Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Sister Conferences Best Papers Track

Speech Recognition Using RFID Tattoos (Extended Abstract)

Jingxian Wang* , **Chengfeng Pan** , **Haojian Jin** , **Vaibhav Singh** , **Yash Jain** , **Jason I. Hong** , **Carmel Majidi** and **Swarun Kumar**
Carnegie Mellon University

Abstract

This paper presents a radio-frequency (RF) based assistive technology for voice impairments (i.e., dysphonia), which occurs in an estimated 1% of the global population. We specifically focus on acquired voice disorders where users continue to be able to make facial and lip gestures associated with speech. Despite the rich literature on assistive technologies in this space, there remains a gap for a solution that neither requires external infrastructure in the environment, battery-powered sensors on skin or body-worn manual input devices. We present RFTattoo, which to our knowledge is the first wireless speech recognition system for voice impairments using batteryless and flexible RFID tattoos. We design specialized wafer-thin tattoos attached around the user's face and easily hidden by makeup. We build models that process signal variations from these tattoos to a portable RFID reader to recognize various facial gestures corresponding to distinct classes of sounds. We then develop natural language processing models that infer meaningful words and sentences based on the observed series of gestures. A detailed user study with 10 users reveals 86% accuracy in reconstructing the top-100 words in the English language, even without the users making any sounds.

1 Introduction

This paper seeks to develop an RF-based assistive technology for persons with voice impairments. In the US, more than 2 million people require digital Adaptive Alternative Communication (AAC) methods to help compensate for speech impairments . While various classes of voice impairments exist, we target acquired conditions where users continue to be able to make facial and lip gestures associated with speech. We aim to learn these gestures over time to produce speech in real-time. Our approach applies to a wide range of temporary and permanent acquired dysphonia (voice disorders) ranging from hoarseness to complete loss of voice that occurs in about 1% of the global population.

While there is rich literature on assistive input-to-speech technologies for speech impairments, state-of-the-art solutions suffer important limitations. Camera-based [Hassanat, 2014; Delmas *et al.*, 2002; Agrawal *et al.*, 2016] visual solutions for real-time lip-reading require users to constantly be within line-of-sight of a camera, which may not be possible when the user is on the move. Audio-based assistive solutions [Muda *et al.*, 2010] only apply to speech impairments where users are able to produce sounds and struggle in noisy environments. Past work has proposed a variety of face-worn sensors for speech sensing, particularly in clinical settings, such as magnets attached to the tongue [Bedri *et al.*, 2015], EEG helmets [Suppes *et al.*, 1997] and EMG electrodes on the face [Janke and Diener, 2017]. Assistive text-to-speech innovations require users to provide constant manual input to the system via keypads or various user interfaces that require training and practice for proficiency. There remains a gap for an everyday intuitive assistive technology for voice impairments that does not require external infrastructure, bulky sensors on the face or manual hand input.

We present RFTattoo, the first wireless speech recognition platform for voice impairments through skin-friendly, wafer-thin, battery-free and stretchable RFID tattoos. We fabricate specialized light RFID tattoos attached to the skin surface of face at known locations. Each tag is fabricated to be stretchable, flexible, wafer-thin, extremely light and made with hypoallergenic materials. The tags are designed to be hidden under makeup and extremely skin-friendly. We track the strain of individual tags over time as they deform in response to motions generated by different intended sounds. However, it is often the case that certain distinct sounds produce similar facial movements. To this end, we build natural language processing models that combine identified facial gestures in context to construct meaningful words and sentences. A detailed user study with 10 users reveals 86% accuracy in recognizing the top-100 words in the English language.

RFTattoo's first challenge is to process signals from RFID tattoos to recognize distinct facial and lip gestures called *visemes*[1][Fisher, 1968], that correspond to sounds the user

*Contact Author, jingxian@cmu.edu; The title of the original paper is "RFID Tattoo: A Wireless Platform for Speech Recognition" which appears in ACM UbiComp 2020.

[1]A viseme is a set of phonemes that look the same, for example, when lip reading.

intends to express. RFTattoo recognizes visemes by modeling the pure stretch of the flexible tag antenna. An intuitive approach to model tag stretch to infer its impact on the frequency at which it resonates. Specifically even a small change, say one millimeter, in the electrical length of an antenna lowers its resonant frequency by as much as 8 MHz in our experiments. Unfortunately, RFID tags in the U.S. operate in the FCC's unlicensed 900 MHz band with an effective bandwidth of 26 MHz. This makes it challenging to accurately capture the large frequency shifts induced by stretch. More importantly, requiring an RFID reader to hop through all frequencies even within the unlicensed band would be too time-consuming ($\sim$ few seconds) to recognize real-time speech.

RFTattoo addresses this challenge by probing multiple specially tuned RFID tags instead of probing multiple frequencies at the reader. In particular, we design an RFID tag that advertises the bits of its own current stretch value even if it is probed at one frequency (e.g. 915 MHz). Our approach to do so attaches multiple RFID chips to a common antenna, each tuned to multiple sets of specially chosen frequencies.

A second challenge RFTattoo must address is the dynamic radio environment – changing orientation of the RFID tags, multipath reflections as well as movement of the user's body. RFTattoo achieves this through a novel tag antenna design that isolates the impact of stretch from other aspects pertaining to the radio environment. Specifically we fabricate two co-located RFID antennas with two materials – one stretchable and one non-stretchable. We then compare the signals received across both RFID tags to isolate any effect from the tag location, orientation and radio environment.

Finally, RFTattoo builds a natural language processing framework to map stretch values of tattoos placed at different points in the face to recognize words and sentences the user intends to speak. A key challenge in this regard is the fact that some sounds produce identical facial and lip gestures (visemes) and therefore cause a high degree of ambiguity in the recognized phonemes. RFTattoo addresses this through two approaches. First, RFTattoo monitors subtle movements of the user's tongue through its impact on the magnitude and phase of the RFID tags on the skin's surface. We show how this allows for disambiguation of certain phonemes that produce identical facial movement. Second, RFTattoo leverages a useful property commonly exploited in natural language processing – the fact that adjacent phonemes are not completely independent but must follow the English dictionary and rules of grammar. Sec. 5 describes our approach to recognize common words and sentences at high accuracy, based on these observations.

Limitations: We emphasize a few important limitations of RFTattoo: (1) RFTattoo achieves highest accuracy when the location of RFID tags on the face are known through a light-weight calibration a priori. This means that for optimal performance, one must re-calibrate should RFTattoo tags be peeled off and on, or with natural wear. (2) RFTattoo may miss visemes should specific tags be unresponsive owing to shadowing from the body relative to the reader. (3) RFTattoo's accuracy is poor in the face of unknown or untrained words (e.g. less common words and proper nouns). This is a common problem shared by voice recognition systems [Johnson, 2019] (e.g. Siri, Alexa, etc.) as well as visual lip reading systems [Hassanat, 2014; Agrawal *et al.*, 2016; Assael *et al.*, 2016].

We implement RFTattoo by building custom tag antennas using stretchable Ag-PDMS conductors on PDMS substrates connected to three RFID chips. We use a meander-line antenna appropriately impedance tuned to respond at the 900 MHz ISM band. We use commodity Impinj RFID readers attached to the user's waist. Our system is attached to the user's face using hypoallergenic stickers and covered with makeup. We conduct a detailed user study with 10 users including two users with temporary dysphonia (loss of voice). We also include results when all users are instructed to mouth words silently. Our results reveal that:

- RFTattoo achieves a median accuracy in stretch of 1.4 mm.
- RFTattoo distinguishes between eleven visemes of the English language at an accuracy of 90%.
- RFTattoo recognizes the most frequently used 100 words of the English language at an accuracy of 86%.

Contributions: Our main contribution is a novel system that recognizes intended speech of users with voice impairments using light-weight RFID tattoos attached to the face. Our contributions include:

- Algorithms that recognize subtle mm-accurate stretches of the tattoos as well as movement of the tongue by processing RF-backscatter signals at a handheld reader.
- A natural language processing framework that recognizes various facial gestures associated with speech to construct meaningful words and sentences.
- A detailed user study that reveals the promise of our approach in recognizing intended speech, even when users do not make any sounds.

2 Primer on RFID Tags

RFID tags are widely used in our daily life, for example, ID cards, contactless key fobs, baggage trackers in airports, and clothing tags in warehouses and markets. RFID tags are batteryless; they rely on a nearby wireless energy source (RFID reader) to operate and send information back to the reader. The communication range of passive RFID tags is limited – at most 5 to 10 meters. Recent work develops a long-range RFID system that extends the range by 8 times [Wang *et al.*, 2019b]. While commercial RFID tags are usually used for asset tracking and person identification, many new applications are developed in the wireless research community. Prior work has shown that commercial RFID tags can be used for body skeleton tracking [Jin *et al.*, 2018a], shape sensing [Jin *et al.*, 2018b], everyday object localization [Wang *et al.*, 2021], etc. In this paper, we design and build a new type of passive RFID tag that is stretchable and skin-friendly.

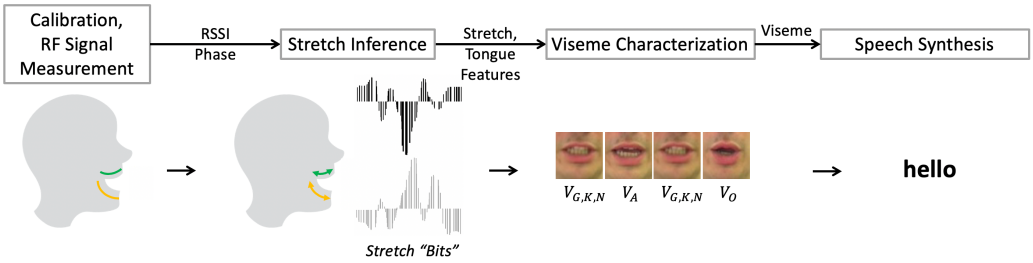

Figure 1: RFTattoo's Architecture: (1) measures the wireless channel of RFID tattoos; (2) infers stretch bits and tongue position based on the reflected signal power and phase of multiple three-chip RFID tattoos; (3) feeds the features into machine learning models to recognize corresponding facial gestures (viseme); images show the corresponding viseme from the GRID dataset [Cooke *et al.*, 2006]; (4) combines visemes to form meaningful words and sentences by natural language processing.

3 Overview of RFTattoo

(1) Inferring RFID stretch and tongue position: RFTattoo actively measures the stretch of each RFID tag and the position of the tongue – two key aspects that help recognize speech. RFTattoo infers stretch by monitoring its effect on impedance due to the fact that a stretched tag is longer and thinner. RFTattoo specifically measures the frequency response, i.e. the change in magnitude of the reflected signal across frequencies to study this effect. We measure this property accurately and in real-time despite the limited bandwidth of commercial reader. We further show how RFTattoo can also infer the position of the tongue through its effect on RFID impedance.

(2) Processing speech: Given the stretch of individual RFID tags, RFTattoo fuses these measurements to infer visemes, that are visual gestures of the face produced by different syllables pronounced by the user. We note that some visemes can be produced by multiple sounds, (e.g. "thee" and "tea" are indistinguishable visually). We show how we can disambiguate many such sounds using the position of the tongue. Sec. 5 describes our system that borrows from natural processing techniques to fuse the resulting phoneme measurements into meaningful words and sentences.

4 Processing RFTattoo Signals

4.1 Inferring Tag Stretch

Our key approach to monitor tag stretch measures the change in impedance as a result of the tattoo elongating. Specifically, as tattoos are stretched, its effecting width decreases and length increases, both of which increases its resistance and reactance. In effect, this causes a change in the resonant frequency of the RFID tag.

Why does resonant frequency shift with stretch?: The resonant frequency of an antenna is the frequency, where the amplitude is higher than at adjacent frequencies. Stretching an RFID tag changes its antenna's electrical length and therefore its resonant frequency. Specifically, as the antenna length increases, the wavelength at which it resonates also increases meaning that the resonant frequency will shift towards lower frequencies. Mathematically, the resonant frequency of a half-wave dipole antenna is written as [Bhartia *et al.*, 1991]:

$$f_{res} = \frac{c}{2\sqrt{\varepsilon_e}} \frac{1}{L + L_e} \tag{1}$$

where L is the effective length of the half-wave dipole antenna and the L_e is the effective elongation of the antenna as stretching, ε_e is the effective relative permittivity of the antenna substrate and c is the speed of light in free-space. The ε_e can be estimated using the method mentioned from [Jackson and Alexopoulos, 1986]. From this equation, we can see the resonant frequency is inversely proportional to its electric length leading to a very simple approach to infer stretch once electric length is found accurately.

Using the principle above, RFTattoo infers the stretch by monitoring the frequency responses of the tags. We would like to refer our full paper that describes the approach in greater details. RFTattoo further measures the tag locations and the tongue positions using the phase and amplitude of the tag signals [Wang *et al.*, 2019a].

5 From RF Signals to Speech

RFTattoo synthesizes speech by processing the stretch, location and tongue position of various points in the skin obtained from the RFID tag signals. It first uses this information to classify between various facial gestures called visemes that are unique to different sounds. RFTattoo then borrows from the rich literature on text-to-speech in natural language processing to synthesize speech in real-time.

5.1 Characterizing Visemes

A *viseme* is a unit of visual speech – more specifically, the visual equivalent of a phoneme (a unit of sound in speech recognition). Each viseme represents the shape of the face when the user attempts to speak a particular phoneme [Fisher, 1968]. Past work on automated lip-reading have widely used visemes to recognize speech based on video input[Bear, 2017]. Recognizing shorter visemes as opposed to longer speech segments has several advantages such as needing less training effort and generalizing well for different speaker identities (speaking styles, accents, etc.).

Choice of visemes: Phonemes map many-to-one to visemes, because many phonemes can not be distinguished using only visual cues (e.g. "p" vs. "m" sounds). Phoneme-to-viseme

mappings have been constructed mainly by two approaches: linguistic and data-driven. In this paper, we use the map from [Lee and Yook, 2002] obtained through a hybrid linguistic and data driven approach – a relatively sparse set which worked well experimentally (see Sec. 6). This map is composed by 38 phonemes and 11 classes (plus a silence class).

Viseme classification: To classify the different viseme sets, we use the resonant frequency shift property of the RFID tattoo tag. RFID tattoos are attached to 4 different locations on the persons face: above the upper lip, below the lower lip, the left cheek and the right cheek. As the person utters the different phoneme sounds, these tags are stretched by different amounts resulting in diverse set of resonant frequencies.

5.2 Speech Synthesis

Now we aim to synthesize the speech that the user intends to speak. At each unit of time, we predict a list of phoneme candidates, derived from the viseme mapping, as well as their likelihood scores. The likelihood scores are obtained by our machine learning model, which outputs the predicted viseme with corresponding probabilities for all possible visemes. Using these phoneme candidates, we can reconstruct words with ambiguities. We note that despite the 90% accuracy in viseme classification and even upon accounting for tongue position, the ambiguity of the phonemes could significantly impact speech reconstruction accuracy.

To address this, RFTattoo draws from a salient advantage of natural language processing – adjacent phonemes and words are not independent – they are limited by the English dictionary and rules of grammar. We leverage this fact to disambiguate the recognition results and recognized the transcript of what the user intends to speak. Finally, we synthesize the speech using a public text-to-speech API[2].

During the operation, our recognition algorithm will produce a prediction stream that contains the recognition result and the corresponding time windows. Each recognition result r_k consists a list of phoneme candidates $r_k c_1, r_k c_2, ..$ and associated likelihoods $r_k l_1, r_k l_2, ...$

Word & sentence segmentation: To perform speech recognition, we first organize the recognized phonemes into words and sentences, based on their recorded time stamps. Here, a word is comprised of phonemes, and words compose a sentence. The lengths of pauses between in-word syllables, words, and sentences often vary. We run a pilot study with four participants and empirically determine the pause thresholds for both word-level separation and sentence level separation. If the pauses between multiple adjacent phonemes are smaller than the word/sentence threshold, we group these phonemes into the same word/sentence.

On-the-fly word disambiguation: Next, for each set of phonemes constituting a word, we derive a set of most likely word candidates using a pronouncing dictionary [Group, 2019]. A pronouncing dictionary defines the mapping between sequences of phonemes and words. We then need to select one word from each group to assemble the final sentence. Choosing the words randomly, or even the most likely word per phoneme sequence often results in gibberish, since the words may not form meaningful sentences in combination. Leveraging this fact, we build a Bayesian model to evaluate the naturalness of the sentence formed by different word sequences. Let $N(w_k|w_1, w_2, ..w_{k-1})$ denote the naturalness score of choosing the word w_n among a group G_k, given that the prior sequence $w_1, w_2, ..w_{k-1}$ is determined. The selection of the incoming word W_k^* is equivalent to finding word that can maximize:

$$w_k^* = \arg\max_{w_k} N(w_k|w_1, w_2, ..w_{k-1}) * l(w_k) \quad (2)$$

where $l(w_k)$ is the word likelihood score from the earlier viseme recognition.

We use a co-occurrence relation to measure the naturalness. We count the frequencies of m consecutive words appear in a large document collection, and use these frequencies to indicate the naturalness. In other words, the more common the word sequence is, the more natural the sequence would be. We measure the co-occurrence in a sliding window of m consecutive words.

$$N(w_k|w_1, w_2, ..w_{k-1}) = \prod_1^m p(w_k, w_{k-1}, .., w_{k-m+1}) \quad (3)$$

where $p(w_k, w_{k-1}, .., w_{k-m+1})$ is the non-zero frequency of these consecutive words in a large document collection. If we cannot find a specific consecutive word sequence in the corpus, We set $p = 1e^{-3}$ to avoid multiplication by zero. Our implementation sets $m = 3$ and measures the frequency in Cornell Movie Dialog Corpus [Danescu-Niculescu-Mizil and Lee, 2011]. While our approach is simple and easy to reproduce, a more specific and contextual corpus, such as including sentences used most commonly in daily conversation can improve performance.

6 Results

We evaluate the performance of our speech recognition system at three different levels: viseme, word and sentence [Wang *et al.*, 2019a]. We present the accuracy in sentence construction below:

Method: We conducted a pilot experiment for the sentence construction in a regular office space.

Result: We first observe that for the 20 commonly used sentences in day-to-day use known to our system, RFTattoo shows an average of 91% accuracy. In contrast, the raw recognition accuracy for sentences unseen by RFTattoo is 35.7%. Integrating natural language processing based correction further boosts the average accuracy to 53.2%. We note that this is within the performance range for unknown sentences of state-of-the-art vision-based lip-reading software that requires line-of-sight (e.g. 46.8% in [Chung *et al.*, 2017]). We also find that RFTattoo works better for longer sentences, which contains more contextual information. Our results reveal that RFTattoo holds promise in reconstructing sentences for users with voice impairments. Our accuracy can be further improved over time with more data to tune to the user's particular speaking habits.

[2]https://cloud.google.com/text-to-speech/

References

[Agrawal *et al.*, 2016] Shreya Agrawal, Verma Rahul Omprakash, et al. Lip reading techniques: A survey. In *Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on*, pages 753–757. IEEE, 2016.

[Assael *et al.*, 2016] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: Sentence-level lipreading. *CoRR*, abs/1611.01599, 2016.

[Bear, 2017] Helen L. Bear. Decoding visemes: improving machine lipreading (phd thesis). *CoRR*, abs/1710.01288, 2017.

[Bedri *et al.*, 2015] Abdelkareem Bedri, Himanshu Sahni, Pavleen Thukral, Thad Starner, David Byrd, Peter Presti, Gabriel Reyes, Maysam Ghovanloo, and Zehua Guo. Toward silent-speech control of consumer wearables. *Computer*, 48(10):54–62, 2015.

[Bhartia *et al.*, 1991] P. Bhartia, K. V. S. Rao, and R. S. Tomar. *Millimeter-wave microstrip and printed circuit antennas / P. Bhartia, K.V.S. Rao, R.S. Tomar*. Artech House Boston, 1991.

[Chung *et al.*, 2017] Joon Son Chung, Andrew W Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, pages 3444–3453, 2017.

[Cooke *et al.*, 2006] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[Danescu-Niculescu-Mizil and Lee, 2011] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics, 2011.

[Delmas *et al.*, 2002] P. Delmas, N. Eveno, and M. Lievin. Towards robust lip tracking. In *Object recognition supported by user interaction for service robots*, volume 2, pages 528–531 vol.2, 2002.

[Fisher, 1968] Cletus G Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11(4):796–804, 1968.

[Group, 2019] CMU Speech Group. The cmu pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2019. (Accessed on 01/17/2019).

[Hassanat, 2014] Ahmad Basheer Hassanat. Visual words for automatic lip-reading. *arXiv preprint arXiv:1409.6689*, 2014.

[Jackson and Alexopoulos, 1986] Davidr Jackson and Nicolaosg Alexopoulos. Analysis of planar strip geometries in a substrate-superstrate configuration. *IEEE transactions on antennas and propagation*, 34(12):1430–1438, 1986.

[Janke and Diener, 2017] Matthias Janke and Lorenz Diener. Emg-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 25(12):2375–2385, 2017.

[Jin *et al.*, 2018a] Haojian Jin, Jingxian Wang, Zhijian Yang, Swarun Kumar, and Jason Hong. Rf-wear: Towards wearable everyday skeleton tracking using passive rfids. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 369–372. ACM, 2018.

[Jin *et al.*, 2018b] Haojian Jin, Jingxian Wang, Zhijian Yang, Swarun Kumar, and Jason Hong. Wish: Towards a wireless shape-aware world using passive rfids. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pages 428–441, 2018.

[Johnson, 2019] Bernadette Johnson. How siri works. https://electronics.howstuffworks.com/gadgets/high-tech-gadgets/siri4.html, 2019. Accessed: 2019-11-01.

[Lee and Yook, 2002] Soonkyu Lee and DongSuk Yook. Audio-to-visual conversion using hidden markov models. In *Pacific Rim International Conference on Artificial Intelligence*, pages 563–570. Springer, 2002.

[Muda *et al.*, 2010] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

[Suppes *et al.*, 1997] Patrick Suppes, Zhong-Lin Lu, and Bing Han. Brain wave recognition of words. *Proceedings of the National Academy of Sciences*, 94(26):14965–14969, 1997.

[Wang *et al.*, 2019a] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I Hong, Carmel Majidi, and Swarun Kumar. Rfid tattoo: A wireless platform for speech recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(4):1–24, 2019.

[Wang *et al.*, 2019b] Jingxian Wang, Junbo Zhang, Rajarshi Saha, Haojian Jin, and Swarun Kumar. Pushing the range limits of commercial passive rfids. In *16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19)*, pages 301–316, Boston, MA, 2019. USENIX Association.

[Wang *et al.*, 2021] Jingxian Wang, Junbo Zhang, Ke Li, Chengfeng Pan, Carmel Majidi, and Swarun Kumar. Locating everyday objects using nfc textiles. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*, pages 15–30, 2021.