Biosignal-based Spoken Communication: A Survey

Tanja Schultz, *Senior Member, IEEE,* Michael Wand, *Member, IEEE,* Thomas Hueber, *Member, IEEE,* Dean J. Krusienski, *Senior Member, IEEE,* Christian Herff, *Member, IEEE,* Jonathan S. Brumberg

***Abstract*—Speech is a complex process involving a wide range of biosignals, including but not limited to acoustics. These biosignals - stemming from the articulators, the articulator muscle activities, the neural pathways, and the brain itself - can be used to circumvent limitations of conventional speech processing in particular, and to gain insights into the process of speech production in general. Research on biosignal-based speech processing is a wide and very active field at the intersection of various disciplines, ranging from engineering, computer science, electronics and machine learning to medicine, neuroscience, physiology, and psychology. Consequently, a variety of methods and approaches have been used to investigate the common goal of creating biosignal-based speech processing devices for communication applications in everyday situations and for speech rehabilitation, as well as gaining a deeper understanding of spoken communication. This article gives an overview of the various modalities, research approaches, and objectives for Biosignal-based Spoken Communication.**

***Index Terms*—biosignals, spoken communication, multimodal technologies, speech recognition and synthesis, speech rehabilitation, electromyography, ultrasound, functional near-infrared spectroscopy, electroencephalography, electrocorticography**

I. Introduction

Human speech production is a complex motor process, that starts in the brain and ends with respiratory, laryngeal, and articulatory gestures for creating acoustic signals of verbal communication. Physiological measurements using specialized sensors and methods can be made at each level of speech processing, including the central and peripheral nervous systems, muscular action potentials, speech kinematics (tongue, lips, jaw, etc), and sound pressure. Together, these physiological measurements are known as speech-related "biosignals" and have been used for decades to better understand the underlying mechanisms of human speech production. Modeling the mapping between physiological parameters and acoustic consequences of speech still remains a very active research field. Propelled by technological advances, an increasing number of studies have investigated speech-related biosignals in applied research focused on developing spoken communication (SC) systems. This field is referred to as "Biosignal-based Spoken Communication," and encompasses two primary tracks for converting: (1) biosignals into text (biosignal-based speech recognition), and (2) biosignals into a synthetic voice (biosignal-based speech synthesis). Examples of these two technical tracks include Brain-Computer Interfaces (BCI) for restoring communication by directly decoding cortical brain activity [1], [2], [3] into speech representations, and Silent-Speech Interfaces (SSI) [4], which offer a way to communicate privately without disturbing bystanders and / or provide voice communication for people with severe speech impairments (e.g., laryngectomy patients). Furthermore, several studies have recently investigated biosignals as a means to provide valuable articulatory biofeedback to speakers about their own voice production for increasing articulatory awareness in speech therapy or language learning (e.g., [5], [6], [7]).

T. Schultz and C. Herff are with the Cognitive Systems Lab, Faculty for Computer Science and Mathematics, University of Bremen, Germany, e-mail: (tanja.schultz@uni-bremen.de).

M. Wand is with The Swiss AI Lab IDSIA, Lugano, Switzerland

T. Hueber is with the GIPSA-lab, CNRS/Grenoble-Alpes Univ., Grenoble France

D.J. Krusienski is with the ASPEN Lab, Biomedical Engineering Institute, Old Dominion University, Norfolk, VA, USA

J.S. Brumberg is with the Speech and Applied Neurosciene Lab, Speech-Language-Hearing Department, University of Kansas, Lawrence, KS, USA

The field of Biosignal-based Spoken Communication has rapidly advanced in recent years and the IEEE Special Issue on this subject is intended as a snapshot and comprehensive review of the current state-of-the-art. This survey paper provides an overview and definition of the methods, sensor technologies, signal processing algorithms, and applications used across the field. We provide specific focus on the processing, analysis, classification, recognition, and interpretation of a large variety of biosignals representing speech and language, including a discussion on advanced machine learning approaches, as well as theory and applications related to spoken language processing. With its broad scope, this survey intends to bridge the gap between the disciplines, provide a linking structure within the special issue, and to generally provide an entry point for readers interested in this very active field of research and development.

The remainder of this survey paper is organized in five sections. Following this introduction, Section II provides a definition of "biosignals", as well as the different modes of speaking. Section III describes methods used to acquire speech-related biosignals, ranging from respiratory, laryngeal, and articulatory kinematics, to muscular and neurological activity. Section IV summarizes processing methods needed to analyze each speech-related biosignal and includes descriptions of relevant features, dimensionality reduction and compression methods. This section also discusses the usage of biosignal-based automatic speech recognition and speech synthesis. The paper ends with a discussion of the wide variety of use cases and existing applications in Section V, and a view toward the future of Biosignal-based Communication in Section VI.

II. General Definitions and Uses of Biosignals

In this section we provide a definition for *biosignals* along with a description of the most important biosignals in speech. We also define a variety of *speaking modes* referred to throughout the article.

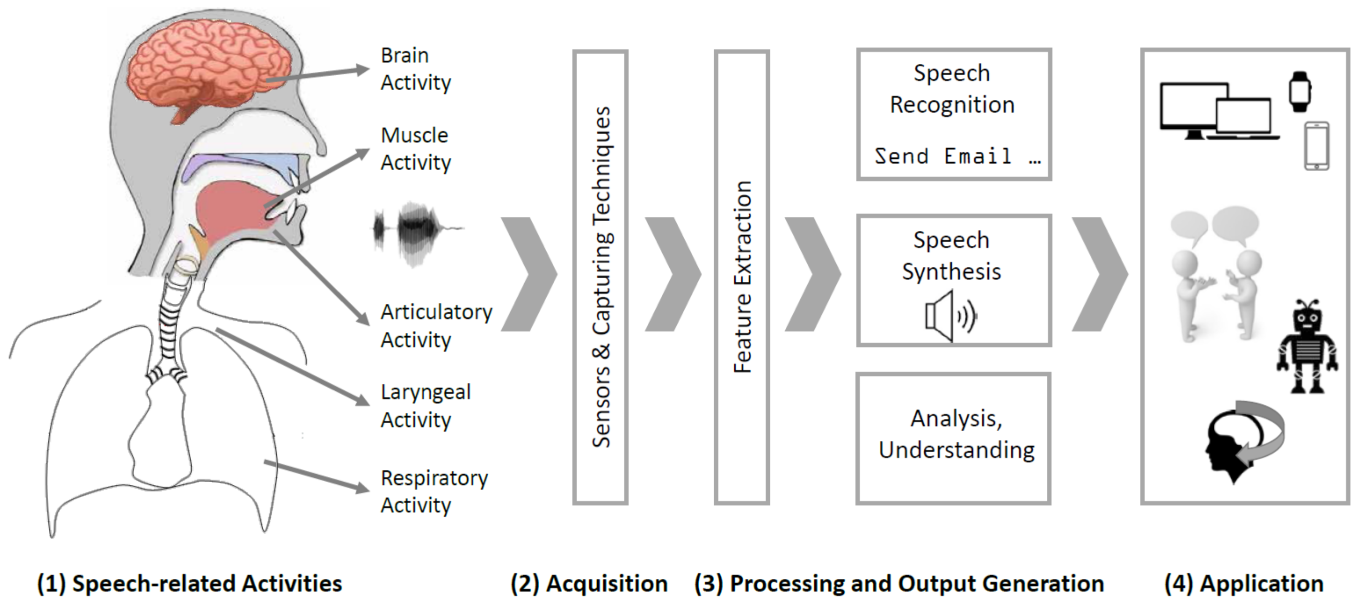


Fig. 1. Biosignal-based Spoken Communication resulting from (1) Speech-related Activities of the Human Body, (2) Signal Acquisition using various activity-dependent Sensor Technologies, (3) Biosignal processing including Feature Extraction followed by Output Generation for (4) various target Applications.

A. Biosignals

We define *Biosignals* as autonomous signals produced by human activities measured in physical quantities using different sensor technologies. Autonomous signals result from chemical, physical, and biological processes of the human organism and serve the functions of control, regulation, and information transmission throughout the body. Sensor technologies can be used to measure each signal, in terms of kinetic (force, torque, movement), kinematic (position, velocity, acceleration), optical (radiance, luminance), chemical (concentration, pH, olfactory), electrical (potential, current, resistance), acoustic (sound pressure and intensity, impedance), and thermal (temperature) quantities, resulting in the corresponding categories of biosignals.

Biosignals have been used in medical diagnostics [8] for decades. More recently, rapid advances in sensor technologies in accuracy, resolution, miniaturization, integration, connectivity, mobility, usability, costs, availability, and many other features, have propelled the application of biosignals to other contexts, including information technologies. In particular, the human-computer interaction (HCI) community has embraced biosignals to extend the number of modalities available for developing robust and intuitive devices. Information obtained from the biosignals is used to interpret not only physical states, but also affective and cognitive states, and activities of a user. Thereby, biosignals provide an *inside* perspective on human mental processes, intentions, and needs that complement traditional means of observing human interaction from the *outside*, and thus enable personalized and adaptive services [9].

In speech and language, biosignals are used for basic and translational research and development, including: voice-driven HCI, human-human interaction and communication, speech therapy, and language learning. At a basic level, biosignals can provide a comprehensive description of speech processing by reflecting all speech-related activities of the human body as depicted Figure 1 (step #1), found in the brain, the peripheral nervous system, the muscles, the speech anatomy of articulation (jaw, lips, tongue, and other orofacial structures), phonation (vocal folds), and respiration. The biosignals of speech then result from being captured through a wide variety of sensors and capturing techniques (step #2 in Figure 1, see Section III).

The remainder of this article focuses on biosignals beyond traditional acoustic waveforms captured by techniques such as electromyography (EMG), electroencephalography (EEG), electrocorticography (ECoG), intracranial microelectrodes, functional near-infrared spectroscopy (fNIRS), ultrasound (US), and permanent magnet articulography (PMA). While acoustic biosignals can only be captured during vocalizations that displace particles of the surrounding medium (usually air) by a vibrating object (usually the vocal folds), kinematic, kinetic, electrical, and optical biosignals do not rely on such air particle displacement and thus extend to many *speaking modes* beyond the audible one.

B. Speaking Modes

Speech results from modulation of the expiratory air flow from the lungs through the glottis, which is filtered by the vocal tract [10]. The acoustic transfer function of the vocal tract depends on the geometry of both oral and nasal cavities, which are configured by positions of the tongue, lips, jaw, and velum. For the purpose of Biosignal-based Spoken Communication, we distinguish different speaking *modes* based on glottal activity and intensity in Tables I & II.

Each speaking mode in Tables I & II produces sound pressure waves that can be captured by traditional acoustic-based sensors resulting in acoustic biosignals. This survey focuses specifically on biosignal acquisition from *speech produced*

TABLE I
SPEAKING MODE DEPENDENT ON GLOTTAL ACTIVITY

Modal speech	The vocal folds vibrate for voiced sounds or do not vibrate for unvoiced sounds.
Whispered speech	Turbulent flow through a constant aperture formed between the vocal folds results in only unvoiced sound.

TABLE II
SPEAKING MODE ACCORDING TO LEVEL OF EFFORT.

Normal modal speech	Modal speech at normal intensity: the "standard" mode of speaking.
Shouted speech	Modal speech characterized by higher intensity, higher pitch, and more open articulation than speech at normal intensity. It shares common properties with Lombard speech produced in noisy environments [11].
Murmured speech	Characterized by very low-intensity (voiced/unvoiced) sounds that are barely perceptible to bystanders. Residual acoustic activity can, however, be recorded using a specific microphone (see Section III-B).

without making any sound. The current literature refers to "speech-without-sound" rather inconsistently, and sometimes equivalently as, imagined, silent, covert, or inner speech, despite differences in their behavioral components. In the context of spoken communication studies, the confusion and inconsistency of terminology might be a result of different instructions given to subjects - or the lack of instructions. In Table III we propose the classification of speech-without-sounds into three levels: silent, imagined, and inner speech.

TABLE III
CLASSIFICATION OF SPEECH MODES WITHOUT ACOUSTIC OUTPUT.

Silent speech	Speakers are instructed to move their articulators as if producing normal modal speech, but to suppress their pulmonary airstream so that no sound is emitted. Silent speech production can be measured by monitoring articulatory movements using motion-capture devices, imaging techniques, or by measuring the activity of muscles (see Section III).
Imagined speech	Similar to silent speech, except movements of the articulators are also suppressed. Imagined speech in this context is identical to first-person motor imagery of speaking in which the speakers should feel as though they are producing speech rather than simply talking to themselves. Since imagined speech is produced without any articulatory movements, this speaking mode requires observations at the neural level.
Inner speech	Though there is a range of descriptions for inner speech (e.g., self-talk, verbal thinking, inner voice, inner dialogue) [12], we adopt Vygotsky's model [13] that defines inner speech as an internalized process in which one thinks in pure meanings. In contrast to imagined and silent speech, no phonological properties and turn-taking qualities of an external dialogue are retained. Thus, inner speech is even more difficult to investigate, even at the neural level of observation.

Each speaking mode in Table III has distinct challenges and opportunities for signal acquisition and application to spoken communication. Some opportunities include: (1) robustness to adverse environments, e.g., measuring articulation is less prone to acoustic noise than airborne signals; (2) less disturbing or more secure, e.g., whispered or silent speech is favored over normal modal speech in quiet environments, and silent or imagined speech allows one to communicate confidentially; and (3) rehabilitation / restoration applications for individuals with voice problems or speech disabilities, e.g., silent speech interfaces as a voice prosthesis for individuals with laryngectomy, and possibly as a neural prosthesis for speech using imagined speech (i.e., speech brain-computer interfaces) for individuals with paralysis and mutism due to neurological disease or trauma (e.g., locked-in syndrome).

In addition to challenges of recording and processing biosignals for speech without sound (section IV-D), a major challenge is precisely due to the lack of auditory feedback and, for imagined and inner speech, a complete lack of behavioral landmarks. Some specific challenges for silent, imagined, and inner speech include: (1) difficulty distinguishing speech from non-speech activity, (2) a lack of temporal information about the speech content, and (3) difficulty for study participants to utter silent speech [14] due to the absence of auditory feedback. Another confound for silent speech is that articulation may change depending on the communication situation: e.g., a silent speaker communicating in a public place may hypo-articulate to prevent lip-reading and maintain privacy. Careful instruction of study participants is therefore necessary to obtain consistent signals.

III. CAPTURING SPEECH-RELATED BIOSIGNALS

This section describes the production of speech as a result of (1) respiratory, laryngeal, and articulatory activity, (2) intraoral residual acoustic activity, (3) muscle activity, and (4) brain activity, and gives an overview of methods and techniques for their acquisition, (see step #2 in Figure 1).

A. Respiratory, Laryngeal, and Articulatory Activity

Breathing is central to speech production by providing the airflow required to generate sounds. Breathing kinematics can be recorded by means of a face/nose mask or by chest and abdominal plethysmography; their properties during speech production have been extensively studied for more than 40 years (e.g., [15]). More recently, Rochet-Capellan et al. [16] revealed that breathing may contribute to timing and coordination between dialogue partners in face-to-face spoken communication.

Laryngeal activity refers to vocal fold oscillations (in modal and murmured speech), and can be estimated either indirectly from speech acoustics using inverse filtering, or directly by an electroglottography (EGG) [17]. With EGG, the degree and rate of vocal fold contact is related to changes in electrical resistance between two electrodes placed around the neck. This technique is very sensitive to the exact positioning of the electrodes relative to the location of the vocal folds.

Articulatory activity refers to the movements of the speech articulators, and can be measured using a number of different techniques. Here, we distinguish techniques based on sensors attached along the vocal tract from imaging techniques.

Magnetic Articulography (EMA/PMA): Two techniques *Electromagnetic Articulography* (EMA) [18] and *Permanent-Magnetic Articulography* (PMA) [19], [20] are available to measure articulator configurations during speech production using magnetic field sensing. The location where magnetic field generation and sensing take place differentiates each approach.

To record EMA, participants are seated with their head inside an alternating magnetic field, generated by transmitter coils. This field induces an electrical current in receiver coils glued to the main articulators (tongue, lips, velum). Multiple transmitter and receiver coils are used to recover real-time articulatory movements in a 2D or 3D Cartesian space. EMA records articulatory data with very high spatial and temporal resolution (<1 mm, ~500 Hz), and is used to model articulatory dynamics during speech production. These data have been explored in different areas of speech technology, such as automatic speech recognition [21], low bit-rate speech coding [22], and speech synthesis [23]. EMA is an invasive procedure and requires wires to be run inside the mouth, which can cause discomfort, and is not portable. As a result, EMA is typically used in laboratory settings.

In PMA, the positions of the sensors are reversed: permanent magnet transmitters are attached to the articulators, and the sum magnetic field is measured by sensors outside the mouth. The resulting field is a superposition of all the transmitter fields, and requires sophisticated analyses to decode the spatial position of articulators. However, since PMA requires only permanent magnets to be fixed inside the mouth without any connecting wires, it is more comfortable than EMA [19].

Palatography: This technique uses sensors embedded inside a pseudo-palate that are placed inside the mouth. In Electropalatography (EPG), contact sensors are used to record the timing and location of palatal contacts during speech. A modification by Birkholz et al. added optical distance sensors to the pseudo-palate (Optopalatography) [24] to record tongue positions for phonemes that do not involve palatal contacts (e.g., vowels), and lip movements.

Imaging techniques (IMG): Video imaging is a straightforward way to capture the movements of the visible speech articulators (i.e., lips and jaw) during speech production. Several sizeable (audio-)visual data corpora are available, such as GRID [25] and the “Lip Reading in the Wild” corpus [26].

Medical imaging techniques can be used to capture the movements of the intraoral articulators. Magnetic Resonance Imaging (MRI) is widely used in phonetic research [27], and obtains high-contrast images of the vocal tract showing all articulators and internal structures. Moreover, recent advances in real-time dynamic MRI (RT-MRI) can be used to record sequences of vocal tract images at 100 fps with acceptable spatial resolution [28]. However, MRI requires a bulky and expensive equipment which prevents its use as a portable communication device.

Ultrasound imaging of the vocal tract is a clinically safe technique that records images of tongue movements during speech in the mid-sagittal or coronal planes with good spatial and temporal resolution (~1 mm, ~80 Hz), see [29] for a complete review. Data is recorded by placing an ultrasonic transducer beneath the chin (held manually or using a head strap) to emit ultrasonic waves and detect reflections from the upper surface of the tongue. Ultrasound images have relatively low quality due in part to the presence of speckle noise and to a loss of signal from tongue structures with poor alignment to the ultrasound beam (i.e., non-orthogonal). However, lightweight ultrasound scanners are now available making this technology suitable for practical communication systems.

B. Intraoral Acoustic Activity

The acoustic output of very soft vocal productions such as murmured speech is too small to be recorded using a conventional microphone, though it can be captured using a stethoscopic (i.e., tissue-conducted), non-audible murmur (NAM) microphone [30]. The device is placed just below the ear, and is capable of detecting very low-amplitude sounds generated inside the vocal tract by a soft laryngeal airflow. The main application for NAM microphones is the design of silent speech interfaces. Intraoral acoustic activity can also be exploited for spoken communication (in normal speech) in very noisy / adverse environments (e.g., a helicopter cockpit). Some examples include *throat microphones* that detect the acoustic variations propagating through the neck tissues [31], and *bone-conducted* microphones that detect intraoral activity via a sensor placed on the skull [32].

C. Muscle Activity

Muscular activity can be observed using electromyography (EMG) to capture electrical signals generated during muscle fiber contraction [33]. EMG can be recorded in two ways: invasively via needle electrodes inserted into muscle tissue or non-invasively using surface electrodes. Surface electrodes are most common in the context of speech processing systems since using needle electrodes requires medical expertise and hygiene precautions, and they are susceptible to dislocation when applied to moving tissue [33].

Surface Electromyography (EMG): Speech-related surface EMG is acquired using electrodes attached to the face positioned either over specific muscles [34], [35] or arranged in a grid [36]. Signals are acquired as a potential difference between two electrodes, measured either in a monopolar (reference-versus-active) or bipolar (active-versus-active) configuration. The recorded voltage potentials are separated from their generators (i.e., motor units) by layers of tissue with varying conductivity; therefore, they represent a superposition of many activity sources, possibly even several muscles. The EMG signal is further attenuated by skin tissue and the skin-electrode interface, which both act as a low-pass filter [37]. However, EMG is advantageous for speech synthesis and recognition because the signal appears approximately 60 ms *before* actual articulatory movements [38], [39].

D. Brain Activity

Brain activity can be measured based on its hemodynamic (fMRI, fNIRS) or electrophysiological (EEG, MEG, ECoG, microelectrodes) dynamics.

Functional Magnetic Resonance Imaging (fMRI): Neural activity can be acquired using fMRI by observing the changing concentrations of oxygenated and deoxygenated hemoglobin, which are related to the increased demand for oxygen as neurons are active and engaged. Oxygenated and deoxygenated hemoglobin have different magnetic properties that can be detected by the strong magnetic fields produced in the MRI environment. Due to its high spatial resolution over the entire brain, fMRI is the de-facto standard in neuroimaging and has been instrumental in a variety of studies investigating speech and language, for reference, see [40] for a review. The slow nature of the hemodynamic response, noisy environment, and the large chamber required for fMRI significantly limits the utility for practical communication interfaces.

Functional Near Infrared Spectroscopy (fNIRS): fNIRS is a brain imaging technique pioneered by Jobsis [41] that also detects changes in the amount of hemoglobin present in the brain as an indirect marker of neural activity. Light in the near infrared spectrum is absorbed by hemoglobin, but not by biological tissue (e.g., bones, skin, muscle). Therefore, the amount of hemoglobin present can be estimated by placing near infrared light emitters and detectors around the head and calculating the amount of light absorbed. Similar to fMRI paradigms, neural activity increases the demand for energy, which is supplied by fresh oxygenated blood that carries hemoglobin to the site of neural processing. fNIRS is well suited to investigate speech processes in non-clinical populations as it is less affected by motion artifacts that plague EEG [42] and can quickly be set up in non-laboratory environments. fNIRS emitters can also be easily realized using LEDs [43], which enable low-cost fNIRS devices [44]. Additionally, the light emitters and detectors do not require additional skin preparation steps common to EEG (e.g., skin abrasion and application of conductive gel), which simplifies acquisition.

Electroencephalography (EEG): EEG is the measurement of the electrical activity of the brain using electrodes placed on the surface of the scalp. EEG signals observed at individual electrode sites are the result of the simultaneous activation of millions of neurons whose summed voltage is conducted through the brain volume, skull, and scalp layers [45]. The large number of neurons contributing to the EEG signal, combined with the low-pass filter properties of the skull and scalp, result in a spatial resolution on the order of centimeters and spectral bandwidth on the order of 80 Hz. As a non-invasive measure of electrophysiological activity, EEG has desirable temporal properties to adequately characterize the neural processing of speech production. Unfortunately, EEG is highly susceptible to myoelectrical, motion, and environmental artifacts, which interfere with EEG recordings made during overt speech production (e.g., modal, whispered, and silent speech) [46]. Though some methods have been developed to cancel this interference (e.g., [47]), validation is still needed to ensure only artifacts are removed from the EEG signal. An alternative is to record EEG during imagined speech (see Table III), or to restrict analysis to the speech motor planning and preparation phases (e.g., [48]). See [49] for a comprehensive review of the EEG components involved in speech and language processing.

EEG is typically recorded and processed using time-locked averages (i.e., event-related potentials, ERPs) to overcome its comparatively low signal-to-noise ratio [50]. However, EEG can also be analyzed as single-trial ERPs and for changes in spectral content over time (e.g., event-related (de)synchronization) [51]. Despite the disadvantages for studying speech, EEG remains the most common technique used in BCIs for communication [1].

Microwire Electrodes & Microarrays: Intracranial wire microelectrodes and microarrays, such as the Utah array [52], provide unparalleled spatial and temporal resolution down to single neuron action potentials. The electrodes are typically 1–2 mm long and have recording surfaces that range from 20 – 80 microns [53]. They record extracellular potentials of only those neurons nearest to the recording tip, and as an array they can record small brain areas of a few square millimeters simultaneously. Extracellular recordings contain both neural spiking data (action potentials, 300 – 6000 Hz) and the local field potential ($<$300 Hz), which represents the neural activity from a larger area around the electrode tip [54].

The invasive procedure to implant microarrays or microwire electrodes into the cortex is only rarely performed with humans and few studies exist investigating speech processes using this technique. In these few examples, implants in cortical areas for speech-motor control have been used to analyze and decode intended phone production [55], [56], and to control a vowel speech synthesizer [57], [58].

Electrocorticography (ECoG): ECoG is an invasive technique for measuring the electrical activity of the brain from sites directly on the cortical surface. The opportunity to measure ECoG in humans is most common in patients with severe cases of epilepsy, who require temporary implantation of electrode grids for pre-surgical planning, or intra-operative monitoring [59]. The implanted sub-dural electrode grids can remain on the brain surface for a period of several days to two weeks, during which patients consent to participate in scientific experiments.

Clinical ECoG recordings typically have an electrode spacing on the order of 10 mm, while micro-ECoG recordings can have spacing on the order of 1 mm. In contrast to scalp EEG, ECoG signals measured on the brain surface do not suffer from spatial blurring from dura matter, skull, and scalp [60], record electrical activity from neural tissue directly underneath each electrode, and are less susceptible to muscle and environmental artifacts. ECoG recordings have a spectral bandwidth in excess of 200 Hz, and special emphasis has been placed on the high-gamma band ($>$70 Hz), which is not readily observable in scalp EEG [61]. The high-gamma range is very spatially localized and highly correlated with cognitive functions and behavioral output, including speech processes [62].

IV. Processing Biosignals for Spoken Communication

While the analysis of the described speech-related biosignals can be used to gain a better understanding of speech processes in general, the development of biosignal-based applications for spoken communication requires further processing (see step #3 of Figure 1). Biosignals are first processed to extract suitable features and to handle artifacts, followed by classification or regression methods to generate the output for the targeted application. The classification approach typically consists of using automatic speech recognition for the transformation of spoken commands or continuous speech into text (e.g., phones, words, phrases or complete sentences), which then can be displayed on a screen or synthesized using text-to-speech synthesis. The regression method typically involves using speech synthesis for the direct mapping of biosignal-captured spoken input to audible speech output. While the boundaries between these three steps are sometimes blurred in practice, for simplicity, we describe them separately. Thus, this section starts with a summary of feature extraction methods for each biosignal, followed by short overviews of speech recognition and synthesis approaches applied to biosignal-based spoken communication with emphasis on the peculiarities of non-acoustic speech-related biosignals. Finally, we summarize the current status of these systems and discuss open challenges.

A. Extracting Relevant Features from Biosignals

Following the acquisition of speech-related biosignals (Section III), relevant features are extracted according to mode-specific standards in physiological signal processing.

Acoustic signals, limited here to body conduction microphones (including NAM), are typically processed similarly to standard speech signals from normal (modal) speech. For example, Mel-Frequency Cepstral Coefficients (MFCC) plus context features can be estimated from NAM recordings [30].

Visual articulatory data (e.g., video images of the lips, ultrasound images of the tongue, etc.) are usually acquired as high-resolution 2D or 3D data. We briefly review three main approaches that have been investigated in the context of audio-visual and visual-only speech recognition, silent speech interfaces, and articulatory biofeedback. See [63] for a more detailed review.

In one approach, automatic segmentation of the articulators in each video image (i.e., the extraction of their contours) has been used to track lip movements using the active shape model (ASM) [64], active appearance model (AAM) [65], and more recently constrained local neural fields (CLNF) [66]. For ultrasound images, the robust and fully automatic tracking of the tongue is still an unsolved issue and has been investigated using ASM [67], AAM [68], and neural networks (shallow [69], deep [70]). A second approach uses dimensionality reduction techniques in which an entire region-of-interest is processed without focusing on a particular object (e.g., lip or tongue contours). Some examples of this approach include the discrete cosine transform to process lip images [71] and principal components analysis for lip [72], and tongue images [73]. A third approach has recently emerged using the deep learning paradigm in which both discriminative feature extraction and classification can be jointly achieved. One powerful deep architecture is the so-called *Convolutional Neural Network* (CNN) [74], which has been used in a few recent studies for encoding lips [75], and for extracting high-level articulatory abstractions from the joint observation of lips and tongue images [76].

Magnetic-articulography techniques (i.e., EMA or PMA) commonly provide a low-dimensional data vector representing the positions of the speech articulators, and requires only minimal pre-processing. EMA systems directly measure the 2D or 3D coordinates of the receiver coils attached to the articulators, and are usable in raw form by a classifier or a regression model. PMA data are less explicit and may require more preprocessing such as low-pass filtering, background cancellation, and normalization for proper identification of articulatory positions and movements [19].

Palatography (i.e., EPG and OPG) EPG provides an exact 2D plus time representation of tongue-palate contact patterns and does not require any post-processing. Additional procedures are required for OPG in order to calibrate the distance sensors (the user must touch each sensor once with the tongue while the pseudopalate is in the mouth) and to compensate for measurement errors made when the tongue is not oriented perpendicular to the axes of the optical sensors [77]. Once completed, no further data post-processing is required.

Electromyography provides a timecourse of muscular activation for each recording electrode. Initial approaches used simple thresholding techniques [78] and comparisons of whole-word EMG averages between channels [79] to identify muscles active during speech. Modern approaches now use *time-domain* features [39] similar to the *Hudgins* feature set [80] and *spectral* features [81], [34], [82], [35], [83].

Hemodynamic responses measured by fMRI and fNIRS depend on metabolic processes and are relatively slow as a result. Simple features such as the linear data trend well describe neural activity in fNIRS [84], and newer approaches subsample the hemodynamic response and employ classification methods to determine information bearing spatio-temporal filters [85].

Electrical brain signals measured by EEG and ECoG use similar techniques for analyzing their respective signals to describe the neural processes underlying speech production, and focus on spatial, temporal, and spectral properties. With EEG, it is common to apply a bandpass filter from 1–30 Hz since signals $>$30 Hz are often unreliable due to low SNR. After filtering, ERPs can be aligned to the onset of speech production and averaged to focus on either the time periods preceding or following production onset. The times preceding speech have been well studied and have revealed two major slow-wave potentials that systematically vary with speech production: (1) the *bereitschaftspotential (BP)*, a negativity that occurs in the 1–2 s prior to self-paced speech production [48], and (2) the *contingent negative variation*, a negativity that occurs prior to cued speech production [86]. Analysis of the intervals during speech production is difficult due to EMG contamination (cf. [46]); Alternatively, EEG can be used to intepret neural processes involved in cued imagined speech using both the broadband (1–30 Hz) ERP [87] and narrowband (alpha, beta, and theta) power modulations [88]. The amplitude envelope of the high-gamma band (>70 Hz) in ECoG closely tracks aspects of the acoustic speech signal and can provide an even more detailed view of the spatio-temporal progression of brain activity during speech processes [3], [89]. Similar analyses can be applied to microarray recordings using features such as rate codes and tuning curves [55], [56].

B. Biosignal-based Speech Recognition

Automatic speech recognition (ASR) systems convert speech (typically audio) into text, i.e., a sequence of written words. The ASR task is characterized by its multi-level sequential nature: small units, usually (context-dependent) phones, are concatenated into words, which in turn are concatenated into continuous sentences. In addition, prior probabilities are assigned to word sequences by means of *language models*. For more than 30 years ASR has been dominated by multi-level statistical modeling schemes, in particular hidden Markov models (HMMs) [90] and n-gram language models [91]. Recent applications of artificial neural networks have revolutionized ASR with the development of hybrid Deep Neural Network Hidden Markov Model systems [92], and end-to-end systems that directly map speech features into text [93].

Fundamentally, biosignal-based speech recognition can be approached by replacing the acoustic signal processing front-end with methods tailored to each biosignal while leaving the statistical modeling back-end unchanged. Examples of this approach include isolated word recognition using image-specific features for lipreading [94], and continuous phone-based HMM recognition using sEMG signals [39]. However, there are interdependencies at each processing level in the speech recognition pipeline that allow for adaptation / improvement to back-end systems for each biosignal.

One important design aspect of biosignal-based speech recognition is the way in which smaller units are concatenated into words and sentences. Large units may be easier to recognize, but harder to share between different words, leading to difficulty recognizing unseen vocabulary and additional training data requirements. Short units may be unstable due to coarticulatory effects, or they might not contain enough information to reliably identify a pattern of articulation. In visual speech recognition, *viseme* units have been defined by visually grouping phones of similar appearance, or by considering articulatory gestures [95]. However, speech recognition with visemes causes ambiguities that must be resolved, e.g., by language models. *Bundled Phonetic Features* [96] are a data-driven approach that has been successful for EMG-based speech recognition. Finally, biosignal-based speech recognition has been explored using syllables [97] and context-independent or context-dependent phones [98], [99], [100].

In multi-modal speech recognition, combining sources of information is of particular interest, both for recognition and for possible audio-based bootstrapping. The reliability of each biosignal modality is highly variable, depending on phonetic properties [101] and on the environmental conditions (e.g., noise). Frameworks for dynamic estimation of stream weights have been developed for audio-visual and audio-plus-myoelectric speech recognition [102], [103]. Furthermore, manifestations of the articulation (e.g., brain signal, EMG onset, visible muscle contraction, and sound) are not synchronous [72], [39] due to the multi-step nature of speech motor control and the complex relation between articulatory gestures and speech sounds [104], [105]. Articulatory information (place, manner voicing) can also be used to augment conventional (i.e., audio-based) ASR. Incorporating explicit speech production knowledge in ASR can improve the recognition of spontaneous speech and increase robustness to noise, by modeling more efficiently some co-articulation effects, see [106] for a complete review on this line of research.

Visual articulatory data: Audiovisual speech recognition (AVSR) combines video of a speaker's lips / face with traditional speech audio signals to improve ASR performance in adverse conditions (i.e., background noise) [107]. AVSR continues to be widely investigated (see [108] for a complete review) and has been extended to *purely visual* speech recognition (VSR). In that case, no audio signal is used and speech recognition is performed only from visual information. Lip movements observable by video provide only partial information on speech articulation; therefore, recent efforts have also explored the combination of video and ultrasound imaging to capture both lip and tongue movements [109].

Similar to audio-based speech recognition, classification in AVSR or VSR systems is often accomplished using models that explicitly account for speech dynamics including: HMM [110], [94], [99], deep neural networks [111] and long short-term memory neural networks [112]. Though the addition of visual to audio modalities can augment speech representations, they may be acquired with different temporal structures that must be reconciled, e.g., coupled-HMM [113] and dynamic Bayesian network [114].

Magnetic articulography: Automatic continuous speech recognition from EMA data have been investigated for English [115] and French [116] (in conjunction with the audio signal), and small vocabulary recognition using PMA [117] using standard ASR techniques.

Electromyographic signals: Early EMG-based speech recognition used just three surface EMG electrodes to discriminate Japanese vowels [78] and has since been combined with auditory signals for better performance in noisy environments [118]. More recently, EMG-based recognition has been applied to silent speech applications [81], including whole phrases spoken in silent and normal modal speech [34], [39], syllables [97], and phones [100]. Further developments include modeling context-dependent *Bundled Phonetic Features* to address data sparsity [96], adaptation to cope with recording session discrepancies [119], and development of a hybrid neural network – HMM system for EMG-based ASR [120].

Hemodynamic responses: Both fMRI and fNIRS have mostly been used to study speech neuroscience examining the averaged hemodynamic responses over many repetitions of speech tasks. However, a successful silent speech interface must be able to detect speech events in a single trial. A few studies have investigated this decoding approach using fMRI for decoding three Dutch vowels [121], nouns [122], and functional representations of natural speech [123]. Initial applications of fNIRS to speech recognition have focused on discriminating between the speech modes: modal, silent, and imagined [124], [125]. Using fNIRS, these modes can be discriminated from each other and from intervals without speech activity in single trial. However, the slow nature of the hemodynamic response prohibits investigation on a more fine-grained time scale than whole sentences, and thus does not scale up to spoken communication in any speaking style.

Electrical brain signals: The earliest attempts for speech recognition in EEG were used to predict the word a participant was attending to in a passive listening paradigm, without any speaking involvement (silent, imagined, or other) [126]. This approach has been improved using ECoG to reconstruct a stimulus from auditory cortical activity during passive listening [127]. Additional attempts have focused on speech production (actual or imagined) paradigms for decoding acoustic features and phonemes (EEG: [87], [128], [129]; ECoG: [130], [131]; microelectrodes: [55], [56]), syllables (EEG: [88]; ECoG: [132]), and words (ECoG: [133]). ECoG has also been used to decode speech articulatory features [134], and recently HMM-based ASR was applied to ECoG signals to decode continuously spoken speech [98].

C. Biosignal-based Speech Synthesis

In contrast to speech recognition approaches, speech synthesis is a means to artificially produce human speech from a given input signal. Current biosignal-based approaches usually consist of three processing steps: (1) features extraction from biosignals and (intended) speech (e.g., Mel-cepstral features), (2) biosignal features are mapped to speech features, and (3) speech is synthesized from the predicted speech features, for example by a *vocoder* (a digital filter that models the spectral envelope and is excited with a proper signal). In this section we focus on the second step.

The mapping between biosignal and speech features is usually described as a regression problem between multidimensional continuous variables. Gaussian Mixture Regression is a classical approach inspired by statistical voice conversion [135], [136] and has been used for EMG [137], PMA [20] and US [138] applications. Specifically, the joint probability density function of biosignal inputs and speech outputs is modeled by a Gaussian Mixture Model (GMM). The mapping from biosignal to speech is accomplished either frame-by-frame using the conventional mean square error estimator [135], or sequence-by-sequence using a maximum-likelihood estimator [136]. Artificial Neural Networks are also powerful regressors that can be used for direct biosignal-to-speech mapping, and have been used for EMA-to-speech [139], [140], EMG-to-speech [141], and ultrasound/video-to-speech [73]. An HMM-based regression technique based on full-covariance GMM has been proposed to explicitly model phoneme-specific dynamics of articulation, and to use linguistic priors for regularizing biosignal-to-speech conversion [138]. A performance comparison of different mapping approaches in terms of real-time capabilities and conversion quality has been carried out for EMG-to-speech in [142].

Brain-based biosignals have primarily been used for speech recognition applications (Section IV-B), and do not directly incorporate speech synthesis into their decoding models. Only very few attempts have been made for direct speech synthesis using electrophysiological signals from the human brain. In one example, a microelectrode device implanted into the speech motor cortex was used to control a formant frequency speech synthesizer [57], [58]. This BCI-speech synthesizer converted changes in neural activity into the first two formant frequencies using an adaptive filter neural decoder, followed by synthesis for immediate audio output. A recent study has extended the BCI-based formant synthesizer technique for use with EEG instead of implanted microelectrodes [143]. ECoG has also been used for direct synthesis BCIs by applying regression methods to reconstruct the speech spectrogram which can then be converted to an audio waveform [144].

Finally, we note that biosignal-based speech synthesis can be achieved by performing speech recognition followed by applying conventional text-to-speech systems. This method produces high-quality output, but is constrained by the limitations of the underlying recognizer. In particular, speech recognition operates on limited vocabulary, produces recognition errors, and there is an unavoidable delay between articulation and synthesis since words must be completely articulated and recognized before synthesis is possible. This delay is often undesired, particularly in biofeedback applications (see section V-B), which encourages further work on direct speech synthesis from biosignals.

D. Current Research and Open Challenges

All systems described above have made substantial progress in recent years, particularly in algorithm improvements, system miniaturization, and field studies. This section summarizes the status of different biosignal processing systems, their current applicability, and open challenges identified from both the literature and our own work.

The reviewed technologies can be grouped into two categories based on their practical applicability and maturity: (i) a stable baseline system is being tested in field studies with a *substantial* number of potential users (including patients where applicable); (ii) studies are performed on a small number of subjects under laboratory conditions. Technology in category (i) must necessarily provide an easy-to-use recording system and a reasonable speech recognition / synthesis quality.

Visual articulatory data are typically in category (i); large visual speech corpora exist and have been used in various AVSR and VSR studies [111], [112], [26], benefiting from the fact that data can be recorded without special equipment, or is already available. One study achieved 65.4% word accuracy for the recognition of 333 word classes *without* use of a language model applying television broadcasts data [26]. This result is considered state-of-the-art given the large variation in the data and the size of the vocabulary. Yet, improvements can be made considering that short words were excluded due to the presence of visual ambiguities.

EMG-based speech recognition is also considered in category (i); it has been used with large amounts of data from many subjects [145] and has even been extended to speech restoration applications for individuals with laryngectomies [83]. Also, PMA-based speech recognition [117] and EMA-based speech synthesis [140] have been successfully applied to speech restoration, and EPG-based biofeedback to speech therapy [146]. Recognition accuracy has sufficiently improved for feasibility in basic communication scenarios using EMG, e.g., Word Error Rate of 10.3% on a 2000-word vocabulary [83]. In addition, EMG recording systems are mobile and non-intrusive, though it requires time and experience to properly attach the recording electrodes, and best results are obtained when training and test data are recorded in *one* session, without intermediate removal of the electrodes. Unsupervised adaptation schemes show potential to compensate for these session dependencies [119].

Speech recognition from electrical brain signals has so far been limited to laboratory environments (category (ii)), due to methodological complexity and in some cases surgical intervention is required to be performed in specialized hospital environments (e.g., ECoG). Progress using hemodynamic approaches is limited by portability constraints and limited temporal resolution of metabolic processes (e.g., fMIR, fNIRS).

Research foci in biosignal-based speech processing naturally follow from the observed limitations of the existing systems, and include the following:

- *Robust and Portable Recording systems*: Portability has been achieved for PMA [20], EMG [34], and to some extent video/ultrasound [138]. In the case of lipreading, a system could also be based on fixed cameras (e.g., for forensic purposes) instead of a personal device, but this only works if there is an unobstructed view of the subject's face. EEG data can in theory be obtained with a portable device, however in practice high-quality signals are only obtained under laboratory conditions. ECoG, as described above, requires a specialized hospital environment.
- *Feedback*: In Section II-B we note that silent, imagined, and inner speech may be difficult to utter reliably and consistently. Real-time feedback [57], [140], is considered a promising approach to resolve this issue, though it is both a technical challenge (since data must be processed very quickly) and a modeling challenge (due to the asynchronicity between different manifestations of the speech process, see section IV-B).
- *System Adaptation for Silent or Inner/Imagined Speech*: Instead of relying on speakers to properly use real-time feedback, an alternative approach proposes algorithmic adaptation to account for variability in speaking modes, as has been done for an EMG-based speech recognizer [147], [148]. While discrimination of speaking modes from brain activity has been shown to be possible [124], understanding the difference between modal and imagined speech processes and creating large-scale recognizers for imagined speech are still significant open issues.
- *Multi-Session and Multi-Speaker systems*: Most existing systems are speaker-specific, with the exception of some lipreading systems [26]. Even when data is only taken from one speaker, there may be inter-session differences due to a variety of factors (e.g., environmental artifacts, sensor positioning, etc.), which can be remedied by standard methods (adaptation [119], recalibration [140], integration of session independence as a neural network training objective [149]).
- *Sufficiency of speech representations*: Visual speech recognition using only lip images (i.e., lipreading) is insufficient, and suffers from ambiguities, which can be resolved by including ultrasound images of the vocal tract as an additional input. EMA/PMA and EMG are more sufficient, though EMA/PMA do not represent facial gestures, and without needle electrodes, EMG can not represent specific tongue muscles. That said, these methods provide a fairly complete representation of the speech process with ambiguities in voicing only (cf. [101]). Acquiring appropriate and sufficient signals directly from the speech- and language-related areas of the brain should also provide a complete representation of the processes needed to understand and generate speech, though there is a practical limit on signal acquisition and interpretation. For brain-based techniques, sufficiently sampling the speech and language-related areas of the brain remains an open and intriguing challenge.

A comparison between biosignal-based speech processing systems is difficult at this time since available data corpora differ in size, vocabulary, recording setup, etc., and benchmark data have not been established yet. For speech recognition with medium-sized vocabularies, the three major articulation-based systems (PMA, video+ultrasound, EMG) all perform reasonably and similarly, and further improvements are likely in the near future. The availability of large data corpora will be crucial in extending these systems to truly large-scale speech recognition (with tens of thousands of vocabulary words), and equally to high-quality speech synthesis. Ultimately, the "best" system will be the one which most convincingly resolves the issues summarized above, and will depend upon the constraints of the intended application, including factors such as user preference, performance, reliability, environment, comfort, aesthetics, etc., see section V.

V. Use Cases of Biosignal-based Spoken Communication

Capturing, processing, and interpretation of biosignals related to speech in the absence of an intelligible airborne acoustic signal opens up novel use cases in spoken communication (see step # 4 in Figure 1). A survey on Silent Speech Interfaces (SSI) [4] introduced relevant human-computer interfaces developed before 2010. Sensor technologies and machine learning advanced this field in the past few years. Published use cases and applications of "Biosignal-based Spoken Communication" fall into four main categories, (1) voice prostheses and devices to *restore spoken communication*, for individuals unable to speak due to impairment, disease, or accident; (2) methods to deliver articulatory biofeedback of voice production to increase articulatory awareness for *therapy and training for spoken communication*, such as speech therapy and language learning; (3) approaches to enhance speech recognition and synthesis performance for *robust spoken communication in noisy environments*, like the fusion of complementary speech-related biosignals to compensate for signal corruptions under adverse noise conditions; and (4) strategies for *mute spoken communication* in situations, when audible communication is prohibited or unwanted, e.g., avoiding disruptions in quiet environments or securing against eavesdropping. The concrete systems which we describe in this section frequently address several of these challenges, but often target just a single application. This strategic approach affects the direction of research, requires diverse ethical considerations (e.g., for working with patients), and also influences the design of the communication system: for example, individuals with speech impairments may be willing to invest a significant amount of time into the optimization of their personal communication system, whereas healthy users typically expect little or no enrollment time.

A. Restoring Spoken Communication

An important goal for biosignal-based speech synthesis techniques is to restore spoken communication for individuals with disordered or absent vocalization. Each of the modalities described has specific applications and is most appropriate for specific clinical populations (e.g., individuals with dysarthria, laryngectomy, or paralysis). In laryngectomy, an individual's larynx is surgically removed, and traditional options to restore voice include: using an electrolarynx device resulting in a very robotic voice, using oesophageal speech, or using a tracheosophageal prosthesis, which has to be replaced every few months. Biosignal-based alternatives for this population include PMA-synthesis [19], [117] and EMG-based speech recognition [83].

For individuals with the most severe speech and motor impairments, the objective is to supplement or bypass the speech-motor pathways using available biosignals for improved speech output. In this use case, current research focuses on synthesizing speech during imagined speech, or speech attempts by individuals with total paralysis, directly from brain signal recordings. The superior spatial resolution and signal fidelity of invasive techniques such as microelectrodes and ECoG make them promising approaches for the design of practical speech-based BCIs and neuroprosthetics [57], [58], [98]. Such systems may perform a continuous reconstruction of speech or a discrete classification and output of sounds, words, etc. (see Section IV-B), depending on the objective and constraints of the system. While it may be possible to decode individual words or phrases discretely, scaling this approach to a larger vocabulary can become intractable. Alternatively, the ability to decode basic units of speech, such as formants or phones [98], will enable the creation of generative models that are not limited to a fixed vocabulary. In any case, effectively developing and transferring models trained on normal modal speech to imagined speech remains an active research challenge since the neural representations of normal modal and imagined speech are not identical.

B. Therapy and Training for Spoken Communication

The methods developed for biosignal-based speech processing can also be used for multimodal biofeedback in order to study speech production, facilitate second language learning, and rehabilitate speech impairments. Visual feedback of the articulators (e.g., lip reading) can have a dramatic effect on perception [150], and can even improve speech perception and comprehension by individuals with hearing impairments [5]. Articulatory kinematics captured using EMA have been used for speech training with an emphasis on improving second language learning [6], and investigating articulatory deficits in dysarthria [7]. EPG has also been successfully used as a biofeedback tool for speech therapy [146] and L2 pronunciation training [151]. Promising results using ultrasound imaging have also been obtained for rehabilitation of the English /r/ [152] and persisting speech sound disorders [153].

Notably, biosignal-based speech recognition and synthesis performance declines for silent compared to normal speaking, even when the ASR system is trained and tested exclusively on the respective speaking modes [148]. Speakers report difficulties to steadily producing silent speech [14], in part due to the absence of auditory feedback that is critical for normal speech production [154], [155]. Biofeedback created by real-time speech output could provide an optimal solution to alleviate the challenges in BCI and SSI (see section V-D below). In a closed-loop paradigm, speakers can rely on synthetic speech for auditory feedback and exploit it to regulate their own production, as in [140] for SSI and [57] for BCI.

C. Robust Spoken Communication in Noisy Environments

Improving spoken communication under adverse noise conditions has long been a challenge for speech research and development. Large-scale DARPA programs (e.g., ASE, SPINE, RATS) targeted improvements to speech processing in military and civilian contexts, such as in combat situations, air traffic control, search-and-rescue operations, and security scenarios. Beside the development of noise-robust algorithms, this led to the creation of new sensors like throat and bone-conduction microphones, which can be combined with traditional microphones for improved, fused biosignal ASR [156]. EMG is a natural extension of these techniques and has been used for small vocabulary recognition in acoustically harsh environments [157], and spoken communication for firefighters, pilots, and astronauts through electrodes integrated into self-contained breathing apparatuses [81], [158]. Like subaudible microphones, the EMG combined with conventional acoustic signals can further improve ASR performance in noisy environments [118].

D. Mute-Spoken Communication

In many situations, spoken communication is desired but making any sounds is prohibited or socially inappropriate. For example, carrying out phone conversations may disturb bystanders in quiet environments like libraries or is inappropriate during group meetings. Eavesdropping is a risk when communicating private information in public places. Furthermore, safety and security settings may require a silent communication. Several different biosignal-based systems address the challenges of mute-spoken communication. For instance, silent speech interfaces have been developed using ultrasound imaging, combined with a conventional video camera to capture tongue and lip movements simultaneously, without any audio signals [109], [138]. Importantly, several studies show that performance drops when speaking modes are mixed in training and testing [138]. In the case of surface electromyography, signal-based adaptation methods are proposed to reduce the differences between speaking modes [147] and EMG-based speech recognizers are designed which are trained *and* tested on silent speech [148], [35]. Another way to alleviate the impact of articulatory differences between modal and silent speech is to provide a silent speaker with a synthetic auditory feedback, in real-time [140], [57]. One example involves an articulatory synthesizer that converts EMA data into spectral features using a deep neural network, and can be controlled in real-time by naive subjects articulating silently [140].

PMA-based speech recognition and synthesis has now been achieved in a highly portable manner [20]. While most published research focuses on the aim of restoring speech communication to speech-impaired persons (see section V-A above), PMA was originally proposed for mute communication of individuals without impairment [117]. However, a full study on using silent speech to drive a PMA-based synthesizer has not yet been published.

Studies toward EEG-based speech recognition on silent or imagined speech include classification of single phonemes [87], [88], [129]. Alternative approaches use limb motor imagery to control a formant frequency speech synthesizer without the presence of an acoustic speech signal [143]. Imagined speech decoding has been accomplished with a greater range of speech output using intracortical recording methods including formant frequency prediction using microelectrodes [57], [58], phoneme classification with microelectrodes [55], spectrotemporal features using ECoG [159], and word pairs using ECoG [160].

Beyond mute communication, the technologies described in this survey may be combined with speech translation to bridge the language barrier [39]. Using current procedures, simultaneous translation of a spoken conversation results in the overlap of two voices (one voice from the speaker in the source language and one voice in the target language, coming either from a human interpreter or from the synthesized output of a speech translation system). To avoid such inconvenient scenarios, speakers could instead silently speak (or imagine) in their native tongue, while listeners hear only the translated output. Thus, the combination of mute communication plus translation creates the illusion of speaking in a foreign tongue.

VI. Conclusions and Perspectives

Biosignal-based Spoken Communication is a rapidly evolving cross-disciplinary field. Research and development takes place at the intersection of engineering, computer science, medicine, psychology, and neurosciences. It requires the mastering of sensor technologies, signal, speech and language processing, as well as human-machine interfaces.

This survey paper is intended to provide an entry point for readers interested in this very active field, to define and describe terminology, to recite relevant publications, and thereby to bridge the gap between disciplines. It presents a broad overview over the state-of-the-art technologies, methods, and applications. Table IV summarizes the applicability of biosignals for use cases and speaking modes described in this survey (table cells are grayed out for those speaking modes prohibited by a certain use case). Cell entries in normal font identify techniques that have been reported for capturing speech-related activities of the respective speaking mode and have successfully applied the resulting biosignals to the use cases; these studies are cited in this survey. Italic font indicates applicability but no published results yet, while "-" mark cases when a capturing technique is not applicable.

TABLE IV
APPLICABLE TECHNOLOGIES FOR USE CASES AND SPEAKING MODES (GRAYED OUT CELLS = NO TARGET SPEAKING MODE FOR USE CASE, *italic font* = APPLICABLE BUT NO PUBLICATIONS YET, "-" = NOT APPLICABLE)

Use Cases	Speaking Modes (Section II, Table 1-3)				
(Section V)	modal	murmer	whisper	silent	imagine
(A) Restore SC			*EMG* *PMA* *IMG* *ECoG*	EMG PMA *IMG* ECoG	- - - ECoG
(B) Therapy & Training	EMA EPG IMG *intraoral*	*EMA* *EPG* *IMG* *intraoral*	*EMA* *EPG* *IMG* -	EMA *EPG* *IMG* -	- - - -
(C) Robust SC	EMG *EPG* *PMA* *IMG* intraoral	EMG *EPG* *PMA* *IMG* intraoral	EMG *EPG* *PMA* *IMG* -		
(D) Mute SC		NAM		EMG EMA *PMA* IMG EEG ECoG	- - - - *EEG* ECoG
Insights in SC	All biosignals captured by described technologies including fMRI, fNIRS, MEG, and their combination				

Driven by recent advances in sensor technologies (resolution, accuracy, miniaturization, energy consumption, connectivity, mobility, and costs, to name only a few), the large attention and developments in neurosciences, and the impact of deep learning approaches to automatic speech processing, we expect major breakthroughs in the years to come.

Acknowledgments

The authors gratefully acknowledge the contributions of many of their past and present collaborators, including X. Alameda-Pineda, P. Badin, G. Bailly, AW Black, F. Bocquelet, G. Chollet, B. Denby, F. Elisei, D. Fabre, L. Girin, C. Guan, M. Janke, S-C. Jou, K. Nakamura, P. Perrier, K. Prahallad, M. Pouget, P. Roussel, G. Schalk, J.L. Schwartz, E. Tatulli, A. Toth, and B. Yvert. Their insights and hard work have made the subject of Biosignal-based Spoken Communication a very fruitful area in recent years. Special thanks to L. Diener, who helped with the write-up of this material and to the anonymous reviewers for their valuable comments.

TS, DJK, and CH acknowledge joint funding by the Federal Ministry of Education and Research (BMBF) in Germany and the National Science Foundation (NSF) in the USA for the project "RESPONSE - REvealing SPONtaneous Speech processes in Electrocorticography" under references 01GQ1602 (BMBF) and 1608140 (NSF). MW acknowledges funding from the EU H2020 programme (#687795). JB acknowledges funding by the National Institutes of Health (R03-DC011304).

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clin Neurophysiol*, vol. 113, no. 6, pp. 767–791, 2002.

[2] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: A focused review," *Front Neurosci*, vol. 10, no. 429, 2016.

[3] S. Chakrabarti, H. M. Sandberg, J. S. Brumberg, and D. J. Krusienski, "Progress in speech decoding from the electrocorticogram," *Biomed Eng Lett*, vol. 5, no. 1, pp. 10–21, 2015.

[4] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Commun*, vol. 52, no. 4, pp. 270 – 287, 2010.

[5] N. P. Erber, "Auditory-visual perception of speech," *J Speech Hear Disord*, vol. 40, no. 4, pp. 481–492, 1975.

[6] W. F. Katz and S. Mehta, "Visual feedback of tongue movement for novel speech sound learning," *Front Hum Neurosci*, vol. 9, 2015.

[7] J. J. Berry, C. North, B. Meyers, and M. T. Johnson, "Speech sensorimotor learning through a virtual vocal tract," *J Acoust Soc Am*, vol. 133, no. 5, pp. 3342–3342, 2013.

[8] E. Kaniusas, *Biomedical Signals and Sensors I*. Springer, 2012.

[9] T. Schultz, C. Amma, D. Heger, F. Putze, and M. Wand, "Biosignalebasierte Mensch-Maschine-Schnittstellen," *at - Automatisierungstechnik, 2013*, vol. 61, no. 11, pp. 760 – 769, 2013.

[10] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete Time Processing of Speech Signals*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[11] M. Garnier, N. Henrich, and D. Dubois, "Influence of sound immersion and communicative interaction on the Lombard effect," *J Speech Lang Hear Res*, vol. 53, no. 3, pp. 588–608, 2010.

[12] B. Alderson-Day and C. Fernyhough, "Inner speech: Development, cognitive functions, phenomenology, and neurobiology," *Psychol Bull*, vol. 5, no. 141, pp. 931–965, 2015.

[13] L. Vygotsky, R. Rieber, and A. Carton, *The Collected Works of L.S. Vygotsky: Volume 1: Problems of General Psychology, Including the Volume Thinking and Speech*, ser. Cognition and Language: A Series in Psycholinguistics. Plenum, 1987.

[14] C. Herff, M. Janke, M. Wand, and T. Schultz, "Impact of different feedback mechanisms in EMG-based speech recognition," in *12th Annu. Conf. Int. Speech Communication Association*, Florence, Italy, 2011, pp. 2213 – 2216.

[15] T. J. Hixon, M. D. Goldman, and J. Mead, "Kinematics of the chest wall during speech production: Volume displacements of the rib cage, abdomen, and lung," *J Speech Lang Hear Res*, vol. 16, no. 1, pp. 78–115, 1973.

[16] A. Rochet-Capellan and S. Fuchs, "Take a breath and take the turn: how breathing meets turns in spontaneous dialogue," *Philos T R Soc B*, vol. 369, no. 1658, p. 20130399, 2014.

[17] M. Rothenberg, "A multichannel electroglottograph," *J Voice*, vol. 6, pp. 36–43, 1992.

[18] P. W. Schönle, K. Gräbe, P. Wenig, J. Höhne, J. Schrader, and B. Conrad, "Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract." *Brain Lang*, vol. 31, pp. 26 – 35, 1987.

[19] J. M. Gilbert, S. I. Rybchenko, R. Hofe, S. R. Ell, M. J. Fagan, R. K. Moore, and P. Green, "Isolated word recognition of silent speech using magnetic implants and sensors," *Med Eng Phys*, vol. 32, pp. 1189 – 1197, 2010.

[20] J. A. Gonzalez, L. A. Cheah, J. M. Gilbert, J. Bai, S. R. Ell, P. D. Green, and R. K. Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Comput Speech Lang*, vol. 39, pp. 67 – 87, 2016.

[21] K. Livescu, "Feature-based pronunciation modeling for automatic speech recognition," PhD, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2005.

[22] S. Chennoukh, D. Sinder, G. Richard, and J. Flanagan, "Articulatory based low bit-rate speech coding," *J Acoust Soc Am*, vol. 102, no. 5, pp. 3163–3163, 1997.

[23] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1171–1185, 2009.

[24] P. Birkholz and C. Neuschaefer-Rube, "Combined optical distance sensing and electropalatography to measure articulation." in *12th Annu. Conf. Int. Speech Communication Association*, 2011, pp. 285–288.

[25] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J Acoust Soc Am*, vol. 120, no. 5, pp. 2421 – 2424, 2006.

[26] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *13th Asian Conf. Computer Vision*, 2016.

[27] A. D. Scott, M. Wylezinska, M. J. Birch, and M. E. Miquel, "Speech MRI: morphology and function," *Physica Medica*, vol. 30, no. 6, pp. 604–618, 2014.

[28] P. W. Iltis, J. Frahm, D. Voit, A. A. Joseph, E. Schoonderwaldt, and E. Altenmüller, "High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players," *Quantitative imaging in medicine and surgery*, vol. 5, no. 3, p. 374, 2015.

[29] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clin Linguist Phonet*, vol. 19, no. 6-7, pp. 455 – 501, 2005.

[30] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *2003 IEEE Int. Conf. Acoustics Speech Signal Processing*, Hong Kong, 2003, pp. 127 – 130.

[31] S. A. Patil and J. H. Hansen, "The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification," *Speech Commun*, vol. 52, no. 4, pp. 327–340, 2010.

[32] J. C. Bos, D. W. Tack, and L. L. Bossi, "Speech input hardware investigation for future dismounted soldier computer systems," *DRCD Toronto CR*, vol. 64, 2005.

[33] A. J. Fridlund and J. T. Cacioppo, "Guidelines for human electromyographic research," *Psychophysiology*, vol. 23, pp. 567 – 589, 1986.

[34] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel, "Session independent non-audible speech recognition using surface electromyography," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico, 2005, pp. 331–336.

[35] Y. Deng, J. T. Heaton, and G. S. Meltzner, "Towards a practical silent speech recognition system," in *15th Annu. Conf. Int. Speech Communication Association*, Singapore, 2014, pp. 1164 – 1168.

[36] M. Wand, C. Schulte, M. Janke, and T. Schultz, "Array-based electromyographic silent speech interface," in *Proc. Biosignals*, 2013.

[37] R. Merletti and P. A. Parker, *Electromyography: physiology, engineering, and non-invasive applications, Chapter 4.2*. John Wiley & Sons, 2004, vol. 11.

[38] R. Netsell and B. Daniel, "Neural and mechanical response time for speech production," *J Speech Hear Res*, vol. 17, pp. 608 – 618, 1974.

[39] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *9th Int. Conf. Spoken Language Processing*, Pittsburgh, PA, USA, 2006, pp. 573 – 576.

[40] C. J. Price, "A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading," *Neuroimage*, vol. 62, no. 2, pp. 816–847, 2012.

[41] F. Jobsis, “Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters,” *Science*, vol. 198, no. 4323, pp. 1264–1267, 1977.

[42] G. Strangman, D. Boas, and J. Sutton, “Non-invasive neuroimaging using near-infrared light,” *Biol Psychiat*, vol. 52, no. 7, pp. 679–693, 2002.

[43] H. Ayaz, B. Onaral, K. Izzetoglu, P. Shewokis, R. McKendrick, and R. Parasuraman, “Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: empirical examples and a technological development,” *Front Hum Neurosci*, vol. 7, p. 871, 2013.

[44] A. von Lühmann, C. Herff, D. Heger, and T. Schultz, “Towards a wireless open source instrument: functional near-infrared spectroscopy in mobile neuroergonomics and BCI applications,” *Front Hum Neurosci*, vol. 9, no. 617, 2015.

[45] P. L. Nunez and R. Srinivasan, *Electric fields of the brain: The neurophysics of EEG*. Oxford University Press US, 2006.

[46] I. I. Goncharova, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, “EMG contamination of EEG: spectral and topographical characteristics,” *Clin Neurophysiol*, vol. 114, no. 9, pp. 1580–1593, 2003.

[47] M. De Vos, , S. Riès, K. Vanderperren, B. Vanrumste, F.-X. Alario, S. Van Huffel, and B. Burle, “Removal of muscle artifacts from EEG recordings of spoken language production.” *Neuroinformatics*, vol. 8, no. 2, pp. 135–50, 2010.

[48] L. Deecke, M. Engel, W. Lang, and H. H. Kornhuber, “Bereitschaftspotential preceding speech after holding breath,” *Exp Brain Res*, vol. 65, no. 1, pp. 219–223, 1986.

[49] P. Indefrey and W. J. M. Levelt, “The spatial and temporal signatures of word production components.” *Cognition*, vol. 92, no. 1-2, pp. 101–44, 2004.

[50] T. W. Picton, S. Bentin, P. Berg, E. Donchin, S. A. Hillyard, R. Johnson, G. A. Miller, W. Ritter, D. S. Ruchkin, M. D. Rugg, and M. J. Taylor, “Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria,” *Psychophysiology*, vol. 37, no. 2, pp. 127–152, 2000.

[51] G. Pfurtscheller and F. Lopes da Silva, “Event-related EEG/MEG synchronization and desynchronization: basic principles,” *Clin Neurophysiol*, vol. 110, no. 11, pp. 1842–1857, 1999.

[52] E. M. Maynard, C. T. Nordhausen, and R. A. Normann, “The Utah Intracortical Electrode Array: A recording structure for potential brain-computer interfaces,” *Electroen Clin Neuro*, vol. 102, no. 3, pp. 228–239, 1997.

[53] A. B. Schwartz, “Cortical neural prosthetics,” *Annu Rev Neurosci*, vol. 27, no. 1, pp. 487–507, 2004.

[54] E. Stark and M. Abeles, “Predicting movement from multiunit activity,” *J Neurosci*, vol. 27, no. 31, p. 8387, 2007.

[55] J. S. Brumberg, E. J. Wright, D. S. Andreasen, F. H. Guenther, and P. R. Kennedy, “Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex,” *Front Neurosci*, vol. 5, p. 65, 2011.

[56] A. Tankus, I. Fried, and S. Shoham, “Structured neuronal encoding and decoding of human speech features,” *Nat Commun*, vol. 3, p. 1015, 2012.

[57] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen, P. Ehirim, H. Mao, and P. R. Kennedy, “A wireless brain-machine interface for real-time speech synthesis,” *PLoS One*, vol. 4, no. 12, p. e8218, 2009.

[58] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, “Brain–computer interfaces for speech communication,” *Speech Commun*, vol. 52, no. 4, pp. 367–379, 2010.

[59] N. E. Crone, D. L. Miglioretti, B. Gordon, J. M. Sieracki, M. T. Wilson, S. Uematsu, and R. P. Lesser, “Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. I. Alpha and beta event-related desynchronization.” *Brain*, vol. 121, no. 12, pp. 2271–2299, 1998.

[60] A. Gevins, B. Cutillo, J. Desmond, M. Ward, S. Bressler, N. Barbero, and K. Laxer, “Subdural grid recordings of distributed neocortical networks involved with somatosensory discrimination,” *Electroen Clin Neuro*, vol. 92, no. 4, pp. 282–290, 1994.

[61] G. Pfurtscheller and R. Cooper, “Frequency dependence of the transmission of the EEG from cortex to scalp,” *Electroen Clin Neuro*, vol. 38, no. 1, pp. 93–96, 1975.

[62] N. Crone, A. Sinai, and A. Korzeniewska, “High-frequency gamma oscillations and human brain mapping with electrocorticography,” *Prog Brain Res*, vol. 159, pp. 275–295, 2006.

[63] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, “A review of recent advances in visual speech decoding,” *Image Vision Comput*, vol. 32, pp. 590 – 605, 2014.

[64] J. Luettin, N. A. Thacker, and S. W. Beet, “Visual speech recognition using active shape models and hidden Markov models,” in *1996 IEEE Int. Conf. Acoustics Speech Signal Processing*, vol. 2, Atlanta, GA, USA, 1996, pp. 817–820.

[65] A. Biswas, P. Sahu, and M. Chandra, “Multiple cameras audio visual speech recognition using active appearance model visual features in car environment,” *Int J Speech Tech*, vol. 19, no. 1, pp. 159–171, 2016.

[66] D. B. Li Liu, Gang Feng, “Automatic dynamic template tracking of inner lips based on CLNF,” in *2017 IEEE Int. Conf. Acoustics Speech Signal Processing*, New Orleans, LA, USA, 2017, pp. 5130 – 5134.

[67] M. Li, C. Kambhamettu, and M. Stone, “Automatic contour tracking in ultrasound images,” *Clin Linguist Phonet*, vol. 19, no. 6-7, pp. 545–554, 2005.

[68] A. Roussos, A. Katsamanis, and P. Maragos, “Tongue tracking in ultrasound images with active appearance models,” in *IEEE Int. Conf. on Image Processing*, 2009, pp. 1733–1736.

[69] D. Fabre, T. Hueber, F. Bocquelet, and P. Badin, “Tongue tracking in ultrasound images using Eigentongue decomposition and artificial neural networks,” in *16th Annu. Conf. Int. Speech Communication Association*, Dresden, Germany, 2015.

[70] I. Fasel and J. Berry, “Deep belief networks for real-time extraction of tongue contours from ultrasound during speech,” in *IEEE 20th Int. Conf. on Pattern Recognition*, 2010, pp. 1493–1496.

[71] M. Heckmann, K. Kroschel, C. Savariaux, and F. Berthommier, “DCT-based video features for audio-visual speech recognition.” in *7th Int. Conf. Spoken Language Processing*, vol. 3, Denver, CO, USA, 2002, pp. 1925–1928.

[72] C. Bregler and Y. Konig, “Eigenlips for robust speech recognition,” in *1994 IEEE Int. Conf. Acoustics Speech Signal Processing*, vol. 2, Adelaide, Australia, 1994, pp. 669–672.

[73] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone, “Eigentongue feature extraction for an ultrasound-based silent speech interface,” in *2007 IEEE Int. Conf. Acoustics Speech Signal Processing*, Honolulu, HI, USA, 2007, pp. I–1245 – I–1248.

[74] Y. LeCun and Y. Bengio, “The handbook of brain theory and neural networks,” M. A. Arbib, Ed. MIT Press, 1998, ch. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258.

[75] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, “Lipreading using convolutional neural network.” in *15th Annu. Conf. Int. Speech Communication Association*, Singapore, 2014, pp. 1149–1153.

[76] E. Tatulli and T. Hueber, “Feature extraction using multimodal convolutional neural networks for visual speech recogntiion,” in *2017 IEEE Int. Conf. Acoustics Speech Signal Processing*, New Orleans, LA, USA, 2017, pp. 2971–2975.

[77] S. Stone and P. Birkholz, “Angle correction in optopalatographic tongue distance measurements,” *IEEE Sensors J.*, vol. 17, no. 2, pp. 459–468, Jan 2017.

[78] N. Sugie and K. Tsunoda, “A Speech Prosthesis Employing a Speech Synthesizer-Vowel Discrimination from Perioral Muscle Activities and Vowel Production,” *IEEE Trans. Biomed. Eng.*, vol. BME-32, no. 7, pp. 485–490, July 1985.

[79] M. S. Morse, S. H. Day, B. Trull, and H. Morse, “Use of myoelectric signals to recognize speech,” in *11th Annu. Conf. IEEE Engineering in Medicine and Biology Society*, 1989, pp. 1793 – 1794.

[80] B. Hudgins, P. Parker, and R. Scott, “A new strategy for multifunction myoelectric control,” *IEEE Trans. Biomed. Eng.*, vol. 40, pp. 82 – 94, 1993.

[81] C. Jorgensen, D. D. Lee, and S. Agabon, “Sub auditory speech recognition based on EMG/EPG signals,” in *Int. Joint Conf. Neural Networks*, Portland, OR, USA, 2003, pp. 3128 – 3133.

[82] Y. Deng, R. Patel, J. T. Heaton, G. Colby, L. D. Gilmore, J. Cabrera, S. H. Roy, C. J. D. Luca, and G. S. Meltzner, “Disordered speech recognition using acoustic and sEMG signals,” in *10th Annu. Conf. Int. Speech Communication Association*, Brighton, UK, 2009, pp. 644 – 647.

[83] G. S. Meltzner, J. T. Heaton, Y. Deng, G. D. Luca, S. H. Roy, and J. C. Kline, “Silent speech recognition as an alternative communication device for persons with laryngectomy,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. Special Issue Biosignal-based Spoken Communication, 2017.

[84] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS," *Front Hum Neurosci*, vol. 7, p. 935, 2014.

[85] D. Heger, C. Herff, and T. Schultz, "Combining feature extraction and classification for fNIRS BCIs by regularized least squares optimization," in *36th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*. Chicago, IL, USA: IEEE, 2014, pp. 2012–2015.

[86] J. Prescott and G. Andrews, "Early and late components of the contingent negative variation prior to manual and speech responses in stutterers and non-stutterers," *Int J Psychophysiol*, vol. 2, no. 2, pp. 121–130, 1984.

[87] C. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural Networks*, vol. 22, no. 9, pp. 1334–1339, 2009.

[88] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward EEG sensing of imagined speech," in *Human Computer Interaction*, J. A. Jacko, Ed. Berlin Heidelberg: Springer-Verlag, 2009, pp. 40–48.

[89] J. S. Brumberg, D. J. Krusienski, S. Chakrabarti, A. Gunduz, P. Brunner, A. L. Ritaccio, and G. Schalk, "Spatio-temporal progression of cortical activity related to continuous overt and covert speech production in a reading task," *PLoS One*, vol. 11, pp. 1–21, 11 2016.

[90] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Readings in Speech Recognition*, A. Waibel and K.-F. Lee, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990, pp. 267–296.

[91] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Comput Linguist*, vol. 18, no. 4, pp. 467–479, 1992.

[92] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *12th Annu. Conf. Int. Speech Communication Association*, Florence, Italy, 2011, pp. 437–440.

[93] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE Int. Conf. Acoustics Speech Signal Processing*, Shanghai, China", 2016, pp. 4945 – 4949.

[94] G. I. Chiou and J.-N. Hwang, "Lipreading from color video," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1192 – 1195, 1997.

[95] L. Cappelletta and N. Harte, "Viseme definitions comparison for visual-only speech recognition," in *Proc. EUSIPCO*, 2011, pp. 2109 – 2113.

[96] T. Schultz and M. Wand, "Modeling coarticulation in large vocabulary EMG-based speech recognition," *Speech Commun*, vol. 52, no. 4, pp. 341 – 353, 2010.

[97] E. Lopez-Larraz, O. M. Mozos, J. M. Antelis, and J. Minguez, "Syllable-based speech recognition using EMG," in *32nd Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*, Buenos Aires, Argentina, 2010, pp. 4699 – 4702.

[98] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Front Neurosci*, vol. 9, no. 217, 2015.

[99] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Visuo-phonetic decoding using multi-stream and context-dependent models for an ultrasound-based silent speech interface." in *10th Annu. Conf. Int. Speech Communication Association*, Brighton, UK, 2009, pp. 640–643.

[100] M. Walliczek, F. Kraft, S.-C. Jou, T. Schultz, and A. Waibel, "Sub-word unit based non-audible speech recognition using surface electromyography," in *9th Int. Conf. Spoken Language Processing*, Pittsburgh, PA, USA, 2006, pp. 1487 – 1490.

[101] M. Wand and T. Schultz, "Analysis of phone confusion in EMG-based speech recognition," in *2011 IEEE Int. Conf. Acoustics Speech Signal Processing*, Prague, Czech Republic, 2011, pp. 757 – 760.

[102] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 5, pp. 863 – 876, 2015.

[103] K.-S. Lee, "SNR-adaptive stream weighting for audio-MES ASR," *IEEE Trans. Biom. Eng.*, vol. 55, pp. 2001 – 2010, 2008.

[104] F. H. Guenther, S. S. Ghosh, and J. A. Tourville, "Neural modeling and imaging of the cortical interactions underlying syllable production," *Brain Lang*, vol. 96, pp. 280 – 301, 2006.

[105] J.-L. Schwartz and C. Savariaux, "No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag," *PLoS Comput Biol*, vol. 10, no. 7, p. e1003743, 2014.

[106] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *J Acoust Soc Am*, vol. 121, no. 2, pp. 723–742, 2007.

[107] E. D. Petajan, "Automatic lipreading to enhance speech recognition (speech reading)," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1984.

[108] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, p. 23, 2004.

[109] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Commun*, vol. 52, no. 4, pp. 288–300, 2010.

[110] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[111] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 2130–2134.

[112] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *2016 IEEE Int. Conf. Acoustics Speech Signal Processing*, Shanghai, China", 2016.

[113] G. Gravier, G. Potamianos, and C. Neti, "Asynchrony modeling for audio-visual speech recognition," in *Proc. Int. Conf. on Human Lang. Tech. Research*, 2002, pp. 1–6.

[114] J. N. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, "DBN based multi-stream models for audio-visual speech recognition," in *2004 IEEE Int. Conf. Acoustics Speech Signal Processing*, vol. 1, Montreal, Quebec, Canada, 2004, pp. 993–996.

[115] A. Wrench and K. Richmond, "Continuous speech recognition using articulatory data," *6th Int. Conf. on Spoken Language Processing*, pp. 145–148, 2000.

[116] P. Heracleous, P. Badin, G. Bailly, and N. Hagita, "A pilot study on augmented speech communication based on electro-magnetic articulography," *Pattern Recogn Lett*, vol. 32, no. 8, pp. 1119–1125, 2011.

[117] M. Fagan, S. Ell, J. Gilbert, E. Sarrazin, and P. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Med Eng Phys*, vol. 30, no. 4, pp. 419 – 425, 2008.

[118] A. D. C. Chan, K. B. Englehart, B. Hudgins, and D. F. Lovely, "Myoelectric signals to augment speech recognition," *Med Biol Eng Comput*, vol. 39, pp. 500 – 506, 2001.

[119] M. Wand and T. Schultz, "Towards real-life application of EMG-based speech recognition by using unsupervised adaptation," in *15th Annu. Conf. Int. Speech Communication Association*, Singapore, 2014.

[120] M. Wand and J. Schmidhuber, "Deep neural network frontend for continuous EMG-based speech recognition," in *17th Annu. Conf. Int. Speech Communication Association*, San Francisco, CA, USA, 2016, pp. 3032 – 3036.

[121] E. Formisano, F. De Martino, M. Bonte, and R. Goebel, ""Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech," *Science*, vol. 322, no. 5903, pp. 970–973, 2008.

[122] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just, "Predicting human brain activity associated with the meanings of nouns," *Science*, vol. 320, no. 5880, pp. 1191–1195, 2008.

[123] A. Huth, W. de Heer, T. Griffiths, F. Theunissen, and J. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.

[124] C. Herff, F. Putze, D. Heger, C. Guan, and T. Schultz, "Speaking mode recognition from functional near infrared spectroscopy," in *34th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*, San Diego, CA, USA, 2012, pp. 1715–1718.

[125] C. Herff, D. Heger, F. Putze, C. Guan, and T. Schultz, "Cross-subject classification of speaking modes using fNIRS," in *Neural Information Processing*, ser. Lecture Notes in Computer Science, T. Huang, Z. Zeng, C. Li, and C. Leung, Eds., vol. 7664. Springer Berlin Heidelberg, 2012, pp. 417–424.

[126] P. Suppes, Z. Lu, and B. Han, "Brain wave recognition of words," *Proc Natl Acad Sci USA*, vol. 94, no. 26, p. 14965, 1997.

[127] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex." *PLoS Biol*, vol. 10, no. 1, p. e1001251, 2012.

[128] N. Yoshimura, A. Nishimoto, A. N. Belkacem, D. Shin, H. Kambara, T. Hanakawa, and Y. Koike, "Decoding of covert vowel articulation us-

ing electroencephalography cortical currents," *Front Neurosci*, vol. 10, 2016.

[129] A. Porbadnigk, M. Wester, J. P. Callies, and T. Schultz, "EEG-based speech recognition - impact of temporal effects," in *2nd Int. Conf. Bio-inspired Systems Signal Processing*, Porto, Portugal, 2009.

[130] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "Direct classification of all American English phonemes using signals from functional speech motor cortex." *J Neural Eng*, vol. 11, no. 3, p. 035015, 2014.

[131] T. Blakely, K. J. Miller, R. P. N. Rao, M. D. Holmes, and J. G. Ojemann, "Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids." in *30nd Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*, 2008, pp. 4964–7.

[132] K. E. Bouchard and E. F. Chang, "Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography," in *36th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*. Chicago, IL, USA: IEEE, 2014, pp. 6782–6785.

[133] S. Kellis, K. Miller, K. Thomson, R. Brown, P. House, and B. Greger, "Decoding spoken words using local field potentials recorded from the cortical surface." *J Neural Eng*, vol. 7, no. 5, p. 056007, 2010.

[134] F. Lotte, J. S. Brumberg, P. Brunner, A. Gunduz, A. L. Ritaccio, C. Guan, and G. Schalk, "Electrocorticographic representations of segmental features in continuous speech," *Front Hum Neurosci*, vol. 9, p. 97, 2015.

[135] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Audio, Speech, Language Process.*, vol. 6, no. 2, pp. 131–142, 1998.

[136] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Audio, Speech, Language Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[137] A. Toth, M. Wand, and T. Schultz, "Synthesizing speech from electromyography using voice transformation techniques," in *10th Annu. Conf. Int. Speech Communication Association*, Brighton, UK, 2009.

[138] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Comput Speech Lang*, vol. 36, pp. 274–293, 2016.

[139] C. T. Kello and D. C. Plaut, "A neural network model of the articulatory-acoustic forward mapping trained on recordings of articulatory parameters," *J Acoust Soc Am*, vol. 116, no. 4, pp. 2354–2364, 2004.

[140] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-time control of an articulatory-based speech synthesizer for brain computer interfaces," *PLoS Comput Biol*, vol. 12, no. 11, p. e1005119, 2016.

[141] L. Diener, M. Janke, and T. Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *Int. Joint Conf. Neural Networks*, Killarney, Ireland, 2015, pp. 1–7.

[142] M. Janke and L. Diener, "EMG-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. Special Issue Biosignal-based Spoken Communication, 2017.

[143] J. S. Brumberg, J. D. Burnison, and K. M. Pitt, "Using motor imagery to control brain-computer interfaces for communication," in *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*. Spring International Publishing Switzerland, 2016.

[144] C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz, "Towards direct speech synthesis from ECoG: A pilot study," in *38th Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society*, Orlando, FL, USA, 2016.

[145] M. Wand, M. Janke, and T. Schultz, "The EMG-UKA corpus for electromyographic speech processing," in *15th Annu. Conf. Int. Speech Communication Association*, Singapore, 2014.

[146] F. Gibbon and A. Lee, "Electropalatography for older children and adults with residual speech errors," in *Seminars in Speech and Language*, vol. 36. Thieme Medical Publishers, 2015, pp. 271–282.

[147] M. Janke, M. Wand, and T. Schultz, "A spectral mapping method for EMG-based recognition of silent speech," in *Proc. B-INTERFACE*, 2010, pp. 22 – 31.

[148] M. Wand, M. Janke, and T. Schultz, "Tackling speaking mode varieties in EMG-based speech recognition," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2515 – 2526, 2014.

[149] M. Wand and J. Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," in *18th Annu. Conf. Int. Speech Communication Association*, 2017, pp. 3662 – 3666.

[150] H. McGurk and J. Macdonald, "Hearing lips and seeing voices." *Nature*, vol. 264, pp. 691–811, 1976.

[151] F. Gibbon, "Bibliography of electropalatographic (EPG) studies in English (1957-2013)," Staeno 2013-05-21. http://www. articulateinstruments.com/EPGrefs.pdf, Report, 2011.

[152] M. Cavin, "The use of ultrasound biofeedback for improving English /r/," *Working Papers of the Linguistics Circle*, vol. 25, no. 1, pp. 32–41, 2015.

[153] J. Cleland, J. M. Scobbie, and A. A. Wrench, "Using ultrasound visual biofeedback to treat persistent primary speech sound disorders," *Clin Linguist Phonet*, pp. 1–23, 2015.

[154] J. Perkell, M. Matthies, H. Lane, F. Guenther, R. Wilhelms-Tricaricoa, J. Wozniaka, and P. Guioda, "Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models," *Speech Commun*, vol. 22, pp. 227 – 250, 1997.

[155] J. A. Tourville, K. J. Reilly, and F. H. Guenther, "Neural mechanisms underlying auditory feedback control of speech," *Neuroimage*, vol. 32, pp. 1429 – 1443, 2008.

[156] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 10, no. 3, pp. 72–74, 2003.

[157] B. J. Betts, K. Binsted, and C. Jorgensen, "Small-vocabulary speech recognition using surface electromyography," *Interact Comput*, vol. 18, no. 6, pp. 1242–1259, 2006.

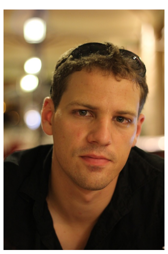
[158] C. Jorgensen and S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Commun*, vol. 52, no. 4, pp. 354–366, 2010.

[159] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. N. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex." *Front Neuroeng*, vol. 7, no. 14, 2014.

[160] S. Martin, P. Brunner, I. Iturrate, J. d. R. Millán, G. Schalk, R. T. Knight, and B. N. Pasley, "Word pair classification during imagined speech using direct brain recordings," *Sci Rep-UK*, vol. 6, p. 25803, 2016.

Tanja Schultz (M'01-SM'17) received her doctoral and diploma degree in Informatics from University of Karlsruhe, Germany, in 2000 and 1995. She joined Carnegie Mellon University, Pittsburgh, PA in 2000 and is an adjunct Research Professor at the Language Technologies Institute. From 2007 to 2015 she was a Full Professor in Informatics at the Karlsruhe Institute of Technology (KIT) in Germany before she became a Professor for Cognitive Systems at the University of Bremen, Germany in April 2015. Since 2007, she directs the Cognitive Systems Lab, where her research activities focus on the processing, recognition, and interpretation of biosignals for human-centered technologies and applications. She is an ISCA Fellow and member of the European Academy of Sciences and Arts.

Michael Wand received a diploma degree in Mathematics and a doctoral degree in Computer Science from Karlsruhe Institute of Technology (formerly University of Karlsruhe). His dissertation, defended in 2014, substantially contributed to a Silent Speech recognizer based on electromyography. Since 2014, he is a postdoctoral researcher at the Swiss AI Lab IDSIA, where he continues to investigate Silent Speech recognition, as well as lipreading, biosignal processing for physiological tasks, and applications of neural networks.



Thomas Hueber received an engineering degree in Electronics, Telecommunication and Computer Science from CPE Lyon (France) and a M.Sc. in Image Processing from University of Lyon in 2006. He worked towards his Ph.D. in Computer Science on *silent speech interfaces* and obtained it from Pierre and Marie Curie University (Paris) in 2009. In 2010, he joined GIPSA-lab (Grenoble, France) as a post-doctoral researcher and became a tenured CNRS researcher in 2011. His research activities deal with multimodal speech processing (recognition, synthesis, conversion), with a special interest in speech biosignals (such as the articulatory movements, muscle and brain activities), their modeling using machine learning techniques, and their use in assistive technologies.

Dean Krusienski (M'01-SM'14) received the B.S., M.S., and Ph.D. degrees in electrical engineering from The Pennsylvania State University. He conducted his post-doctoral research in the Brain-Computer Interface (BCI) Laboratory, Wadsworth Center of the New York State Department of Health. He is currently a Professor of Electrical and Computer Engineering at Old Dominion University (ODU), Norfolk, VA, USA, where he directs the Advanced Signal Processing in Engineering and Neuroscience (ASPEN) Lab. He is also the Graduate Program Director and a founding member of the biomedical engineering program at ODU. His research interests include biomedical signal processing, pattern recognition, brain-computer interfaces, and neural engineering.

Christian Herff received his Diploma in Informatics from Karlsruhe Institute of Technology in 2011 and his Ph.D. from University of Bremen in 2016. He is currently a post-doctoral researcher at the Cognitive Systems Lab at the University of Bremen. His research focuses on the interpretation and analysis of neural signals using machine learning techniques.

Jonathan Brumberg received the B.S. and B.A. degrees in Computer and Information Sciences, and Philosophy in 2002 from the University of Delaware, and the Ph.D. degree in Cognitive and Neural Systems from Boston University in 2009. He completed post-doctoral research and served as a Research Assistant Professor in the Department of Speech-Language-Hearing Sciences at Boston University focusing on brain-computer interfaces for speech synthesis. He is currently an Assistant Professor in the Department of Speech-Language-Hearing at the University of Kansas (KU) where he directs the Speech and Applied Neuroscience Lab. His research interests are in the neurological mechanisms underlying speech and communication and their use in brain-computer interfaces for instantaneous speech output.