Silent-Speech Command Word Recognition using Electro-Optical Stomatography

Simon Stone, Peter Birkholz
Institute of Acoustics and Speech Communication, Technische Universität Dresden
simon.stone@tu-dresden.de, peter.birkholz@tu-dresden.de

Abstract
In this paper that accompanies a live Show & Tell demonstration at INTERSPEECH 2016, we present our current speaker-dependent silent-speech recognition system. Silent-speech recognition refers to the recognition of speech without any acoustic data. To that end, our system uses a novel technique called electro-optical stomatography to record the tongue and lip movements of a subject during the articulation of a set of isolated words in real-time. Based on these data, simple articulatory models are learned. The system then classifies unseen articulatory data of learned isolated words spoken by the same subject. This paper presents the system components and showcases the silent-speech recognition process with a set of the 30 most common German words. Since the system is language-independent and easy to train, the demonstration will also show both training and recognition of any other words on demand.

Index Terms: silent-speech recognition, electro-optical stomatography

1. Introduction
Silent-speech recognition refers to the concept of classifying speech without any acoustic data, e.g., based entirely on articulatory information. It is one of the two major components of a silent-speech interface (SSI) [1], which enables speech communication without any audible signal. Possible users include senior citizens, cancer patients whose larynx had to be removed, and jet pilots, fire fighters or similar jobs in high-noise environments.

The main obstacle in silent-speech recognition is the collection of the articulatory data. Numerous technologies have been employed to capture the speech activity (see [1] for a review). Even though the results are generally encouraging, no system has emerged as clearly superior and word error rates are still relatively high compared to conventional speech recognition (e.g., up to 60% in [2]) in spite of sophisticated data processing and classification techniques.

2. Electro-Optical Stomatography
Our proposed system uses a new technology specifically developed for the capture of tongue and lip movements called electro-optical stomatography (EOS) [3, 4, 5]. It uses a combination of two established techniques to measure articulation: electropalatography (EPG) and glossometry, also known as optopalatography (OPG).

EPG is a well-established technique that is commonly used in diagnostics and therapy of speech pathologies. For a good review of EPG literature see [6], while the technical state-of-the-art is thoroughly presented in [7]. The basic principle of EPG is to record the contact pattern of the tongue and hard palate by using an array of electric contact sensors positioned on a pseudopalate (a thin plastic foil form-fitted to a subject’s hard palate) inside the anterior mouth cavity. This “tongueprint” on the hard palate is characteristic of numerous speech sounds but yields no information on sounds without any palatolingual contact, e.g., open vowels.

Glossometry or OPG, as first introduced by [8], also uses a pseudopalate, but in this case it is outfitted with a number of optical distance sensors that emit light unto the tongue surface and measure the reflected light intensity. The distance sensors are usually arranged along the midsagittal line of the hard palate, as the tongue contour in this section is particularly well-suited to study different vowels.

In EOS, we combine and extend the two complimentary techniques of EPG and OPG. Our system uses a pseudopalate with 124 contact sensors, 5 laser-optical distance sensors (consisting of a laser diode and a phototransistors) along the mid-sagittal line and one additional lip sensor, which consists of a laser diode and two detectors (phototransistors) to capture the lip opening and protrusion. It gathers data at a frame rate of 100 Hz. An example EOS palate is shown in Figure 1.

3. Silent-speech Command Word Recognition
The speaker-dependent system that is presented in this paper uses EOS data of a subject to learn and recognize isolated command words. The recording of EOS data, training of the word models, and evaluation of the recognition is done in a PC software written in C++ (using the free wxWidgets library available from www.wxwidgets.org). Figure 2 shows a screenshot of
the interface.

3.1. Data
The EOS data for training and evaluation of the silent-speech recognizer consists of 12-bit A/D converted light intensity values from the distance sensors that are related to a distance in cm using the calibration scheme described in [5]. Each EOS data frame also contains two 12-bit A/D converted values from the two detectors of the lip sensor and 124 binary values that represent the palatolingual contact pattern. The measured distances and the lip sensor values are directly used as features, but in order to reduce the dimensionality of the feature vectors, the EPG coefficients proposed by [9] are used instead of the entire contact pattern. All features are then transformed to the same range by normalization.

3.2. Training
The silent-speech word models are trained using a standard Viterbi algorithm and the Mahalanobis distance measure. Training data can be recorded on a word by word basis with a user-specified number of repetitions, or using a pre-defined set of word prompts from a text file. The features used in the training can be chosen by the user to investigate their discriminative impact.

3.3. Evaluation
The software currently supports evaluation of the word models by recording a single utterance (the use case of a command word recognizer) or by recording a set of samples defined by a randomized word list from a text file for statistical testing. The best matching model is found in a Nearest Neighbor scheme. Example results using a set of 10 each of the most common German nouns, adjectives and verbs in 5 repetitions (for a total of 150 samples) are shown in Figure 3.

4. Conclusions and Outlook
Our system shows that EOS is well-suited to perform silent-speech recognition. Even with the currently implemented very basic training and matching algorithms, competitive performance accuracies can be achieved. The simplicity of the system allows the quick training of a small vocabulary command word recognizer.

The system is a work-in-progress that is continuously extended: More state-of-the-art training and classification al-

---

Figure 2: Screenshot of the user interface. The “SSI” tab holds buttons for training and testing the word models.

Figure 3: Confusion matrix of a test corpus of 5 repetitions of 30 German words vs. the word models obtained with a different training corpus of 5 repetitions of the same words articulated by the same speaker. In this case, only the distance sensor data were used resulting in an overall Performance Accuracy of 52 %.