Minnesang: Speak Medieval German

Daniel Spelmezan
Media Computing Group
RWTH Aachen University
52056 Aachen, Germany
spelmezan@cs.rwth-aachen.de

Jan Borchers
Media Computing Group
RWTH Aachen University
52056 Aachen, Germany
borchers@cs.rwth-aachen.de

CHI 2006, April 22–27, 2006, Montréal, Québec, Canada.
ACM 1-59593-298-4/06/0004.

Abstract

We present a prototype of the Minnesang exhibit that translates visitors' utterances into medieval German in their own voice. This lets visitors experience how they would have spoken in medieval times. The project illustrates new variants of voice conversion and their use in human-computer interaction.

Keywords

Cross-language voice conversion, voice imitation, interactive exhibit, medieval language

ACM Classification Keywords

C.3 Special-purpose and application-based systems: Signal processing systems. H.5.2 User Interfaces: Voice I/O. H.5.5 Sound and Music Computing: Signal analysis, synthesis, and processing.

Introduction

The Regensburg Experience (REX) is a visitor center in Regensburg, Germany, documenting the city's rich medieval history (see http://www.rex-regensburg.de/). Opening in late 2006, REX will highlight various aspects of medieval life, from architecture and science to literature and music, in interactive exhibits. One of them is the Minnesang exhibit that will let visitors learn about the medieval art form of "Minnesang" love poetry as it was practiced in Regensburg.

For a more interactive experience, we decided we wanted to give visitors the opportunity to hear themselves speak medieval German, as in the following scenario:

Jane walks up to a large display that shows a verse of a medieval Minnesang poem in German, her native language. She reads the verse out aloud. Promptly, a video of her appears on the display: Jane sees and hears herself reciting the poem in medieval German.

Note that this scenario requires two kinds of transformation to happen: First, the text Jane spoke needs to be translated into medieval German. This would normally require speech recognition, translation, and speech synthesis. However, we already know what needs to be said, and can use a recording of a professional speaker reciting that verse in medieval German—essentially ignoring *what exactly* Jane said.

The second transformation, however, is much more difficult: The professional recording needs to be modified to *sound like Jane's voice.* In other words, Jane's voice characteristics—*how* she said the verse—need to be imprinted onto the existing medieval spoken audio signal. This *voice conversion*—modifying a source speaker's voice (the pro in our scenario) to sound like a target speaker's voice (Jane) when saying the same thing—is a field of very active research and the focus of much of this paper.

Related work

Voice conversion changes the characteristics of a speaker's voice. This technique has many applications in voice output systems. It can generate different voices in text-to-speech synthesis systems or produce a convincing movie dub of an actor's voice. A criminal, however, can conceal the identity of his voice for deceitful purposes. For more information on voice conversion and its application areas see [3].

Mathematically speaking, voice conversion requires computing a transformation function that converts the source speaker's voice to the target speaker's voice. A voice carries many unique characteristics, such as speech tempo, pauses, rhythm, dialect, accentuation and pronunciation of certain words, but also pitch frequency, formant structure, and other vocal tract characteristics. These factors are important to closely imitate a speaker's voice [2], but not all of them are required to merely identify a voice. Average pitch frequency, formant structure, and some characteristics of the vocal tract can already determine speaker identity. This greatly simplifies the task of voice conversion and is what most voice conversion systems do [7].

Almost all procedures for voice conversion assume that sufficient training data exists in the form of a parallel speech corpus from the source and target speakers (i.e., both speakers utter exactly the same sentence in the same language) to estimate the transformation function [3,12]. When this parallel speech corpus is not available (i.e., the two speakers utter different sentences), the relationship between the source and target voices cannot be captured in as much detail because the speech frames do not correspond. This case is known as text-independent or non-parallel voice conversion. Cross-language voice conversion is even more challenging due to inter-language differences between the source and target speech besides inter-speaker differences [1].

In our scenario, we can use parallel training if our professional speaker also speaks Jane's language (German), and if we have recorded the speaker reciting the verse in modern German. If Jane were English, we would have to use non-parallel, cross-language training, using her English utterance as training data to imprint her voice onto the speaker's medieval German sentences.

Only few approaches exist that deal with text-independent and cross-language voice conversion. Ye and Young [6] use a speech recognizer to index the utterance of an unknown source speaker and select similar speech frames from a large database of an already indexed target speaker. Kumar and Verma [9,10] focus on creating a set of descriptors of a person's voice and convert the phonemes of the source speaker to the corresponding phonemes of the target speaker. Mashimo et al. [11] use a parallel corpus for training and apply the conversion function to an utterance in a different language. Mouchtaris et al. [8] adapt the conversion parameters for a given pair of source and target speakers, for which a parallel corpus exists, to the non-parallel corpus of a different pair of speakers. Sündermann et al. [4,5] define artificial phonetic classes in both the source and target utterance and determine for each source class the most similar target class. The class mapping serves as the basis for the transformation.

Common to all approaches is that they need a large training database from the target speaker to accurately model the target speaker's voice. Only [4,5,8] also try to address the case with sparse training data.

Design

We tried to meet the following requirements with our exhibit: (1) The visitor utters only a short verse from the poem in her native language. (2) The conversion finishes ideally in real-time to provide the result immediately. (3) The converted medieval German utterance sounds like the target speaker, natural, and is free of distortions. (4) We can support different target speaker languages to address visitors from different countries.

To better understand the particular way we are using voice conversion, let us first look at the more traditional ways it is performed.

Voice conversion is usually applied in text-to-speech synthesis or speech-to-speech translation systems. Figure 1 illustrates voice conversion for text-to-speech synthesis, which assumes that a large parallel database for off-line training exists.

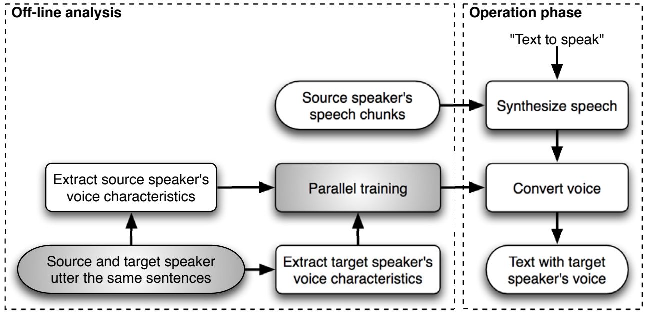
Off-line analysis
Operation phase
"Text to speak"
Source speaker's speech chunks
Synthesize speech
Extract source speaker's voice characteristics
Parallel training
Convert voice
Source and target speaker utter the same sentences
Extract target speaker's voice characteristics
Text with target speaker's voice

Figure 1: Voice conversion for text-to-speech synthesis uses off-line parallel training on equivalent utterances.

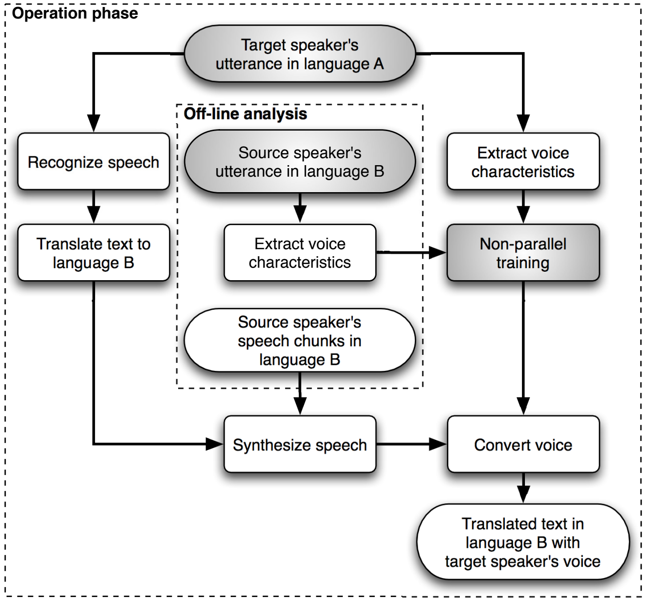
Operation phase
Target speaker's utterance in language A
Off-line analysis
Recognize speech
Source speaker's utterance in language B
Extract voice characteristics
Translate text to language B
Extract voice characteristics
Non-parallel training
Source speaker's speech chunks in language B
Synthesize speech
Convert voice
Translated text in language B with target speaker's voice

Figure 2: Voice conversion for speech-to-speech translation uses real-time non-parallel training between different languages.

Speech-to-speech translation systems, as illustrated in Figure 2, produce the sentence uttered by the target speaker in language A in his own voice after the sentence has been translated to and synthesized in the listener's language B. The source speaker's voice—the standard voice for speech synthesis in language B—is analyzed off-line. Non-parallel training on few target speaker words takes place in the operation phase. The transformation function refines as the target speaker continues to speak. Compared to text-to-speech synthesis, the achieved conversion quality is still unsatisfactory because the transformation function is too rough.

Figure 3 explains voice conversion as applied in our Minnesang exhibit. The visitor is the target speaker. She reads and utters only four sentences of a poem in her native language A, such as German. The professional speaker, our source speaker, speaks medieval German (language B). We have to use non-parallel training to estimate the conversion function. In some cases, however, the source speaker also speaks the target speaker's language A. In those cases we can combine parallel and non-parallel training to improve the conversion quality.

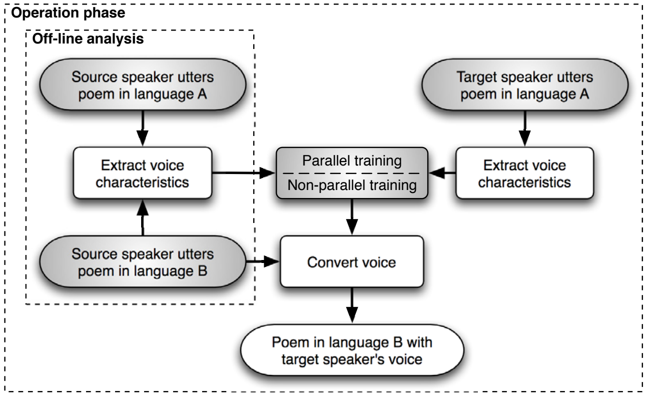
Operation phase
Off-line analysis
Source speaker utters poem in language A
Target speaker utters poem in language A
Extract voice characteristics
Parallel training
Non-parallel training
Extract voice characteristics
Source speaker utters poem in language B
Convert voice
Poem in language B with target speaker's voice

Figure 3: Voice conversion for our system uses real-time non-parallel training on few utterances. If the source speaker speaks language A, parallel training can be applied.

Prototype and implementation

We built a first prototype to investigate the possible interaction with our Minnesang system. Because the algorithms we had available at that time were not able to perform voice conversion in this direction, and with this small training database, with satisfactory quality, this system did not feature live voice conversion yet. Instead, it displayed text in German or English, and contained prerecorded verses in medieval German spoken by those people who were demoing the exhibit. This Wizard-of-Oz study helped us to understand how visitors reacted to the general idea of this exhibit before the technology was ready. Figure 4 shows our colleague in front of the prototype. The console hosts the computer, microphone, and the projector that projects the poem and the user's image onto a large lute-shaped screen that includes a small camera. The user first presses a button to receive on-screen instructions in his language, then presses a red button on the desk to record the poem in his voice.

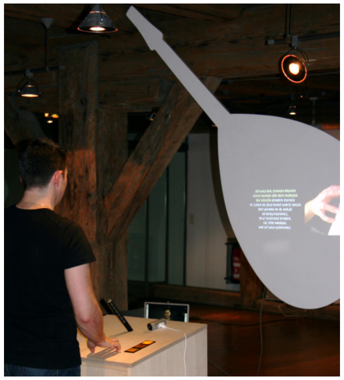

Figure 4: A user interacts with the prototype.

We presented this prototype to the public in Regensburg during a preview in July 2005. Two trained users interacted with the prototype as described in our scenario (see video at http://www.rex-regensburg.de/fileadmin/images/Video/Minnesang.wmv).

The following sentences in medieval German, German, and English represent a line from the medieval poem:

- Wie sol man rehte triuwe gerehticlîch erkennen?
- Woran erkennt man wahrhaft wahre Treue?
- How can one truly recognize a love that's true?

Evaluation

The Wizard-of-Oz prototype worked successfully, and we received positive and helpful informal feedback from the visitors watching the demonstrations. They were fascinated by the possibility to hear their own voice speaking medieval German. To improve our system, they suggested the addition of medieval music and medieval scenery to the video. Visitors would also like to take their poem recording with them as a movie.

Summary and future work

We introduced our ongoing work on an interactive exhibit that will allow visitors to speak a medieval language. Our project illustrates a new variant of voice conversion and how it can be used in human-computer interaction. We built a prototype system and successfully presented it to the public.

Our next steps are to overcome the problem of limited vocabulary size to build the voice profile and to create satisfactory voice renditions even in those cases where parallel training is not possible because the visitor’s native language is not part of our repertoire. We hope that this work raises the interest in voice conversion as a new tool in HCI and inspires other novel applications areas.

Acknowledgements

We would like to thank Nils Beck for his invaluable help in programming the software framework, assembling the prototype, and for his inspiring ideas. We also thank Peter "Gecko" Götz for building the exhibit hardware, including an insanely large lute.

References

[1] Abe, M., Shikano, K., Kuwabara, H. Cross Language Voice Conversion. In *Proc. ICASSP 1990*.

[2] Zetterholm, E. A case study of successful voice imitation. In *Logopedics Phoniatrics Vocology 2002*.

[3] Suendermann, D. Voice Conversion: State-of-the-Art and Future Work. In *Proc. DAGA 2005*.

[4] Suendermann, D., Bonafonte, A., Ney, H., and Hoege, H. A First Step Towards Text-Independent Voice Conversion. *In Proc. ICSLP 2004*.

[5] Suendermann, D., Ney, H., and Hoege, H. VTLN-Based Cross-Language Voice Conversion. *In Proc. ASRU 2003*.

[6] Ye, H. and Young, S. Voice Conversion for Unknown Speakers. In *Proc. ICSLP 2004*.

[7] Ye, H. and Young, S. High Quality Voice Morphing. In *Proc. ICASSP 2004*.

[8] Mouchtaris, A., Van der Spiegel, J., and Mueller, P. Non-Parallel Training for Voice Conversion by Maximum Likelihood Constrained Adaptation. In *Proc. ICASSP 2004*.

[9] Kumar, A. and Verma, A. Using phone and diphone based acoustic models for voice conversion: a step towards creating voice fonts. In *Proc. ICASSP 2003*, 720-723.

[10] Verma, A. and Kumar, A. Articulatory class based spectral envelope representation for voice fonts. In *Proc. International Conference on Multimedia and Expo 2004*, 1647-1650.

[11] Mashimo, M., Toda, M., Shikano, K., and Campbell, N. Evaluation of Cross-Language Voice Conversion Based on GMM and STRAIGHT. In *Proc. Eurospeech 2001*, 361-364.

[12] Stylianou, Y., Cappé, O., and Moulines, E. Continuous Probabilistic Transform for Voice Conversion. In *IEEE Transactions on Speech and Audio Processing 1998*, Vol. 6, No. 2, 131-142.