Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 274–293

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Statistical conversion of silent articulation into audible speech using full-covariance HMM☆

Thomas Hueber[a,b,*], Gérard Bailly[a,b]

[a] *Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France*
[b] *CNRS, GIPSA-Lab, F-38000 Grenoble, France*

Received 1 July 2014; received in revised form 20 February 2015; accepted 15 March 2015
Available online 3 April 2015

Abstract

This article investigates the use of statistical mapping techniques for the conversion of articulatory movements into audible speech with no restriction on the vocabulary, in the context of a silent speech interface driven by ultrasound and video imaging. As a baseline, we first evaluated the *GMM-based mapping considering dynamic features*, proposed by Toda et al. (2007) for voice conversion. Then, we proposed a 'phonetically-informed' version of this technique, based on *full-covariance HMM*. This approach aims (1) at modeling explicitly the articulatory timing for each phonetic class, and (2) at exploiting linguistic knowledge to regularize the problem of silent speech conversion. Both techniques were compared on continuous speech, for two French speakers (one male, one female). For modal speech, the HMM-based technique showed a lower spectral distortion (objective evaluation). However, perceptual tests (transcription and XAB discrimination tests) showed a better intelligibility of the GMM-based technique, probably related to its less fluctuant quality. For silent speech, a perceptual identification test revealed a better segmental intelligibility for the HMM-based technique on consonants.

Keywords: Silent speech interface; GMM; HMM; Ultrasound; Articulatory–acoustic mapping

1. Introduction

Silent Speech Interfaces (SSIs) have emerged as a new research field in the last few years (Denby et al., 2010). SSI can be defined as devices that enable oral speech communication without vocalization. With a SSI, the 'silent' speaker articulates normally but does not produce any sound. SSI could be used to preserve the privacy of conversations, for discreet hand-free communication (as in a military operation), or on the contrary, in very noisy environments (where the audio speech signal is too degraded). Since silent speech does not involve vocal folds vibration, SSI could potentially be used after a total laryngectomy, as a temporary alternative to the esophageal voice, which takes time to master, or to the tracheoesophageal voice, which may require an additional surgery. So far, different technologies have been proposed to capture the articulatory activity during silent speech, such as surface electromyography (sEMG) with sensors placed on the face and neck (Schultz and Wand, 2010), or permanent-magnetic articulography (PEMA)

☆ This paper has been recommended for acceptance by S. Narayanan.

* Corresponding author at: GIPSA-lab, 11 rue des Mathématiques, 38402 Saint Martin d'Hères, France. Tel.: +33 4 76 57 49 40; fax: +33 4 76 57 47 10.
E-mail address: thomas.hueber@gipsa-lab.fr (T. Hueber).

http://dx.doi.org/10.1016/j.csl.2015.03.005
0885-2308/

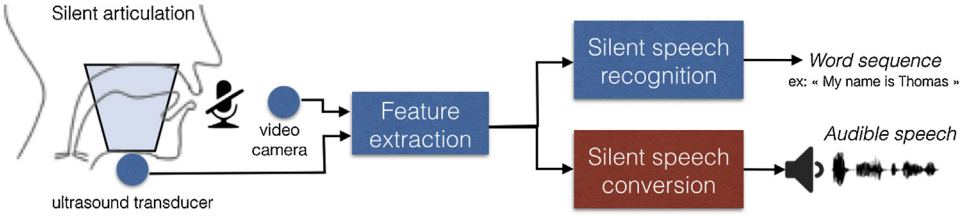

Fig. 1. Silent speech interface driven by ultrasound and video imaging. The present study focuses on the direct conversion of silent articulation into audible speech without any restriction on the vocabulary size (contrary to silent speech recognition).

with magnets glued on the tongue and lips (Fagan et al., 2008). Another approach is to capture and post-process a so-called Non-Audible-Murmur (NAM) using a stethoscopic microphone (Nakajima et al., 2003). In our approach (Denby et al., 2006; Hueber et al., 2010b), articulatory movements are captured using a medical ultrasound transducer placed beneath the chin, and a video camera in front of the lips, as shown in Fig. 1. This sensor set provides relatively complete information on tongue (via ultrasound), lips and jaw movements,[1] while remaining totally non-invasive.

Several studies addressed the problem of *silent speech recognition*, i.e. the identification of a sequence of words from silent articulation: (Wand and Schultz, 2011) for sEMG, (Nakajima et al., 2006) for NAM, (Gilbert et al., 2010) for PEMA and (Hueber et al., 2009) for ultrasound. In this study, we addressed the problem of *silent speech conversion*, i.e. the direct reconstruction of the speaker's voice from his/her silent articulation, without any restriction on the vocabulary size.

In our previous work (Hueber et al., 2010b), this problem was addressed using a 'recognition-followed-by-synthesis' approach. The system was composed of two chained modules: (1) a HMM-based decoder that predicts the most likely phonetic sequence from the observed articulatory movements, and (2) a unit selection algorithm that generates the spectral trajectories from the decoded phonetic sequence. The intermediate phonetic decoding step was motivated by the introduction of linguistic knowledge to regularize the problem of silent speech conversion. Such information might help recover some of the missing information in the silent articulatory data, such as the voicing feature. This approach gave encouraging results but presented some drawbacks. First, the quality of the synthesis depended strongly on the performance of the phonetic decoding: an error during that step systematically corrupted the synthesis. Second, since articulatory and acoustic modalities were processed separately during training, the dependencies between articulatory and spectral features were not explicitly modeled. As a consequence, the spectral targets depended on the decoded phonetic labels only, and did not take into account the articulatory variability within each phonetic class. Therefore the first goal of this new study was to investigate mapping techniques that should be able to explicitly model these local acoustic–articulatory relationships.

Furthermore, in all our previous studies, articulatory-to-acoustic mapping was not performed on actual silent speech: the converted articulatory data were acquired while the speaker was still vocalizing. However, recent studies such as (Hueber et al., 2010a) and (Janke et al., 2010) suggested that silent speech articulation differs from that of modal speech, probably due to the lack of acoustic feedback. Therefore, the second goal of this new study was to evaluate our system on actual silent speech.

The problem of speech synthesis from articulatory movements, commonly called "articulatory synthesis" has been originally addressed by the use of a two or three-dimensional articulatory model of the vocal tract (Birkholz et al., 2006; Maeda, 1990), coupled with an acoustic simulation method (Sondhi and Schroeter, 1987). In the past few years, supervised machine learning techniques have brought significant improvements in articulatory-to-acoustic mapping. These techniques seem to be well adapted to tackle the non-uniqueness and the non-linear aspects of the acoustic–articulatory relationships. Most studies exploit articulatory data recorded using electromagnetic articulography (EMA) (Hiroya and Honda, 2004; Richmond, 2006; Toda et al., 2008; Zhang and Renals, 2008; Zen et al., 2011; Youssef et al., 2011; Hueber et al., 2012). This motion-capture device enables the very accurate tracking of a set of sensors glued on the main speech articulators (tongue, lips, jaw, velum). Several approaches have been proposed in the literature to model the relationship between articulatory positions captured by EMA and the corresponding speech spectrum. They aim at addressing either the direct mapping problem (articulatory

[1] But no systematic information about the velum position, as discussed in Section 4.1.