NASR: NonAuditory Speech Recognition with Motion Sensors in Head-Mounted Displays

Jiaxi Gu[1(✉)], Kele Shen[2], Jiliang Wang[2], and Zhiwen Yu[1]

[1] School of Computer Science and Engineering,
Northwestern Polytechnical University, Xi'an 710072, P.R. China
gujiaxi@mail.nwpu.edu.cn
[2] School of Software, Tsinghua University, Beijing 100084, P.R. China

Abstract. With the growing popularity of Virtual Reality (VR), people spend more and more time wearing Head-Mounted Display (HMD) for an immersive experience. HMD is physically attached on wearer's head so that head motion can be tracked. We find it can also detect subtle movement of facial muscles which is strongly related to speech according to the mechanism of phonation. Inspired by this observation, we propose NonAuditory Speech Recognition (NASR). It uses motion sensor for recognizing spoken words. Different from most prior work of speech recognition using microphone to capture audio signal for analysis, NASR is resistant to acoustic noise of surroundings because of its nonauditory mechanism. Without using microphone, it consumes less power and requires no special permissions in most operating systems. Besides, NASR can be seamlessly integrated into existing speech recognition systems. Through extensive experiments, NASR can get up to 90.97% precision with 82.98% recall rate for speech recognition.

Keywords: Head-Mounted Display, Motion Sensor, Speech Recognition, Machine Learning

1 Introduction

With the growing popularity of Virtual Reality (VR), more and more people put their heads in various Head-Mounted Display (HMD) for an immersive experience, especially for watching 360-degree videos and playing VR games. An HMD is attached on one's head to track his/her head movement to provide dynamic visual content or interaction. We find that the subtle head movement caused by phonation can also be recorded and then used for speech recognition. Inspired by this observation, we propose NonAuditory Speech Recognition (NASR).

In comparison with traditional audio-based methods, NASR is low-cost and loose-constraint. It is based on the anatomic structure motion caused by phonation. When one speaks, his/her facial muscles moves to shape the sound and air stream into recognizable speech [1]. As Figure 1 shows, the movement of facial muscles can be captured by the sensors of HMD. Different from microphones, motion sensors are always available without consuming much power.

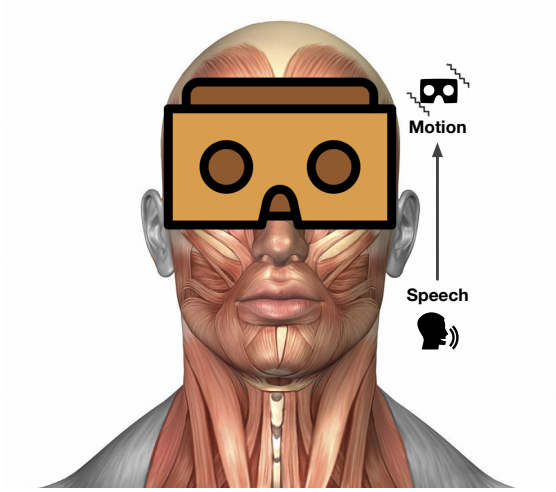

Fig. 1. Human speech is produced with the movement of relevant facial muscles. Through the motion sensors in HMD, speech can be recognized.

Besides, NASR is inherently resistant to acoustic noise of surroundings. It can be seamlessly integrated into existing speech recognition models. To the best of our knowledge, NASR is the first work to merely use motion signal to do speech recognition.

It is however nontrivial to implement NASR because of the following challenges. Firstly, as speech is produced mainly by the muscles near lips, the motion of muscles near HMD is relatively weak. Secondly, different persons have more or less diverse accents. It makes the motion patterns of facial muscles hard to recognize. Last but not least, the speech tone and speed make difference even for the same individual. Therefore, multiple variables have to be considered to make NASR robust and accurate for speech recognition.

In summary, we have the following contributions in this paper:

1. We propose a creative speech recognition method for HMD by merely using motion sensors without touching any audio signal.
2. We present a multidimensional motion data structure and extract relevant features for machine learning.
3. We implement NASR in the real world using an HMD equipped with a smart phone and evaluate it with extensive experiments.

2 Methodology

The application scenario of NASR is speech recognition with motion sensors in HMDs. The source data is motion signal recorded during the speaking of an HMD wearer. The objective is to recognize the speech content. For this objective, we use a machine learning method to build a classifier. By extracting the

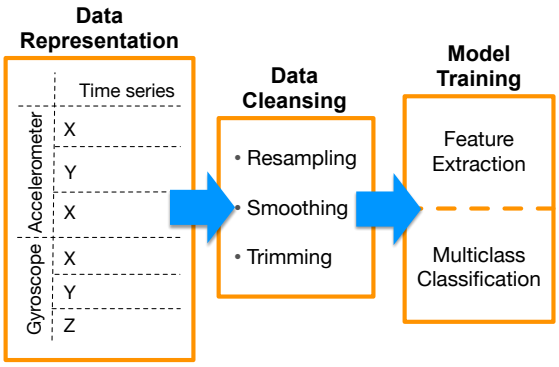

Fig. 2. The whole process of NASR is divided into three main phases.

relevant features from the motion signal, we can build a classifier for recognize the incoming utterance. As Figure 2 shows, the whole process of NASR can be divided into three main phases: *data representation*, *data cleansing* and *model training*.

2.1 Data Representation

The accelerometer and gyroscope are two typical motion sensors in consumer HMDs. Either sensor has 3 axes so we have total 6 time series. We denote it as:

$$\mathbf{S} = \begin{array}{c} \\ acce_x \\ acce_y \\ acce_z \\ gyro_x \\ gyro_y \\ gyro_z \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & ... & t & ... & n \end{array} \\ \begin{pmatrix} S_{1,1} & S_{1,2} & ... & S_{1,t} & ... & S_{1,n} \\ S_{2,1} & S_{2,2} & ... & S_{2,t} & ... & S_{2,n} \\ S_{3,1} & S_{3,2} & ... & S_{3,t} & ... & S_{3,n} \\ S_{4,1} & S_{4,2} & ... & S_{4,t} & ... & S_{4,n} \\ S_{5,1} & S_{5,2} & ... & S_{5,t} & ... & S_{5,n} \\ S_{6,1} & S_{6,2} & ... & S_{6,t} & ... & S_{6,n} \end{pmatrix} \end{array} \in \mathbb{R}^{6\times n}$$

The rows of $\mathbf{S}$ represent different axes of motion sensors. The columns of $\mathbf{S}$ represent samples of sensor readings on the timeline. The n is determined by the sampling rate and duration. For simplicity, we limit a separate word utterance into a fixed time span, i.e., 2 seconds. Thus, the sampling rate is a critical factor. As different types of motion sensors or sensors made by different manufacturers may have different sampling rate, we use a linear interpolation to do resampling to make all the six sensor data have the same number of samples in the fixed sampling time. As a result, it is reasonable to represent the motion pattern with a matrix $\mathbf{S}$.

2.2 Data Cleansing

First, for making the motion signal uniform and reasonable, we resample the data from axes of different sensors in a specific sampling rate. Second, for removing

the interference from physiological activities such as heartbeat and respiration, we use a low-pass filter to remove the high-frequency component for smoothing. Last but not least, the time span of an utterance is uncertain. To generate valid training data for our model, we need to trim each sample to contain exactly one spoken word. Ideally, the time span of a sample needs to start and stop at the edges of the spoken word. We use a threshold of motion amplitude and remove the period of time in which any motion is barely detected.

2.3 Model Training

Feature engineering on time series is a classic but still challenging process. Possible relevant features include basic features of the time series such as the mean value, the standard deviation, the number of peaks or more complex features such as the time reversal symmetry statistics. After feature extraction, we use a subset of samples as a training set. Since the target words are predefined, it is a typical multiclass classification. There are various classification algorithms for this purpose such as Decision Tree and Random Forest. We will evaluate the performance of several different classification algorithms.

3 Implementation and Evaluation

The HMD kit of our experiments consists of two components including a smart phone and a plastic VR headset. The smart phone we use is Samsung Galaxy S7 running Android 8.0 and the VR headset is a plastic version of Google Cardboard which costs about $10. We develop a VR application using Google VR SDK to get motion sensor data in real time. For model training, we generate speech samples from five adults including three males and two females. Our predefined word set contains three common English words: “private”, “secret” and “password”. For each word, we collect 280 samples in total from different people. We also collect unknown-word samples while the wearer is irrationally speaking for 4 minutes. Therefore, we have 4 labels in total.

For feature extraction, we use a development tool called TSFRESH which stands for *Time Series Feature extraction based on scalable hypothesis tests*. It can automatically extract features from time series. In our experiments, 500 relevant features are used. Basically, we adopt Random Forest for multiclass classification and other popular machine learning methods are evaluated too.

In the preliminary experiments, we use a constrained version of samples. Each sample lasts exact 2 seconds and contains at most one spoken word. We use 50% of samples as training instances and the other 50% as testing instances. We start with the subsets containing samples generated by every single person. The model training and testing are both based on the subsets. The recognition accuracy of different words is shown in Table 1. Then, we consider the performance of NASR between different persons. The results are shown in Table 2.

In the near-pilot experiments, we use the whole set of samples. Similarly, the 50% of samples are used for training and the rest are for testing. The results of

Table 1. The performance of NASR on different words.

Word	*private*	*secret*	*password*	unknown-word
Precision	81.25%	96.43%	96.55%	94.83%
Recall	83.87%	71.05%	84.85%	96.49%
F1-score	82.54%	81.82%	90.32%	95.65%

Table 2. The performance of NASR on different persons.

Person	#1	#2	#3	#4	#5
Precision	80.00%	80.00%	88.24%	100%	96.30%
Recall	100%	87.50%	83.33%	88.24%	92.86%
F1-score	88.89%	83.58%	85.71%	93.75%	94.55%

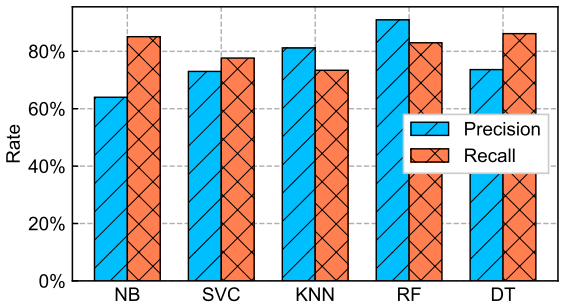

Fig. 3. The performance of NASR using different machine methods. (NB: Naive Bayes; SVC: Support Vector Classification; KNN: K-Nearest Neighbors; RF: Random Forest; DT: Decision Tree.)

different machine learning methods are shown in Figure 3. The Random Forest gets the best performance and its precision is 90.97% and recall rate is 82.98%.

4 Related Work

With the development of wearable devices, motion signal can be used for various applications. Some works use accelerometers or gyroscopes on smart phones for medical purposes such as monitoring heart rate or respiration [6]. Different from using motion signal directly, Guha Balakrishnan et al. use subtle head motion in videos to extract heart rate and beat lengths [2]. Their work points out that the Newtonian reaction to the influx of blood at each heart beat cause recognizable head motion. Javier Hernandez et al. design a series applications for measuring physiological signs using head-mounted glasses [3], smart phone [4] and smart watch [5]. Besides, Reham Mohamed et al. show that the gyroscope sensor is

the most sensitive sensor for measuring the heart rate [7]. In addition to motion sensors, there are also works using wireless signal such as WiFi [8] and RFID [9] for detecting human behaviors. These works may bring more inspirations to NASR in the future.

5 Conclusion

In this paper, we propose NASR which is a nonauditory speech recognition system. Merely relying on the consequent motion during the usage of HMD for VR, the wearer's spoken words can be recognized without touching microphones. NASR shows advantages from various aspects including low battery consumption and resistance to surrounding noise. It can also help to make several potential applications such as voice command assistant like Siri and accessibility-related products. In the future work, we will test NASR on a larger scale and extend it to real-time for continuous speech recognition.

References

1. Abbs, J.H., Gracco, V.L., Blair, C.: Functional muscle partitioning during voluntary movement: Facial muscle activity for speech. Experimental Neurology 85(3), 469–479 (1984)
2. Balakrishnan, G., Durand, F., Guttag, J.: Detecting pulse from head motions in video. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3430–3437 (2013)
3. Hernandez, J., Li, Y., Rehg, J.M., Picard, R.W.: Bioglass: Physiological parameter estimation using a head-mounted wearable device. In: International Conference on Wireless Mobile Communication and Healthcare - Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH). pp. 55–58 (2014)
4. Hernandez, J., McDuff, D.J., Picard, R.W.: Biophone: Physiology monitoring from peripheral smartphone motions. In: International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 7180–7183 (2015)
5. Hernandez, J., McDuff, D., Picard, R.W.: Biowatch: estimation of heart and breathing rates from wrist motions. In: Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare. pp. 169–176 (2015)
6. Kwon, S., Lee, J., Chung, G.S., Park, K.S.: Validation of heart rate extraction through an iphone accelerometer. In: International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5260–5263 (2011)
7. Mohamed, R., Youssef, M.: Heartsense: Ubiquitous accurate multi-modal fusion-based heart rate estimation using smartphones. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1(3), 97 (2017)
8. Wu, C., Yang, Z., Zhou, Z., Liu, X., Liu, Y., Cao, J.: Non-invasive detection of moving and stationary human with wifi. IEEE Journal on Selected Areas in Communications 33(11), 2329–2342 (2015)
9. Zhou, Z., Shangguan, L., Zheng, X., Yang, L., Liu, Y.: Design and implementation of an rfid-based customer shopping behavior mining system. IEEE/ACM Transactions on Networking 25(4), 2405–2418 (2017)