions, or substitutions. The insertion, deletion, and substitution costs between a phoneme and other phonemes, including a "silence" phoneme, is measured by the Kullback-Leibler divergence (KLD) distortion[17] between the statistical distributions of corresponding HMM models.

KLD, or relative entropy, is an information-theoretic measure of (dis)similarity between two probability distributions. If M and $\tilde{M}$ represent two models with continuous probability distributions, the symmetric KLD between them can be computed as

$$D_s\left(M\|\tilde{M}\right)=D\left(M\|\tilde{M}\right)+D\left(\tilde{M}\|M\right),$$

where

$$D_s\left(M\|\tilde{M}\right)=\int_{R^N}P(x|M)\log\frac{P(x|M)}{P\left(x|\tilde{M}\right)}dx$$

Phonetic trie

A structured, easy-to-search, and compact data structure is needed to efficiently generate phonetic candidates. Figure 7 shows an example of a phonetic trie, in which the keys along a path are phoneme strings and the values are letter strings. A node in the tree stores a phoneme (key), all the descendants of a node have a common prefix of the phoneme string, the root node is associated with the empty string, and the values are associated with terminal leaves.

A dynamic programming algorithm searches along a phonetic trie by measuring the phonetic distortions between a given query's phoneme string and the phoneme strings of different paths in the trie. The algorithm performs the search efficiently by pruning out any unfinished branches when the accumulated KLD exceeds a preset threshold.

Phonetic similarity search examples

Figure 8 shows screenshots of phonetic similarity searches on Engkoo.

The first query is input as "shampin," a common Chinese transliteral way of pronouncing the word "champagne." Although a typical spell-checker cannot generate the correct candidate, a phonetic similarity search finds the correct word at the top of the Sounds Like list. All the other candidates—"jumping," "shampoo," and "shopping"—sound very similar to the query.

The phonetically similar candidates generated by Engkoo for the second example query, "fiziks," are also satisfactory: the correct word "physics" was the top choice, and all other candidates are phonetically plausible.

The third query example is "randevu." The intended French-origin word, "rendezvous," is difficult for a non-French speaker to spell correctly. A phonetic similarity search produces the correct word as the top choice along with other similar-sounding candidates.

Figure 8. Screenshots of phonetic similarity searches on Engkoo.

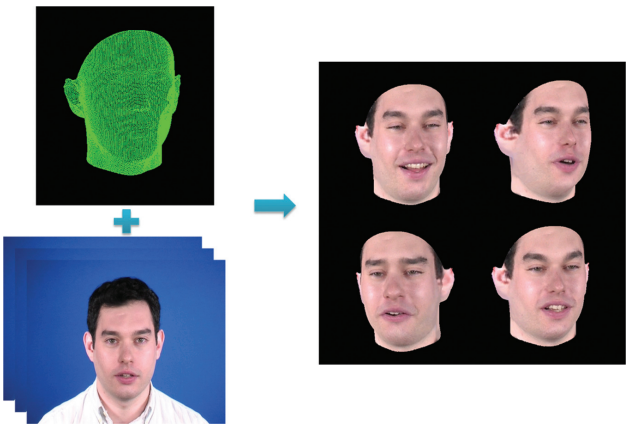

Figure 9. A 3D photorealistic talking head created by combining 2D image samples with a 3D face model.

FUTURE RESEARCH

Engkoo's current karaoke function can be enhanced to reach our long-term goal of creating a low-cost, Web-based, lifelike computer assistant that is useful in many language-learning scenarios ranging from interactive pronunciation drills to conversational training.

Toward that end, we have developed a new 3D photorealistic, real-time talking head with a personalized appearance.[18] First, we record approximately 20 minutes of audiovisual 2D video with prompted sentences uttered by a human speaker. As Figure 9 shows, we then use a 2D-to-3D reconstruction algorithm to automatically "wrap" the 3D geometric mesh with 2D video frames and construct a training database. Next, we form superfeature vectors consisting of 3D geometry, texture, and speech to train a statistical, multistreamed HMM, and then use this model to synthesize both the geometry animation and dynamic texture trajectories.

With regard to talking head's synthesized audio (speech) output, we are working to make the TTS system more personalized, adaptive, and flexible. For example, we have created and tested an algorithm that can teach the talking head to speak authentic English sentences that sound like a Chinese ESL learner.[19] It also would be helpful if the TTS system could synthesize more natural and dynamic prosody patterns for ESL learners to mimic.

The system can control the 3D talking head animation via rendered geometric trajectories, while it renders facial expressions and articulation movements with dynamic 2D

image sequences. The system also can control head motions and facial expressions by separately manipulating their corresponding parameters. These capabilities make it possible to create photorealistic talking heads using video recordings of movie stars or other types of celebrities.

Phonetic similarity search can be improved by collecting more text and speech data to generate phonetic candidates that include generic and localized spelling and pronunciation errors committed by language learners at different levels. Such a database will make it possible to discriminatively train a more powerful LTS module to predict and correct errors.

Other future work will focus on increasing the computer assistant's interactivity, enabling it to hear (via speech recognition) and speak (via TTS synthesis), read and compose, correct and suggest, or even guess the learner's intention.

E-learning is an emerging worldwide trend, catalyzed by the increasing availability of Internet access. Pioneering content providers include universities such as MIT and Stanford, which offer college- and graduate- level online education free of charge, and Khan Academy, which provides thousands of free online K-12 educational videos. But still missing is a service that automatically generates content for language learning and evolves with the Web.

The Engkoo service uses novel Web mining techniques and exposes its content through advanced features based on natural-language and speech-processing technologies. Engkoo leverages talking head and phonetic similarity search to facilitate language learning related to listening to, speaking, and writing words the way they sound. These technologies have been enhanced iteratively through deployment-driven research: release, incorporate implicit and explicit user feedback, improve, and redeploy. The use of these technology components and the models developed to improve them has been extremely successful, as evidenced by an exponential increase in traffic and consistently positive feedback from millions of Chinese ESL users.

Acknowledgments

The authors thank Yuki Arase, Xianjun Huang, Henry Li, Mu Li, Xiaohua Liu, Hao Wei, Weijiang Xu, Dongdong Zhang, and Ming Zhou of Microsoft Research Asia for their contributions to the research and development of the Engkoo project.

References

1. B. Seidlhofer, "Common Ground and Different Realities: World Englishes and English as a Lingua Franca," *World Englishes*, June 2009, pp. 236-245.
2. L.J. Zhang, R. Rubdy, and L. Alsagoff, "Englishes and Literatures-in-English in a Globalised World," *Proc. 13th Int'l Conf. English in Southeast Asia* (ESEA 08), National Inst. of Education, Singapore, 2008, pp. 42-58.
3. M.R. Scott, X. Liu, and M. Zhou, "Towards a Specialized Search Engine for Language Learners," *Proc. IEEE*, Sept. 2011, pp. 1462-1465.
4. D.W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press, 1998.
5. P. Badin et al., "Visual Articulatory Feedback for Phonetic Correction in Second Language Learning," *Proc. Workshop Second Language Learning Studies: Acquisition, Learning, Education and Technology* (L2WS 10), Int'l Speech Comm. Assoc., 2010; www.isca-speech.org/archive/L2WS_2010/papers/lw10_P1-10.pdf.
6. M. Eskenazi, "An Overview of Spoken Language Technology for Education," *Speech Comm.*, Oct. 2009, pp. 832-844.
7. K. Tokuda et al., "Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis," *Proc. 2000 IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 00), IEEE, 2000, pp. 1315-1318.
8. S. Sako et al., "HMM-Based Text-to-Audio-Visual Speech Synthesis," *Proc. 6th Int'l Conf. Spoken Language Processing* (ICSLP 00), Int'l Speech Comm. Assoc., 2000, pp. 25-28.
9. L.J. Wang et al., "Synthesizing Visual Speech Trajectory with Minimum Generation Error," *Proc. 2011 IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 11), IEEE, 2011, pp. 4580-4583.
10. Z.-J. Yan, Y. Qian, and F.K. Soong, "Rich-Context Unit Selection (RUS) Approach to High Quality TTS," *Proc. 2010 IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 10), IEEE, 2010, pp. 4798-4801.
11. Y. Qian et al., "An HMM Trajectory Tiling (HTT) Approach to High Quality TTS," *Proc. Blizzard Challenge 2010 Workshop, Language Technologies Inst.*, Carnegie Mellon Univ., 2010; http://festvox.org/blizzard/bc2010/MSRA_%20Blizzard2010.pdf.
12. L.J. Wang et al., "Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection," *Proc. 11th Ann. Conf. Int'l Speech Comm. Assoc.* (Interspeech 10), Int'l Speech Comm. Assoc., 2010, pp. 446-449.
13. B.-J. Theobald et al., "LIPS2008: Visual Speech Synthesis Challenge," *Proc. 9th Ann. Conf. Int'l Speech Comm. Assoc.* (Interspeech 08), Int'l Speech Comm. Assoc., 2008, pp. 2310-2313.
14. K. Kukich, "Techniques for Automatically Correcting Words in Text," *ACM Computing Surveys*, Dec. 1992, pp. 377-439.
15. B. Peng et al., "A New Phonetic Candidate Generator for Improving Search Query Efficiency," *Proc. 12th Ann. Conf. Int'l Speech Comm. Assoc.* (Interspeech 11), Int'l Speech Comm. Assoc., 2011, pp. 1117-1120.
16. D. Wang and S. King, "Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields," *IEEE Signal Processing Letters*, Feb. 2011, pp. 122-125.
17. P. Liu and F. Soong, *Kullback-Leibler Divergence Between Two Hidden Markov Models*, tech. report, Microsoft Research Asia, 2005.
18. L.J. Wang, W. Han, and F.K. Soong, "High Quality Lip-Sync Animation for 3D Photo-Realistic Talking Head," *Proc. 2012 IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 12), IEEE, 2012, pp. 4529-4532.

19. Y. Qian, J. Xu, and F.K. Soong, "A Frame Mapping Based HMM Approach to Cross-Lingual Voice Transformation," *Proc. 2011 IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 11), IEEE, 2011, pp. 4568-4571.

Lijuan Wang *is a researcher in the Speech Group at Microsoft Research Asia. Her research interests include talking head technology, speech synthesis, and audio-visual signal processing. Wang received a PhD in electrical engineering from Tsinghua University, China. She is a member of IEEE. Contact her at lijuanw@microsoft.com.*

Yao Qian *is a researcher in the Speech Group at Microsoft Research Asia. Her research interests include speech and singing voice synthesis, speech recognition, voice transformation, and computer-assisted language learning. Qian received a PhD in electronic engineering from the Chinese University of Hong Kong. She is a member of IEEE. Contact her at yaoqian@microsoft.com.*

Matthew R. Scott *is a senior development lead in the Innovation Engineering Group at Microsoft Research Asia. His research interests include computer-assisted language learning, natural language processing, and human-computer interaction. Scott received a BA in computer science from Boston University. He is a member of IEEE. Contact him at mrscott@microsoft.com.*

Gang Chen *is a development manager in the Innovation Engineering Group at Microsoft Research Asia. His research interests include computer-assisted language learning, natural language processing, human-computer interaction, computer graphics, and image processing. Chen received an MS in computer science from the University of Science and Technology Beijing. He is a member of IEEE. Contact him at gangch@microsoft.com.*

Frank K. Soong *is a principal researcher in the Speech Group at Microsoft Research Asia. His research interests include speech analysis, perception, coding, recognition, and synthesis. Soong received a PhD in electrical engineering from Stanford University. He is an IEEE Fellow. Contact him at frankkps@microsoft.com.*