Hand gesture recognition using thin plate radiation and gated-recurrent-unit, based on ultrasound doppler

Paul Glemain, Emmanuel Hardy, Charles Hudin, Pierre-Henri Orefice, Nazih Mechbal

Hand Gesture Recognition Using Thin Plate Radiation and Gated-Recurrent-Unit, Based on Ultrasound Doppler

Paul Glémain*†, Emmanuel Hardy‡, Charles Hudin*, Pierre-Henri Orefice* and Nazih Mechbal†
*Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
†PIMM, Arts et Métiers Institute of Technology, CNRS, CNAM, HESAM University, Paris, France
‡Univ. Grenoble Alpes, CEA, Leti, F-38000 Grenoble, France
Email : {paul.glemain,emmanuel.hardy,charles.hudin,pierre-henri.orefice}@cea.fr nazih.mechbal@ensam.eu

***Abstract*—In recent years, touchless technologies for human-computer interaction have been widely developed. Doppler sonar makes it possible to extract information from hand gestures by emitting/receiving ultrasounds, and gestures recognition is generally achieved using features extracted from a gesture sequence as input to Convolutional Neural Network. This work aims at achieving an accurate and rich acoustical touchless gesture recognition with a low number of transducers and a low complexity real-time classifier. For this purpose, we use a thin plate as an acoustic antenna, excited by a few piezoelectric actuators, and capture the echoes with microphones around the plate. High amplitude emissions on a large bandwidth are achievable with a better-integrated system. Signal features selected to contain meaningful information on rich 3D gestures are computed and used as an input to a small Gated-Recurrent-Unit neural network. We achieve the detection and classification of 11 3D gestures with an accuracy of 93.5% with our system.**

***Index Terms*—Ultrasound, Doppler, hand gesture recognition, machine learning, GRU**

I. Introduction

Over the past years, technologies for contactless human-computer interaction (HCI) have been developed, particularly for gesture recognition (GR). Commands can be recognized by a system through the movement of a user. The solutions developed for this task rely on various technologies, including but not limited to the use of cameras, radar sensors, or acoustic sensors. Vision-based solutions primarily rely on analyzing images provided by cameras (e.g., IR, depth sensors) such as Kinect or Leap Motion but may raise privacy concerns and are sensitive to changes in the surrounding lighting conditions. Radar and acoustic solutions, in addition to overcoming these issues, can offer wavelengths that enable centimeter-level resolution, suitable for detecting hand gestures.

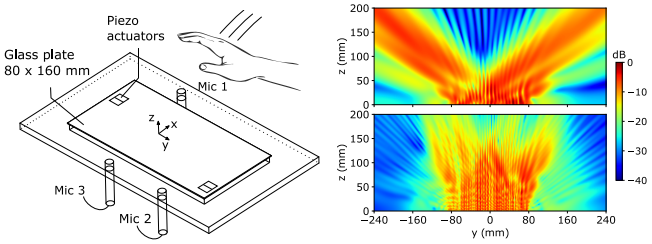
The developed radar or ultrasound solutions employ diverse emission and analysis techniques. Some rely on the study of time of arrival (ToA) and time of flight (ToF) [1], [2]. Other methods utilize Doppler and micro-Doppler analysis, with a continuous wave (CW), frequency-modulated continuous wave (FMCW) or pulse emission with low ultrasound speakers or transducers with a narrow bandwidth for ultrasound solutions [3], [4]. After features extraction, gesture classification is performed in various ways. Many systems employed classification using CNN [5]–[7] while some recent solutions presented the use of Recurrent Neural Networks (RNN) and their derivatives as Long Short-Term Memory (LSTM) or Gated-Recurrent-Unit (GRU) given their effectiveness in handling temporal data. These networks have been used for activity recognition [8] or GR tasks [9], [10]. However, gestures sequences are often processed into a bi-directional network which can prevent low latency streaming operation.

In this article, we present a novel integrated acoustic emitter design for GR task. By utilizing a plate excited by a small number of piezoelectric actuators, we can achieve multi-frequency emission in a large bandwidth. Furthermore, the directivity properties of this system provide gesture localization information. We also propose a set of features extracted not only from isolated microphone signals but also from relative signals. Finally, we show that our system combined with a low-complexity unidirectional GRU network trained using a max-pooling loss function, can both detect and classify gestures in streaming. We perform a 93.5% accuracy in recognizing 11 3D gestures, showing the robustness of the system and making real-time HCI applications feasible.

II. System Design

The setup consists of a thin glass plate equipped with four piezoceramic actuators (PI Ceramic, PIC 151) glued on it as shown in Fig. 1a. The plate is designed to be simply supported on the edges and baffled by an acrylic structure. The signals generated and amplified result in a voltage range of ±30 V applied to the actuators. Additionally, three microphones (PCB Piezotronics, 378C01) are positioned around the plate to capture the reflected echoes. Signals are generated and measured between the system and the computer using an National Instruments NI6363 DAQ card.

Utilizing a plate as an emitter offers several advantages compared to using traditional transducers. Firstly, integrating the emitter into devices such as smartphones is easier, with transducers located behind the screen. Additionally, the plate allows for CW signal emission at different frequencies and high amplitudes. By exploiting radiation directivity of the plate, which depends on the excitation frequency as depicted in



(a) Acoustic setup overview.

(b) Normalized pressure level radiation above the plate for a 40 kHz excitation (top) and 100 kHz excitation (bottom), at x=0.

Fig. 1. System design.

Fig. 1b, specific frequency ranges can convey distinct information about gestures performed above specific parts of the plate. Lower frequencies are more responsive to gestures executed on the sides, while higher frequencies capture gestures on the top surface. Furthermore, the use of multi-frequency emission helps to mitigate the impact of frequency selective fading, which can result in information loss when relying solely on a single frequency emission. Lastly, as this system enables high-frequency emissions, we can measure higher Doppler shifts, enhancing the detection of slower-speed gestures. To ensure effective GR within a 25 cm space surrounding the plate, we have selected three distinct frequencies for CW acoustic emission: 40 kHz, 60 kHz, and 100 kHz.

III. Gesture Recognition Method

A. Data Pre-processing

Before extracting features from the various microphone signals, we initiate their preprocessing as depicted in Fig. 2.

We begin by applying bandpass filtering to each received signal, taking into account the different carrier frequencies, and considering the Doppler shift associated with a maximum hand velocity of 5 m/s. Next, we calculate the baseband complex signal for each carrier frequency and microphone. These signals, which are downsampled to 10 kHz, will be called single microphone channels. We also take an additional step by multiplying the complex signal of one microphone with the conjugate of another microphone for each existing pair. The resulting signals, called relative microphone channels, provide instantaneous relative Doppler shift information between the two microphones through the phase component in the case of a single moving reflector.

Finally, we compute the Fast Fourier Transform (FFT) on 40 ms windows. This operation is performed every 20 ms,

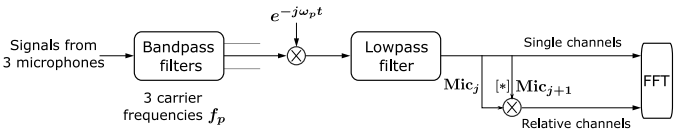


Fig. 2. Pre-processing pipeline with bandpass filtering, baseband conversion and frequency spectrum computation.

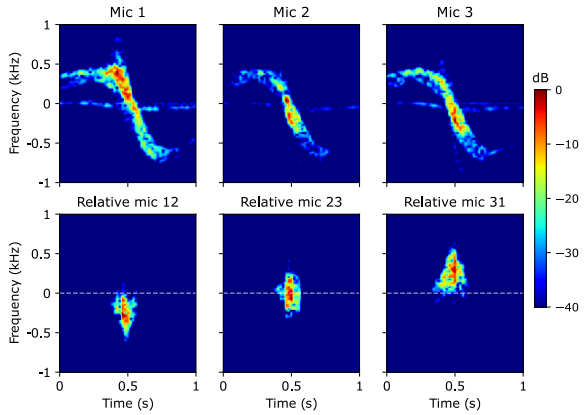


Fig. 3. Spectrograms for a left-to-right swipe at 100 kHz of single microphone channels (first line) and relative microphone channels (second line).

defining the system frame rate, to obtain a time-frequency representation of the signals. The resulting power spectrums, denoted as $S(t, f)$ with t and f representing the time and frequency variables, are depicted in Fig. 3 for a left-to-right swipe at 100 kHz. The relative channels at the bottom provide spatial information about the hand movement between the receivers.

B. Features Extraction

In order to maintain a manageable input vector size for the network without compromising GR performance, we extract informative features related to the gestures, as presented in Table I. While many existing systems perform this extraction on full time-frequency representations of gesture sequences, we have chosen to perform a frame-by-frame features extraction to enable real-time detection and classification tasks.

The spectral flux Fl and spectral entropy H provide valuable information regarding the presence of movement and reflectors above the plate, respectively. These features are summed over different channels and carrier frequencies to reduce overall dimensionality. The spectral centroid provides information about the Doppler shift for each frame, and the relative Doppler shift between two microphones for relative channels. Additionally, the spectral envelopes represent the Doppler frequency signature of each gesture. Extracting these envelopes has demonstrated promising results in various gestures recognition task [11]. They offer supplementary information to other features, particularly for finer gestures that may exhibit multiple Doppler velocities simultaneously, where the centroid alone may not provide sufficient relevant information. The lower and upper envelopes are computed by considering the minimum and maximum frequencies at which the spectrogram energy in the Doppler bandwidth surpasses a threshold empirically chosen. To fully leverage the multiplicity of receivers, the last two features are computed directly on both single microphone channels and relative channels. This approach provides more meaningful information than simply applying basic operations to features extracted from single microphone channels.

TABLE I
FEATURES EXTRACTED FROM THE FREQUENCY SPECTRUMS.

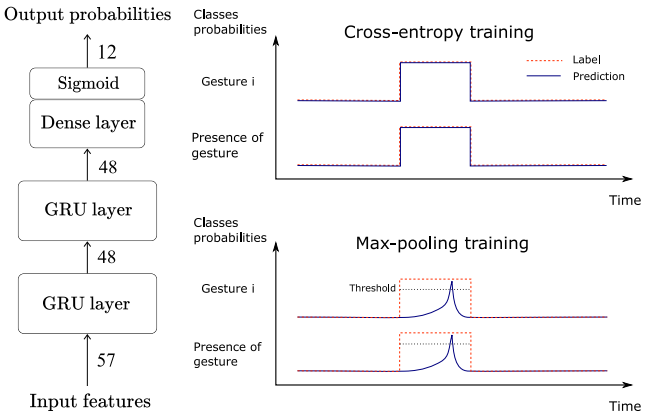
Features	Description
Spectral flux	$Fl(t) = \sqrt{\sum_f \lvert S(t,f) - S(t-1,f)\rvert^2}$
Spectral entropy	$H(t) = -\frac{\sum_f \frac{S(t,f)}{\sum_f S(t,f)} log_2 \frac{S(t,f)}{\sum_f S(t,f)}}{log_2 K}$
Spectral centroid	$c(t) = \frac{\sum_f fS(t,f)}{\sum_f S(t,f)}$
Spectral envelope	$f_{low}(t),\ f_{up}(t)$

C. Neural Network Design

The network architecture, depicted in Fig. 4a, shows the use of GRU layers which are a variant of LSTM. These types of layers were specifically designed to address the issues of exploding or vanishing gradients that could occur when using traditional RNN. Both LSTM and GRU cells incorporate gating mechanisms including a forget gate to effectively process time series data.

For our application, we have decided to employ two unidirectional GRU layers. This choice was motivated by the fact that GRU cells are less complex than LSTM cells and are well-suited for training with small datasets. Additionally, this network architecture only considers information from the past, suitable for streaming application, as opposed to bidirectional layers that incorporate information from both past and future. Each GRU layer consists of 48 units, and the final layer is followed by a dense layer of 12 units with a sigmoid activation function.

The first eleven classes correspond to the gestures to be classified, while the last class indicates the presence of a gesture. This allows the network to detect when a gesture is performed, regardless of its specific class.



(a) Neural Network architecture.

(b) Training steps of the network.

Fig. 4. Neural network design for GR.

The network undergoes two training stages, following the approach described for keyword spotting in [12] and illustrated in Fig. 4b. In the pre-training stage, binary cross-entropy loss is employed to train the system for frame-by-frame classification of gestures and non-gestures. Subsequently, the pre-trained model is further trained using max-pooling loss, transitioning to a segment-level perspective. The system learns to generate spikes when it becomes confident about the presence of a gesture and its corresponding classification. In order to make a decision, we examine frames where the probability of gesture presence surpasses a threshold. We then assess whether this threshold is also exceeded for each specific gesture class.

IV. DATA COLLECTION

The 11 gestures studied are presented in Fig. 5, with four swipes (a) right to left (Srl), (b) left to right (Slr), (c) top to bottom (Stb), (d) bottom to top (Sbt), (e) spread fingers (SF), (f) pinch fingers (PF), (g) finger rotation clockwise (FRc), (h) counter-clockwise rotation (FRcc), (i) push-pull moving toward then away from the plate (PPull), (j) pull-push moving away then toward (PPush) and (k) a double-tap mouvement (Dt). These 3D gestures are widely used and have been selected due to their direct applicability to HCI tasks involving digital systems [13]. Moreover, these gestures include both finer and larger movements, presenting a challenge for the system.

We collected data from a total of 19 participants (8 females and 11 males). Participants were instructed to perform randomly selected gestures, based on their corresponding names and pictures. They were given a single instruction regarding the desired speed of the gestures, ranging from slow to fast. In order to increase the variability in the dataset, no specific guidelines were provided regarding hand placement, approach and removal phases, or the height relative to the system

The collected dataset composed of 1400 gestures was divided into a training and a test set. The training set encompassed gestures performed by 14 participants, whereas the test set comprised gestures performed by the remaining 5 participants. In addition, we included recordings where no gestures

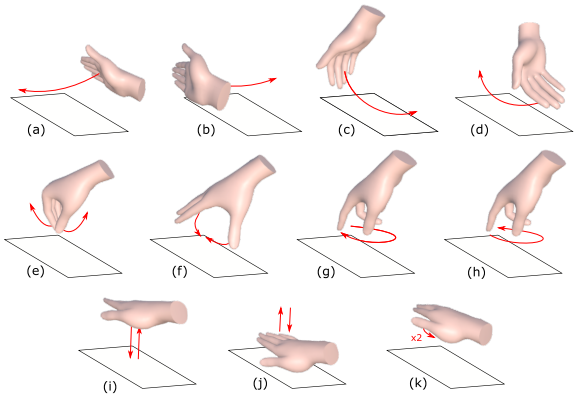


Fig. 5. Illustration of the eleven 3D gestures.

were performed in the training set. The labeling of the data was based on video recordings of the gestures. This labeling approach ensured that the hand's approach and removal phases were excluded from the labeled gestures, thereby preventing the network from learning from these specific points.

V. Results and Discussion

The system accurately detects and classifies 93.5% of gestures performed. For a 5%, gestures are not recognized as learned gestures or are identified as multiple gestures simultaneously (i.e. the probabilities are under the threshold or multiple are above). These cases correspond to the out-of-domain (OOD) category in confusion matrix in Fig. 6. Furthermore, in 1.5% of cases, the system accurately detects the presence of a gesture, but it may confuse it with another gesture. This confusion typically occurs with similar gestures such as swipes.

These results can be compared to systems presented in Table II. As with the works presented, our system retains good accuracy for participants who have not contributed to the training set, despite the large in-class variability with minimal instruction and speed variation. This proves the robustness of our system for detecting a set of diversified 3D gestures.

While confirming the benefits of multi-frequency emission in our system with an average recognition rate difference of approximately 4% between using a single carrier and three carriers in transmission, we found that using more carriers would not enhance accuracy as the frequencies already fully cover the spatial area above the plate.

VI. Conclusion

The system presented allows for efficient detection and classification of 11 3D gestures that can be used for HCI tasks. In addition to relying on a better-integrated transmitter in digital systems, our system exhibits good robustness in handling the significant variability in execution among users. Furthermore, the detection remains highly accurate even for gestures performed by users not included in training set. The proposed set of features and the utilization of a low-complexity GRU network trained with a max-pooling loss appear to be suitable for the task of gesture detection and classification in real-time with a CW Doppler sonar. Further studies can be conducted to explore the improved integration of receivers into our system, as well as the classification performance for a broader range of diverse gestures. Some work can also be done to improve features extraction and make calculation less complex for a low-consumption system design.

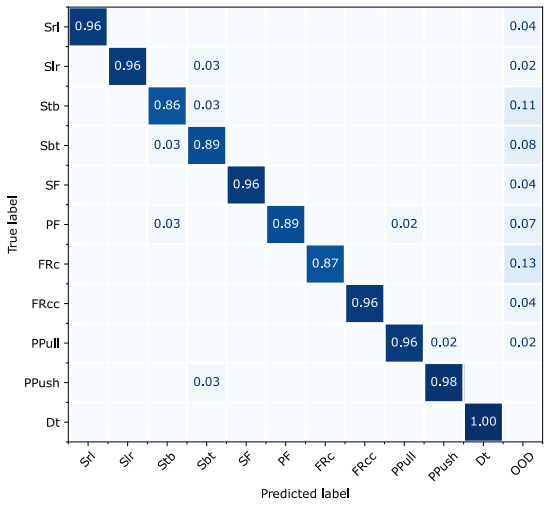


Fig. 6. Confusion matrix for the GR task with an overall accuracy of 93.5%.

TABLE II
Comparison of Our Work with Other Doppler GR Systems.

	Our system	[5]	[10]	[6]
Type of emission	Ultrasound CW	Ultrasound Pulse	Ultrasound CW	Radar FMCW
Method	Doppler	Doppler	Doppler	Doppler
# Receivers	# 3	# 1	# 3	# 2 to 4
Classifier	GRU	CNN	LSTM	CNN
Performance (# Gestures)	93.5% (# 11)	96.5% (# 5)	95.65% (# 6)	83.3% to 98.8% (# 6)
Range of detection	5 to 25 cm	N/A	0 to 1 m	25 cm

References

[1] A. Das, I. Tashev, and S. Mohammed, "Ultrasound based gesture recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 406–410.

[2] E. Hardy *et al.*, "Spike-based beamforming using pMUT arrays for ultra-low power gesture recognition," in *2022 International Electron Devices Meeting (IEDM)*, 2022, pp. 24.4.1–24.4.4.

[3] S. Gupta, D. Morris, S. Patel, and D. Tan, "Soundwave: using the Doppler effect to sense gestures," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin Texas USA, 2012, pp. 1911–1914.

[4] K. Kalgaonkar and B. Raj, "One-handed gesture recognition using ultrasonic Doppler sonar," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1889–1892.

[5] Q. Zeng, Z. Kuang, S. Wu, and J. Yang, "A method of ultrasonic finger gesture recognition based on the micro-Doppler effect," *Applied Sciences*, vol. 9, no. 1111, p. 2314, Jan 2019.

[6] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, "Dynamic hand gesture classification based on multistatic radar micro-Doppler signatures using convolutional neural network," in *2019 IEEE Radar Conference (RadarConf)*, 2019, pp. 1–5.

[7] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-Doppler signatures with convolutional neural network," *IEEE Access*, vol. 4, pp. 7125–7130, 2016.

[8] A. Shrestha, H. Li, J. Le Kernec, and F. Fioranelli, "Continuous human activity classification from FMCW radar with bi-LSTM networks," *IEEE Sensors Journal*, vol. 20, no. 22, pp. 13 607–13 619, 2020.

[9] J.-W. Choi, S.-J. Ryu, and J.-H. Kim, "Short-range radar based real-time hand gesture recognition using LSTM encoder," *IEEE Access*, vol. 7, pp. 33 610–33 618, 2019.

[10] C.-S. Lin, M. Y. Abdul Gaffar, J. Son, and S. Winberg, "Dynamic hand gesture recognition using Doppler sonar and deep learning," in *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 2021, pp. 1–6.

[11] M. G. Amin, Z. Zeng, and T. Shan, "Hand gesture recognition based on radar micro-Doppler signature envelopes," in *2019 IEEE Radar Conference (RadarConf)*, 2019, pp. 1–6.

[12] M. Sun *et al.*, "Max-pooling loss training of long short-term memory networks for small-footprint keyword spotting," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 474–480.

[13] Z. Wang *et al.*, "Hand gesture recognition based on active ultrasonic sensing of smartphone: A survey," *IEEE Access*, vol. 7, pp. 111 897–111 922, 2019.