

Non-invasive silent phoneme recognition using microwave signals

Peter Birkholz, *Associate Member, IEEE*, Simon Stone, Klaus Wolf, and Dirk Plettemeier, *Member, IEEE*

Abstract—Besides the recognition of audible speech, there is currently an increasing interest in the recognition of silent speech, which has a range of novel applications. A major obstacle for a wide spread of silent-speech technology is the lack of measurement methods for speech movements that are convenient, non-invasive, portable, and robust at the same time. Therefore, as an alternative to established methods, we examined to what extent different phonemes can be discriminated from the electromagnetic transmission and reflection properties of the vocal tract. To this end, we attached two Vivaldi antennas on the cheek and below the chin of two subjects. While the subjects produced 25 phonemes in multiple phonetic contexts each, we measured the electromagnetic transmission spectra from one antenna to the other, and the reflection spectra for each antenna (radar), in a frequency band from 2-12 GHz. Two classification methods (k-nearest neighbors and linear discriminant analysis) were trained to predict the phoneme identity from the spectral data. With linear discriminant analysis, cross-validated phoneme recognition rates of 93% and 85% were achieved for the two subjects. Although these results are speaker- and session-dependent, they suggest that electromagnetic transmission and reflection measurements of the vocal tract have great potential for future silent-speech interfaces.

Index Terms—Silent-speech interface

I. INTRODUCTION

WHILE the recognition of audible speech by means of acoustic signals is well established, the recognition of silent speech, i.e., speech without or with hardly audible sound, is a rather new field of research [1]. Silent-speech recognition is still at an experimental stage, but has the potential for a range of new applications. For example, it could enable silent telephone conversations or silent queries to spoken dialog systems, which would provide privacy and prevent the disturbance of nearby people in public places. In contrast to conventional speech recognition, this would also work under very noisy conditions. In the medical domain, silent speech technology could provide a substitute voice for people whose larynx has been removed. For these people, the silent speech movements could be captured and converted into audible speech by means of speech synthesis.

Because the recognition of silent speech relies on the acquisition and interpretation of speech-related biosignals, it is also referred to as biosignal-based speech recognition [2]. Potentially useful biosignals are brain signals, muscle signals, and movement signals, all of which require dedicated measurement methods. For a wide spread and acceptance of

silent speech technology, the methods for the acquisition of the according biosignals should be non-invasive and easy to apply. Furthermore, the recording devices should be portable and the signals should be robust against distortions and allow a good discrimination of speech sounds.

Currently, the main measurement techniques for silent-speech recognition include permanent magnet articulography (PMA) [3], electromagnetic articulography (EMA) [4], video imaging of the face [5], video imaging of the lips in combination with ultrasound (US) imaging of the tongue [6], [7], [8], ultrasound Doppler sensing of facial movements [9], surface electromyography (sEMG) [10], [11], [12], [13], electropalatography (EPG) [14], electro-optical stomatography (EOS) [15], [16], intercranial wire microelectrodes [17], and electrocorticography (ECoG) [18]. Besides the recognition of speech, many of these methods have also been applied to the direct synthesis of speech, i.e., biosignal-based speech synthesis, which requires the mapping of the biosignals to vocoder parameters. Examples are the direct synthesis based on PMA signals [19], [20], [21], EMA signals [22], tongue US signals [23], video images of the face [24], ultrasound Doppler signals of facial movements [25], and EMG signals [26], [27], [28].

Many of the above measurement methods (PMA, EMA, EPG, EOS, and ECoG) are invasive and hence not suited for a wide spread of silent speech interfaces except in the medical domain. Surface EMG, ultrasound Doppler sensing of facial movements, and the (ultrasound) imaging of the tongue and the lips are non-invasive methods, but they have drawbacks of their own. One of the main problems of sEMG is the inter-session variability, which is caused by the sensitivity of sEMG measurements to the electrode-skin impedance and to placement variations of the sensors [29], among others. The main problems with US imaging of the tongue are that the images are rather noisy and the fixation of the ultrasound probe at a constant position is awkward [8]. Imaging of the face or lips using an external camera, as well as ultrasound Doppler sensing of facial movements, have the drawback that the data provide very limited information about the internal state of the vocal tract.

In contrast to the above methods, electromagnetic (EM) waves have been rarely considered to acquire speech-related data. The first who applied EM-wave sensors to the monitoring of speech movements were Holzrichter et al. in 1998 [30]. They used low-power radar sensors that were placed close to the skin of the face and worked at frequencies centered at 2.3 GHz in order to monitor the motion of individual organs like the tongue or the vocal folds. They proposed the use of these sensors for silent speech recognition, but did not

P. Birkholz (peter.birkholz@tu-dresden.de) and S. Stone are with the Institute of Acoustics and Speech Communication, TU Dresden, Germany. K. Wolf and D. Plettemeier are with the Institute of Communication Technology, TU Dresden, Germany.

Manuscript received XXX; revised XXX.

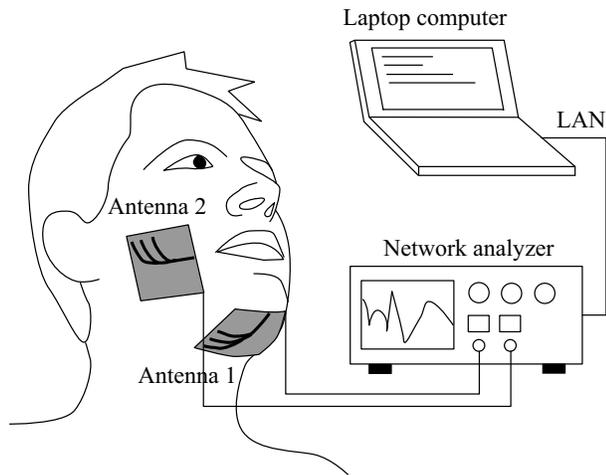


Fig. 1. Setup of the measurement system.

make actual recognition experiments. Their speech sensing approach was presented in an additional paper in 2009 [31], but seems to have gone rather unnoticed by the speech research community. More recently, Eid and Wallace [32] presented an ultra-wideband (UWB) radar system (500 MHz to 10 GHz) for speech sensing using a single antenna that was placed a few centimeters in front of the mouth. In a proof-of-concept, they showed that it was possible to discriminate 10 words spoken by a single speaker by template matching using the complex reflection coefficients. A similar radar system for contactless silent speech recognition was proposed in 2016 by Shin and Seo [33]. They used a commercial off-the-shelf impulse-radio UWB radar module with separate transmitter and receiver antennas that were placed 10 cm in front of the face. In an experiment with five speakers, they achieved an average speaker-dependent word recognition rate of 85 % for 10 isolated words.

In the present study, we propose a UWB speech sensing system that uses two antennas that are placed directly on the cheek and below the chin of a speaker. The state of the vocal tract is captured in terms of the transmission spectrum between the two antennas and the individual reflection spectra of the antennas, for a frequency band of 2-12 GHz. These spectral data are used as “articulatory features” to train two simple classifiers for the discrimination of phonemes. It is shown that a linear discriminant analysis based on these features can achieve very high phoneme recognition rates despite strong contextual variation of the speech sounds.

II. METHOD

A. Data Acquisition

The goal of this study was to find out to what extent different phonemes can be discriminated from the electromagnetic transmission and reflection properties of the vocal tract. To this end, two identical flat foil antennas ($45 \times 40 \text{ mm}^2$) were attached to the speaker’s face using adhesive tape: one below the chin and one on the right cheek, as illustrated in Fig. 1. The antennas were modified antipodal Vivaldi antennas on

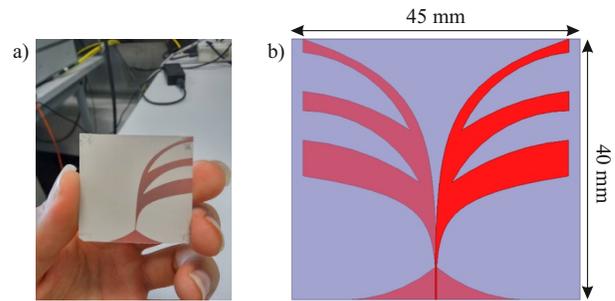


Fig. 2. The antipodal Vivaldi antenna that was used for the experiments. a) Photo of the antenna. b) Layout with the copper tracks on the front side (red) and the back side (matte red).

a flexible substrate (see Fig. 2) that were optimized in size, weight, and bandwidth for on-tissue applications [34].

Both antennas were connected to the two ports of a dual-port vector network analyzer (PNA Series Network Analyzer E8364B by Agilent Technologies) that measures signal amplitude and phase in terms of complex scattering parameters. For the measurements, we used sweep signals with a power band of 2-12 GHz. The lower band limit of 2 GHz was chosen according to the size of the antenna (lower frequencies would have required larger antennas), and the upper band limit of 12 GHz was chosen such that the attenuation of the EM waves in the tissue became not too strong.

The network analyzer was configured to capture an articulatory state of the vocal tract as follows. First, antenna 1 emitted a linear sweep signal (2-12 GHz) with a duration of 6.03 ms and a power of 1 mW (which is far below the transmission power levels of current smartphones). The response of the head and vocal tract to this signal was captured by both antenna 1 (scattering parameter $S_{11}(\omega)$ with the angular frequency ω) and antenna 2 (scattering parameter $S_{21}(\omega)$). Then, antenna 2 emitted the same source signal and the response was recorded by both antenna 1 (scattering parameter $S_{12}(\omega)$) and antenna 2 (scattering parameter $S_{22}(\omega)$). The total recording time for one vocal tract configuration was about 12 ms. From the perspective of the network analyzer, the vocal tract is a passive two-port network that is reciprocal with respect to its transmission properties, i.e., the frequency-dependent damping and delay of the signals from one antenna to the other do not depend on the direction, i.e., $S_{12}(\omega) = S_{21}(\omega)$. Therefore, one articulation measurement consisted of the two reflection measurements $S_{11}(\omega)$ and $S_{22}(\omega)$, and the transmission measurement $S_{21}(\omega)$, i.e., a total of three complex spectra. Each individual spectrum was sampled in terms of 201 discrete frequency components (which is the default setting of the network analyzer), i.e., $S_{11/22/21}(n)$, where $n = 1 \dots 201$ is the frequency index. Hence, 603 complex numbers were obtained during each measurement.

The measurement of the three spectra was triggered by a keystroke on a laptop computer that was connected to the network analyzer via a local area network (LAN). After each measurement, the spectral data obtained by the network analyzer were displayed on the network analyzer screen, and transferred to and saved on the laptop computer. Due to

TABLE I
PSEUDOWORDS OF THE CORPUS, WHICH WERE PRODUCED FOR ALL CONTEXT CONSONANTS $C \in \{/b,d,g,l,r,f,s,j,m,n/\}$ AND ALL CONTEXT VOWELS $V \in \{/a:,e:,i:,o:,u:,ɛ:,ø:,y:/\}$. THE UNDERLINED TARGET PHONEMES WERE SUSTAINED FOR 3 SECONDS TO ALLOW STABLE MEASUREMENTS.

Tense vowels	Lax vowels	Consonants
$/C\underline{a:}d\alpha/$	$/C\underline{i}t\alpha/$	$/V\underline{b}V/$
$/C\underline{e:}d\alpha/$	$/C\underline{ɛ}t\alpha/$	$/V\underline{d}V/$
$/C\underline{i:}d\alpha/$	$/C\underline{a}t\alpha/$	$/V\underline{g}V/$
$/C\underline{o:}d\alpha/$	$/C\underline{ɔ}t\alpha/$	$/V\underline{l}V/$
$/C\underline{u:}d\alpha/$	$/C\underline{u}t\alpha/$	$/V\underline{r}V/$
$/C\underline{ɛ:}d\alpha/$	$/C\underline{y}t\alpha/$	$/V\underline{f}V/$
$/C\underline{ø:}d\alpha/$	$/C\underline{ɔ}t\alpha/$	$/V\underline{s}V/$
$/C\underline{y:}d\alpha/$		$/V\underline{j}V/$
		$/V\underline{m}V/$
		$/V\underline{n}V/$

the relatively slow preprocessing and display of the acquired signals on the network analyzer, each individual measurement (from the keystroke to the storage of the data on the laptop) took about 1 s. Therefore, in order to obtain unbiased spectra, the subjects were asked to sustain the target phonemes in the pseudowords (see below) for about 3 s. For the plosives, which are highly transient in nature, the subjects were asked to sustain the vocal tract configuration with the maximal degree of supraglottal closure. As soon as a subject fully articulated the target phoneme, the measurement was manually triggered by the experimenter.

The main difference between our system and the previously proposed EM-based systems [31], [33], [32] is that we not only measured the reflections of EM waves from the vocal tract (radar mode), but in addition the transmission of the EM waves through the vocal tract from one antenna to the other. The results (Sec. III) demonstrate that the transmission spectrum alone contains more information for the discrimination of phonemes than individual reflection spectra. This advantage comes at the expense that the antennas need to be close/in contact to the skin, while the previous EM-based systems worked completely contactless.

B. Subjects and Corpus

Two male native German speakers (26 and 39 years old) participated in the experiment. None of them reported any history of speech or hearing disorders. The task of the subjects was to produce 230 utterances (audible speech), each consisting of the German article “Eine” followed by a pseudoword that contained a target phoneme. In order to capture the articulatory variation of the speech sounds, each target phoneme was embedded in multiple pseudowords with different context phonemes. The list of pseudowords is given in Table I. The target phonemes were the German tense vowels $/a:,e:,i:,o:,u:,ɛ:,ø:,y:/$, the lax vowels $/ɪ,ɛ,a,ɔ,ʊ,y,æ/$, and the consonants $/b,d,g,l,r,f,s,j,m,n/$. From the consonants that come in voiced-voiceless pairs, only either the voiced or the voiceless one has been included in the target phoneme set, because the voicing information is not available in silent

speech. Each target vowel occurred in the context of each of the consonants $/b,d,g,l,r,f,s,j,m,n/$, and each target consonant occurred in the context of each of the (tense) vowels $/a:,e:,i:,o:,u:,ɛ:,ø:,y:/$. Hence, in total we measured 10 samples of each of 15 vowels, and 8 samples of each of 10 consonants per speaker. Each speaker produced all utterances within the same session, i.e., with identical antenna placements.

C. Classification Experiments

To assess the possibility to discriminate the 25 target phonemes by means of their EM spectra, we applied two classification methods: k -nearest neighbors (kNN) as a non-parametric classification method, and linear discriminant analysis (LDA) as a parametric classification method. These rather basic classifiers were selected because they usually work well with rather small amounts of data like those available in the present study (compared to deep learning methods), and they have no hyperparameters (LDA) or just one hyperparameter (kNN) that need to be tuned for the best possible results. Here, the functions `fitcknn` and `fitcdiscr` of MATLAB R2016b were used as implementations of kNN and LDA, respectively, using their default settings (apart from the number of neighbors for the kNN method, see below). Each of these methods was trained and tested with different feature vectors as input. To assess the contribution of different subsets of spectral data to the classification performance, the following feature vectors were examined:

- the 201 spectral magnitudes of $S_{11}(n)$,
- the 201 spectral magnitudes of $S_{22}(n)$,
- the 201 spectral magnitudes of $S_{21}(n)$,
- the 402 spectral magnitudes of $S_{11}(n)$ and $S_{22}(n)$,
- the 402 spectral magnitudes of $S_{11}(n)$ and $S_{21}(n)$,
- the 402 spectral magnitudes of $S_{22}(n)$ and $S_{21}(n)$,
- the 603 spectral magnitudes of $S_{11}(n)$, $S_{22}(n)$, and $S_{21}(n)$.

For the same combinations of $S_{11}(n)$, $S_{22}(n)$, and $S_{21}(n)$ as in the list above, we also tested feature vectors formed of the real and imaginary parts of the spectral components, instead of the spectral magnitudes, to find out whether the spectral phase contributes to the classification performance. These feature vectors had twice the length of the corresponding magnitude feature vectors. Finally, to find out whether the lower or the upper part of the 2-12 GHz range contributes more to the classification performance, we applied the LDA to feature vectors that contained only the first 100 spectral magnitudes of S_{11} , S_{22} , and S_{21} (2-7 GHz range), and only the last 100 spectral magnitudes (7-12 GHz range).

In the recorded data, the magnitudes of the transmission spectra $S_{21}(n)$ were significantly smaller than the magnitudes of the reflection spectra $S_{11}(n)$ and $S_{22}(n)$. To ensure that all features of the input vectors had a comparable range (and hence a comparable weight for the classification), all transmission spectra $S_{21}(n)$ were scaled up by a factor of 400.

The performance of both classifiers in combination with all types of feature vectors was assessed by leave-one-out cross-validation. This means that for each subject and combination

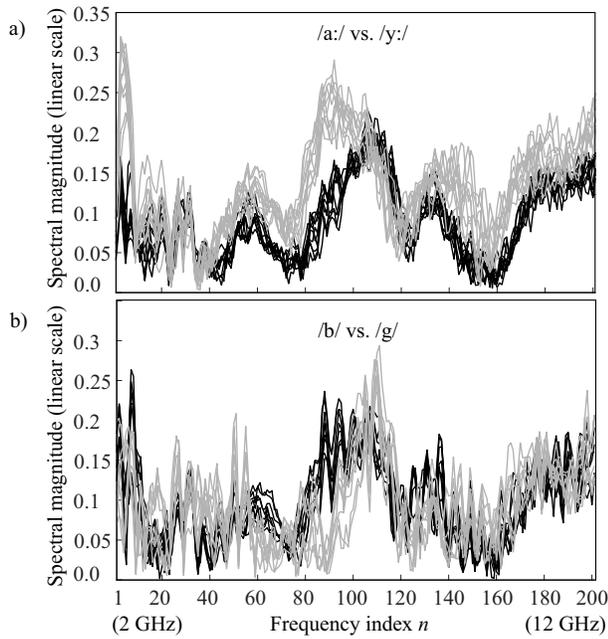


Fig. 3. a) Transmission spectra $S_{21}(n)$ (with 201 frequency points each) for each of the 10 samples of the vowels /a:/ (black) and /y:/ (gray) of speaker 1. b) Transmission spectra $S_{21}(n)$ for each of the 8 samples of the consonants /b/ (black) and /g/ (gray) of speaker 1.

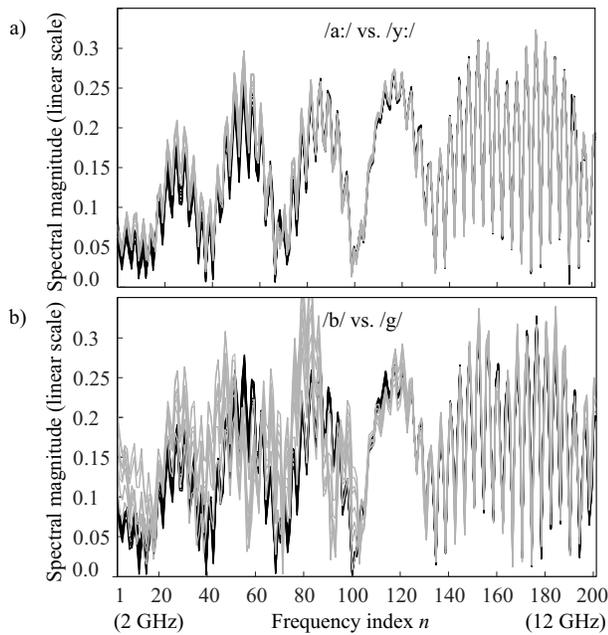


Fig. 4. a) Reflection spectra $S_{11}(n)$ (with 201 frequency points each) for each of the 10 samples of the vowels /a:/ (black) and /y:/ (gray) of speaker 1. b) Reflection spectra $S_{11}(n)$ for each of the 8 samples of the consonants /b/ (black) and /g/ (gray) of speaker 1.

of classifier and feature vector, each of the 230 speech items has been individually used for testing the classifier trained with the 229 other items. The recognition rate in percent was calculated as $(x/230) \cdot 100\%$, where x was the number of correctly classified items. For the kNN classifier, the number of neighbors k was systematically varied for each feature vector condition and only the best results are reported.

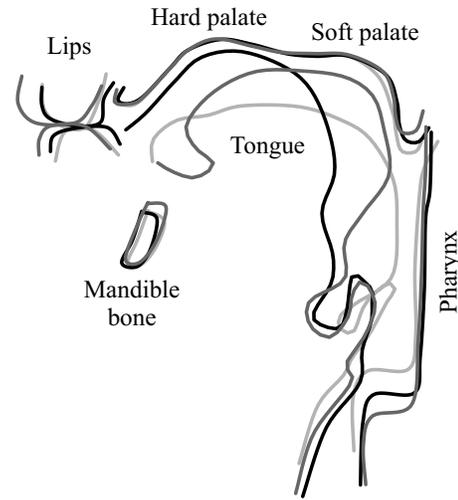


Fig. 5. Vocal tract contours of a male speaker producing the consonant /b/ in the context of the vowels /a:/ (light gray), /i:/ (black), and /u:/ (dark gray).

III. RESULTS AND DISCUSSION

A. Visual Data Analysis

As representative examples of the captured data, Fig. 3 and 4 show magnitude spectra from transmission and reflection measurements, respectively. Figure 3a shows the magnitude spectra of the transmissions from antenna 1 to antenna 2 for the 10 samples of the vowels /a:/ (black) and /y:/ (gray, as in the German word “Tüte” /ty:tə/, engl.: bag), and Fig. 3b shows the transmission spectra for the 8 samples of the consonants /b/ (black) and /g/ (gray). Figure 4 shows the magnitude spectra of the reflections $S_{11}(n)$ for the same phonemes. The spectral differences between /a:/ and /y:/ are well visible, especially for $|S_{21}(n)|$, while the spectral variation of the samples of the same vowel are relatively small. The reason for the rather pronounced spectral differences between /a:/ and /y:/ is that they differ along multiple articulatory dimensions, e.g., with respect to tongue height, lip protrusion and lip opening. For vowels that are more similar from an articulatory point of view, the spectral differences were accordingly less pronounced. In contrast to the vowels, the context-dependent variation of the consonant spectra was generally stronger, both for transmission and reflection spectra. The stronger spectral variation of the consonants is most likely caused by their stronger coarticulatory variation. As an example, Fig. 5 shows the midsagittal vocal tract shapes of /b/ spoken in the symmetric context of the vowels /a:/, /i:/, and /u:/, as determined from dynamic magnetic resonance images of the vocal tract [35]. For these /b/ samples, the only common articulatory feature are the closed lips, while the tongue position strongly varies and essentially mirrors the tongue shape of the context vowel.

For the further visual analysis of the data, we performed a dimensionality reduction of the high-dimensional spectra to a two-dimensional map. Initially, we performed a linear principal component analysis and mapped the feature vectors on the first two principal components. However, in the resulting 2D plot, the speech samples poorly grouped according to their phoneme class. This suggested that the data rather lie

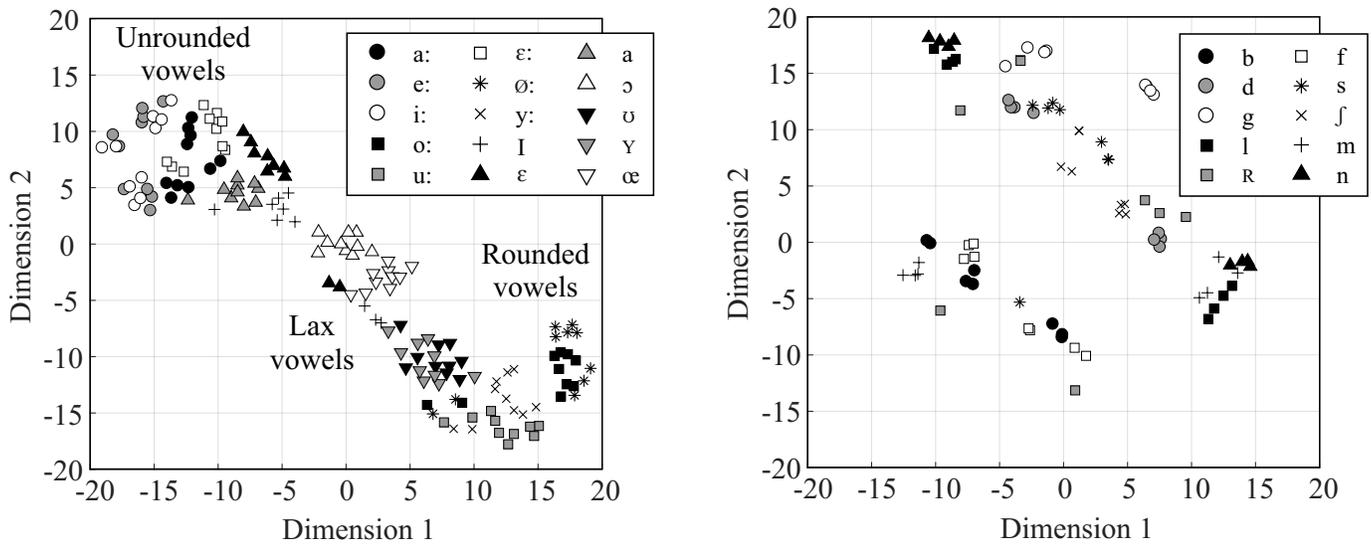


Fig. 6. 2D visualization of the 603-dimensional phone feature vectors (concatenated magnitude spectra of the one transmission measurement and the two reflection measurements of speaker 1) using the method of t-Distributed Stochastic Neighbor Embedding (t-SNE) [36]. There are 10 samples for each vowel and 8 samples for each consonant. For clarity, vowels (left) and consonants (right) are shown separately.

on a *non-linear* manifold within the high-dimensional feature space. Accordingly, we performed a t-Distributed Stochastic Neighbor Embedding (t-SNE, [36]), which is the state-of-the-art for non-linear dimensionality reduction, using the MATLAB implementation by van der Maaten [37]. Because t-SNE is a stochastic algorithm, the result depends on the random initialization of the map points and somewhat differs after each run. A representative result for our data is shown in Fig. 6, where the data points for vowels and consonants have been plotted in two separate maps for clarity. The input data were the 603-dimensional feature vectors formed of the magnitude spectra $|S_{11}(n)|$, $|S_{22}(n)|$, and $|S_{21}(n)|$ of speaker 1. As Fig. 6a shows, the samples of the individual vowels form clusters. There is a certain overlap of clusters, but for these, the according phonemes have similar articulatory features. For example, the rounded vowels, the unrounded vowels, and the lax vowels are located in well separable regions of the map. As shown in Fig. 6b, the consonants do not cluster as well as the vowels. This is likely due to the stronger contextual variation of the articulation of consonants, as discussed above. Nonetheless, the positions of the consonants on the map do hardly overlap with the vowel samples.

B. Classification Results

The accuracy of the phoneme classification using the two methods LDA and kNN in combination with the different high-dimensional feature vectors are summarized in Fig. 7. These plots allow a number of observations. With regard to the spectral data included in the feature vectors, recognition rates usually increase when more data is included. The highest recognition rates are achieved when $S_{11}(n)$, $S_{22}(n)$, and $S_{21}(n)$ are jointly used. This means that all three spectra provide unique information that help to discriminate the speech sounds. However, the contribution of the three spectra to the recognition rate differs. For the feature vectors that contain the

data of only $S_{11}(n)$, only $S_{22}(n)$, or only $S_{21}(n)$, phoneme recognition rates are generally highest for $S_{21}(n)$. This means that the *transmission* spectrum contains the most important information for phoneme recognition. When any two spectra or all three spectra are combined to form the feature vector, the recognition rates are higher than for the included individual spectra alone. Furthermore, it is generally beneficial to use not only the spectral magnitude as features (dark gray bars), but the full spectral information in terms of the real and imaginary parts of the spectral components (light gray bars). This is especially evident in the LDA results when only one individual spectrum is used for the feature vector. When two or all three spectra are used as features, the benefit of using the real and imaginary parts as separate features becomes less pronounced. For the kNN classifier, the recognition rates may even become worse (here for speaker 2).

With regard to the classification method, the performance depends on the dimensionality of the feature vectors. For the feature vectors with the lowest number of dimension, i.e., with the 201 spectral magnitudes of either $S_{11}(n)$, $S_{22}(n)$, or $S_{21}(n)$, kNN performs better than LDA. For feature vectors that combine two or three spectra and/or use the real and imaginary parts of the spectral components as separate dimensions, LDA generally achieves a higher recognition rate than kNN. Hence, while kNN suffers from the “curse of dimensionality” for an increasing number of feature space dimensions [38], LDA benefits from additional dimensions as they facilitate the detection of linear hyperplanes that separate the classes.

With regard to the two examined frequency bands, the feature vectors that contained only the frequency components from 2-7 GHz achieved a higher recognition rate (73.9 % and 77.8 % for the two subjects) than the feature vectors that contained only the frequency components from 7-12 GHz (55.2 % and 59.1 %). This means that a potential extension of the 2-12 GHz range would be most useful towards lower frequencies. However, antennas for lower frequencies would

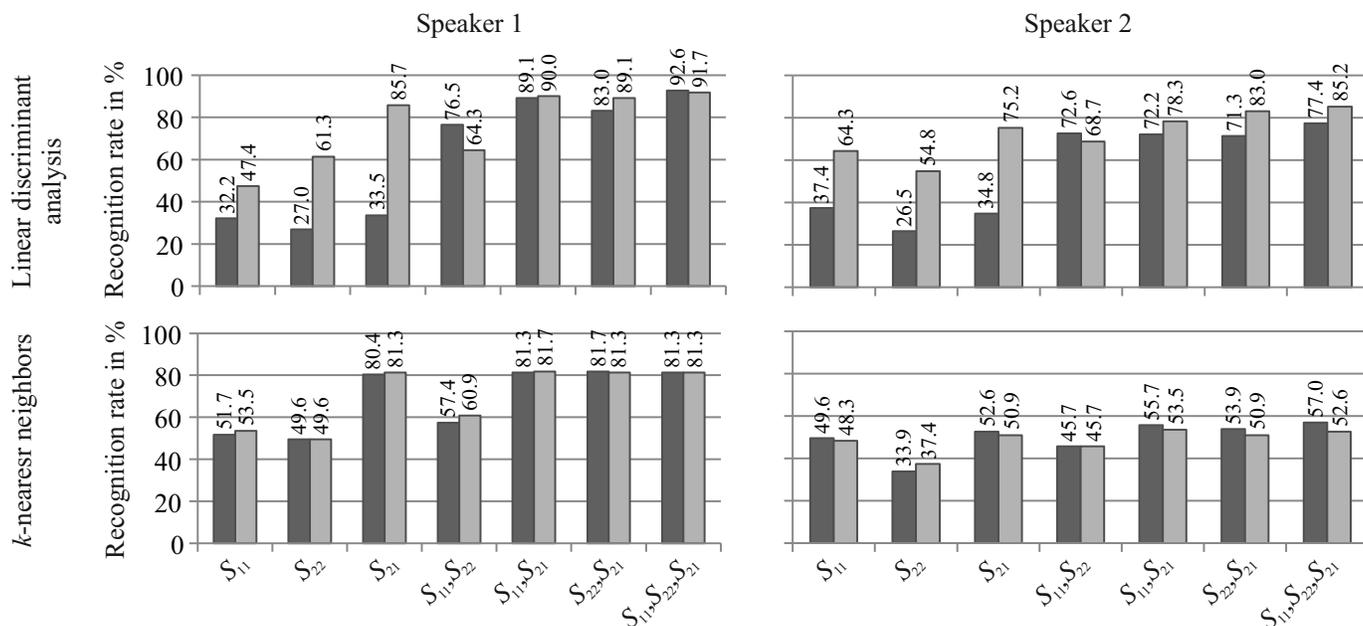


Fig. 7. Phoneme recognition rates for speaker 1 (left) and speaker 2 (right) using linear discriminant analysis (top) and *k*-nearest neighbors (bottom) as classifiers, and using different feature vectors. S_{11} , S_{22} , and S_{21} indicate, which spectral vectors have been concatenated to form the feature vectors: S_{11} is the spectrum of the reflection measurement with antenna 1, S_{22} is the spectrum of the reflection measurement with antenna 2, and S_{21} is the spectrum of the transmission measurement from antenna 1 to antenna 2. For the results shown with dark gray bars, magnitude spectra have been used. For the results shown with light gray bars, the real and imaginary parts of the spectra have been concatenated to form feature vectors of twice the length.

have a bigger size, which would be practical only to a limited extent.

With regard to the subjects, phoneme recognition rates were generally higher for speaker 1 than for speaker 2, independently from the classification method and the feature vectors. Since both speakers produced the speech items subjectively very clear, the most probable explanation is that speaker 2 had a metal dental implant as well as four amalgam (silver) fillings in his right molars, which are likely to impede the penetration of the microwaves through the vocal tract. Subject 1 did neither have dental implants nor metal tooth fillings. However, a definite clarification of this issue needs further investigations.

The recognition rates reported in Fig. 7 are unspecific with regard to the kind of classification errors that have been made. For a more specific picture, Table II shows the confusion matrix of the LDA for the data of speaker 2 using all three magnitude spectra as features. This matrix shows that wrongly recognized speech items were mostly classified as phonemes that are very similar from an articulatory point of view. For the tense vowels, major confusions occurred between /o:/ and /u:/, both of which are rounded back vowels. The lax vowels were generally confused more than the tense vowels, which is probably due to their more similar centralized articulation. Here, the major confusions occurred between /ε/ and /a/, which are also rather similar in articulatory terms. The consonants were hardly confused, which was rather surprising given their strong contextual variation. Notable is that there were no confusions across the boundaries of the three classes of tense vowels, lax vowels, and consonants.

The highest phoneme recognition rates achieved in the present study were 93 % for speaker 1 and 85 % for speaker 2.

To put these numbers in context, Hueber et al. [7] reported phoneme recognition rates of 59 % and 48 % for their two subjects with a silent-speech recognizer that used video images of the lips and ultrasound images of the tongue as sensory data. However, they used about 1 h of continuous speech and distinguished 40 phoneme classes instead of 25, which partly explains the lower performance. Using a silent-speech interface based upon surface electromyography, Jorgensen and Dusan [10] achieved recognition rates of only 18 % for 40 phoneme classes. This suggests that the achievable recognition rates depend to a fair amount on the modality and quality of the sensory data. Most other studies on silent-speech recognition did not report phoneme but word recognition rates for vocabularies of different sizes, which cannot be directly compared to our results (e.g., [3], [15], [4]). When the accuracy of acoustic-based phoneme recognition is taken as the baseline for silent-speech recognizers, representative reference values are 84 % for speaker-dependent phoneme recognition [6], and 77 % for speaker-independent phoneme recognition [39].

IV. CONCLUSIONS

In this study we explored to what extent microwave reflection and transmission measurements of the vocal tract allowed the recognition of speech sounds. With our setup using two antennas attached to the skin of the face, we achieved phoneme recognition rates of 93 % and 85 % for the two examined speakers. However, due to the limitations of the used network analyzer, the phonemes had to be artificially sustained, so that the data set contained only 230 items per speaker. Nevertheless, the corpus was designed in such a way that the speech samples reflected the articulatory variation of

TABLE II

CONFUSION MATRIX FOR SPEAKER 2 AND THE CLASSIFICATION BY LINEAR DISCRIMINANT ANALYSIS. THE FEATURE VECTORS WERE FORMED FROM THE CONCATENATION OF THE MAGNITUDE SPECTRA FROM THE ONE TRANSMISSION MEASUREMENT AND THE TWO REFLECTION MEASUREMENTS. THE DATA CONTAINED 10 SAMPLES FOR EACH OF THE VOWELS, AND 8 SAMPLES FOR EACH OF THE CONSONANTS. THE AVERAGE RECOGNITION RATE WAS 77.4 PERCENT.

		Recognized																										
		a:	e:	i:	o:	u:	ɛ:	ø:	y:	ɪ	ɛ	a	ɔ	ʊ	ʏ	æ	b	d	g	l	r	f	s	ʃ	m	n		
Intended	a:	8	2																									
	e:		9	1																								
	i:			1	9																							
	o:					6	4																					
	u:						3	7																				
	ɛ:	3							7																			
	ø:																											
	y:																											
	ɪ																											
	ɛ																											
	a																											
	ɔ																											
	ʊ																											
	ʏ																											
æ																												
b																												
d																												
g																												
l																												
r																												
f																												
s																												
ʃ																												
m																												
n																												

the phonemes found in continuous speech. The application of the proposed method to the recognition of truly continuous speech requires faster equipment for the data acquisition. The hardware for such real-time data acquisition and processing could be developed based on commercially available integrated circuits, and is the subject of future research. Using sweep durations of about 5 ms, which poses no special problem, it should then be possible to achieve a sampling rate for the articulatory states of about 100 Hz. This would allow to capture much more data and the application of more powerful classification methods like deep belief networks. The system can then also be extended for full-scale large-vocabulary speech recognition based on the large body of algorithms that have been developed for audio-based speech recognition. With regard to the data acquisition, there is further optimization potential in terms of the number, design, and placement of the antennas, and the frequency range of the signals.

REFERENCES

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

[2] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-based spoken communication: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[3] R. Hofe, S. R. Ell, M. J. Fagan, J. M. Gilbert, P. D. Green, R. K. Moore, and S. I. Rybchenko, "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing," *Speech Communication*, vol. 55, no. 1, pp. 22–32, 2013.

[4] J. Wang, A. Samal, J. R. Green, and F. Rudzicz, "Whole-word recognition from articulatory movements for silent speech interfaces," in *Proc. of the Interspeech 2012*, Portland, OR, USA, 2012, pp. 1326–1329.

[5] K. Thangthai and R. Harvey, "Improving computer lipreading via DNN sequence discriminative training techniques," 2017, pp. 3657–3661.

[6] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, and M. Stone, "Phone recognition from ultrasound and optical video sequences for a silent speech interface," in *Proc. of the Interspeech 2008*, Brisbane, Australia, 2008, pp. 2032–2035.

[7] 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 Communication*, vol. 52, no. 4, pp. 288–300, 2010.

[8] B. Denby, J. Cai, T. Hueber, P. Roussel, G. Dreyfus, L. Crevier-Buchman, C. Pillot-Loiseau, G. Chollet, S. Manitsaris, and M. Stone, "Towards a practical silent speech interface based on vocal tract imaging," in *Proc. of the 9th International Seminar on Speech Production (ISSP 2011)*, 2011, pp. 89–94.

[9] S. Srinivasan, B. Raj, and T. Ezzat, "Ultrasonic sensing for robust speech recognition," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010)*, 2010, pp. 5102–5105.

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

[11] J. Freitas, A. Teixeira, and M. S. Dias, "Towards a silent speech interface for Portuguese," *Proc. Biosignals*, pp. 91–100, 2012.

[12] M. Wand and T. Schultz, "Towards real-life application of EMG-based speech recognition by using unsupervised adaptation," in *Proc. of the*

- Interspeech 2014*, Singapore, 2014, pp. 1189–1193.
- [13] G. S. Meltzner, J. T. Heaton, Y. Deng, G. Luca, S. H. Roy, and J. C. Kline, “Silent speech recognition as an alternative communication device for persons with laryngectomy,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2386–2398.
- [14] M. J. Russell, D. Rubin, B. Wigdorowitz, and T. Marwala, “The artificial larynx: A review of current technology and a proposal for future development,” in *14th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics*. Springer, 2008, pp. 160–163.
- [15] S. Stone and P. Birkholz, “Silent-speech command word recognition using electro-optical stomatography,” in *Proc. of the Interspeech 2016*, San Francisco, USA, 2016, pp. 2350–2351.
- [16] —, “Angle correction in optopalatographic tongue distance measurements,” *IEEE Sensors Journal*, vol. 17, no. 2, pp. 459–468, 2017.
- [17] 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 *et al.*, “A wireless brain-machine interface for real-time speech synthesis,” *PLOS ONE*, vol. 4, no. 12, p. e8218, 2009.
- [18] C. Herff, D. Heger, A. de Pestors, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, “Brain-to-text: decoding spoken phrases from phone representations in the brain,” *Frontiers in Neuroscience*, vol. 9, p. 217, 2015. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fnins.2015.00217>
- [19] 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,” *Computer Speech & Language*, vol. 39, pp. 67–87, 2016.
- [20] J. M. Gilbert, J. A. Gonzalez, L. A. Cheah, S. R. Ell, P. Green, R. K. Moore, and E. Holdsworth, “Restoring speech following total removal of the larynx by a learned transformation from sensor data to acoustics,” *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. EL307–EL313, 2017.
- [21] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, “Direct speech reconstruction from articulatory sensor data by machine learning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2362–2374, 2017.
- [22] 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 Computational Biology*, vol. 12, no. 11, p. e1005119, 2016.
- [23] T. G. Csapó, T. Grósz, G. Gosztolya, L. Tóth, and A. Markó, “DNN-based ultrasound-to-speech conversion for a silent speech interface,” *Proc. of the Interspeech 2017*, pp. 3672–3676, 2017.
- [24] T. Le Cornu and B. Milner, “Generating intelligible audio speech from visual speech,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1751–1761, 2017.
- [25] A. R. Toth, K. Kalgaonkar, B. Raj, and T. Ezzat, “Synthesizing speech from Doppler signals,” in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2010)*, 2010, pp. 4638–4641.
- [26] M. Zahner, M. Janke, M. Wand, and T. Schultz, “Conversion from facial myoelectric signals to speech: A unit selection approach,” in *Proc. of the Interspeech 2014*, Singapore, 2014, pp. 1184–1188.
- [27] L. Diener, M. Janke, and T. Schultz, “Direct conversion from facial myoelectric signals to speech using deep neural networks,” in *International Joint Conference on Neural Networks*, 2015, pp. 1–7.
- [28] L. Diener, C. Herff, M. Janke, and T. Schultz, “An initial investigation into the real-time conversion of facial surface EMG signals to audible speech,” in *Proc. of the 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC 2016)*, 2016, pp. 888–891.
- [29] M. Wand, C. Schulte, M. Janke, and T. Schultz, “Compensation of recording position shifts for a myoelectric silent speech recognizer,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, 2014, pp. 2094–2098.
- [30] J. Holzrichter, G. Burnett, L. Ng, and W. Lea, “Speech articulator measurements using low power EM-wave sensors,” *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 622–625, 1998.
- [31] J. F. Holzrichter, “Characterizing silent and pseudo-silent speech using radar-like sensors,” in *Proc. of the Interspeech 2009*, Brighton, UK, 2009, pp. 628–631.
- [32] A. M. Eid and J. W. Wallace, “Ultrawideband speech sensing,” *IEEE Antennas and Wireless Propagation Letters*, vol. 8, pp. 1414–1417, 2009.
- [33] Y. H. Shin and J. Seo, “Towards contactless silent speech recognition based on detection of active and visible articulators using IR-UWB radar,” *Sensors*, vol. 16, no. 11, p. 1812, 2016.
- [34] X. Fang, M. Ramzan, Q. Wang, and D. Plettemeier, “Compact antipodal Vivaldi antennas for body area communication,” in *Proc. of the 12th International Conference on Body Area Networks (BODYNETS 2017)*, Dalian, China, 2017.
- [35] P. Birkholz, “Modeling consonant-vowel coarticulation for articulatory speech synthesis,” *PLoS ONE*, vol. 8, no. 4, p. e60603, 2013.
- [36] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
- [37] L. v. d. Maaten, “Matlab implementation of t-SNE [software],” 2010. [Online]. Available: <https://lvdmaaten.github.io/tsne/#implementations>
- [38] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is “nearest neighbor” meaningful?” in *International Conference on Database Theory*. Springer, 1999, pp. 217–235.
- [39] A.-r. Mohamed, G. Dahl, and G. Hinton, “Deep belief networks for phone recognition,” in *Proc. of the NIPS Workshop Deep Learning for Speech Recognition and Related Applications*, vol. 1, no. 9, Vancouver, Canada, 2009, pp. 39–47.

Peter Birkholz received the diploma in computer science and the Ph.D. degree (with distinction) in signal processing from the Institute for Computer Science at the University of Rostock, Germany, in 2002 and 2005, respectively. He worked as a research associate at the University of Rostock from 2005-2009, and at the Department of Phoniatrics, Pedaudiology, and Communication Disorders of the RWTH Aachen University, Germany, from 2009-2014. Since 2014 he is Junior Professor for Cognitive Systems at the TU Dresden, Germany. His main topics of research include articulatory speech synthesis, computational neuroscience, and measurement techniques for speech research. Dr. Birkholz was awarded the Joachim-Jungius Prize in 2006 by the University of Rostock for his dissertation on articulatory speech synthesis, and the Klaus-Tschira Award for Achievements in Public Understanding of Science in 2006.

Simon Stone (né Preuß) received the diploma degree from the Faculty of Electrical Engineering and Information Technology of the RWTH Aachen University, Aachen, Germany in 2012. He has worked as a research fellow at the University Hospital RWTH Aachen from 2012 to 2014. Since 2014 he is working as a research fellow and pursuing his Ph.D. degree with the Institute of Acoustics of Speech Communication at the TU Dresden, Germany. His research interests include signal processing, instrumental articulatory measurements, and their application to silent-speech interfaces.

Klaus Wolf received the diploma from the Faculty of Electrical Engineering of the TU Dresden in 1988. Since then, he has been working as a laboratory engineer at the chair of Radio Frequency and Photonics Engineering at the TU Dresden.

Dirk Plettemeier received the diploma in Communications Engineering from the University of Applied Sciences, Lemgo, Germany, and the diploma in Electrical Engineering from the Ruhr-University Bochum, Germany. In 2002, he received the Ph.D. degree from the Ruhr-University Bochum and was with the Institute for High-Frequency Technique at this university until 2003. From 2003 to 2007, he was the head of the research group Numerical Computation of High-Frequency Electromagnetic Fields and Waves at the Electrotechnical Institute, Chair and Laboratory for Theory of Electromagnetic Fields and Electromagnetic Compatibility (EMC) at the TU Dresden, Germany. From 2007 to 2011, he was head of the research group High Frequency Engineering - Antennas and Wave Propagation at the TU Dresden. Since 2011, he is Professor for Radio Frequency Engineering at the TU Dresden.