Journal of Neural Engineering

Inductive conformal prediction for silent speech recognition

To cite this article before publication: Ming Zhang *et al* 2020 *J. Neural Eng.* in press https://doi.org/10.1088/1741-2552/ab7ba0

Inductive conformal prediction for silent speech recognition

Ming Zhang[1], You Wang[1], Wei Zhang[1], Meng Yang[2], Zhiyuan Luo[3], Guang Li[1]

[1] State Key Laboratory of Industrial Control Technology, Institute of Cyber Systems and Control, Zhejiang University, Hangzhou 310027, Zhejiang, China
[2] Department of Computer Science and Technology, School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing, 100083, China
[3] Department of Computer Science, Royal Holloway, University of London, Egham Hill, Egham, Surrey TW20 0EX, UK

E-mail: king_wy@zju.edu.cn

August 2019

Abstract. *Objective.* Silent speech recognition based on surface electromyography has been studied for years. Though some progress in feature selection and classification has been achieved, one major problem remains: how to provide confident or reliable prediction. *Approach.* Inductive conformal prediction (ICP) is a suitable and effective method to tackle this problem. This paper applies ICP with the underlying algorithm of random forest to provide confidence and reliability. We also propose a method, test time data augmentation, to use ICP as a way to utilize unlabelled data in order to improve prediction performance. *Main Results.* Using ICP, p-values and confidence regions for individual predictions are obtained with a guaranteed error rate. Test time data augmentation also outputs relatively better conformal predictions as more unlabelled training data accumulated. Additionally, the validity and efficiency of ICP under different significance levels are demonstrated and evaluated on the silent speech recognition dataset obtained by our own device. *Significance.* These results show the viability and effectiveness of ICP in silent speech recognition. Moreover, ICP has potential to be a powerful method for confidence predictions to ensure reliability, both in data augmentation and online prediction.

Keywords: silent speech recognition, inductive conformal prediction, test time data augmentation, guaranteed error rate

Submitted to: *J. Neural Eng.*

1. Introduction

Silent speech recognition (SSR) is an emerging and active research area of brain-computer interface (BCI). It aims to enable recognition of speech when an audible acoustic signal is unavailable, and is generally based on either electromyography (EMG) or electroencephalography (EEG) (Kapur et al., 2018, Anumanchipalli et al., 2019). Surface EMG (sEMG) seems to be more practical in recent years because it requires only a very small number of channels, is noninvasive, and has been properly decoded by muscles (Denby et al., 2010, Jorgensen and Dusan, 2010, Wand et al., 2014, Ji et al., 2018). sEMG involves the recording of the surface muscle activity, which is the results of the spatial and temporal integration of motor unit action potential (MUAP) stemming from neural signals. Since speech is pronounced by the activity of the human articulatory muscles, sEMG allows us to detect this activity and determine the corresponding speech without the need for sound (Katirji, 2007, Kass and Mizrahi, 2010, Jorgensen and Dusan, 2010, Preston and Shapiro, 2012, Hakonen et al., 2015).

Some researchers have made good progress in both isolated words and short sentences (polysyllable) classification during past years, which further demonstrates the potential of silent speech recognition via sEMG (Chan et al., 2001, Manabe et al., 2003, Hueber et al., 2010, Schultz et al., 2017). For example, Jorgensen et al. used a set of six words as discrete commands in a simulated planetary mission, achieving an error rate lower than 10% (Jorgensen et al., 2003). Another research work mentioned the error rate of 32% in continuous recognition of 108 words (Jou et al., 2006). Early in 2018, it was reported the promising applications of SSR which achieved a relatively good result in silent speech recognition (Kapur et al., 2018).

Silent speech recognition has many applications; for example it can be used to build assisting devices for speech-disabled persons, as well as for secure and noise-free communications (Lopez-Larraz et al., 2010, Schultz and Wand, 2010, Fraiwan et al., 2011). However, there are a few important cases that demand very high confidence in risk-sensitive SSR applications such as silent speech in pilots, military operation or controlling peripheral devices. Incorrect recognitions may result in disastrous consequences. Therefore, it is important to be able to measure the risk of misclassification and if possible, to ensure low risk of error. Furthermore in many cases, high quality data without labels need to be selected or examined to augment a training set in order to improve the model and then the predictions. Such situations require a reliable confidence measure of data.

Conformal prediction (CP), first proposed in 2005 (Vovk et al., 2005), can output region predictions (a subset of predicted labels) with guaranteed error rate under the assumption of independent and identically distribution (*i.i.d.*) (Devetyarov and Nouretdinov, 2010). Inductive conformal prediction (ICP) is a variant of CP as it can reduce computational complexity compared to other methods such as transductive conformal prediction (Vovk, 2015). ICP can be implemented on top of any existing machine learning algorithms as the underlying algorithm for calculating nonconformity score, including support vector machines, decision trees, boosting, and neural networks (Papadopoulos et al., 2007, Löfström et al., 2013, Lei et al., 2015, Sun et al., 2017). The main benefit of using ICP method is to have more efficient and reliable classification while at the same time maintaining prediction accuracy for a given significance level (Devetyarov and Nouretdinov, 2010, Nguyen and Luo, 2015, Wang et al., 2019).

In this paper, inductive conformal prediction based on a random forest classifier is applied to silent speech recognition. We also propose a method, test time data augmentation, to use ICP as a way to utilize unlabelled data in order to improve prediction performance. All SSR data is collected by our own device with a good transducer performance from surface electromyography to digital signals. Experiments are carried out to evaluate performance of ICP and demonstrate the confidence and credibility of new unlabelled examples can be used to augment the training set. Our experimental results show the superiority of ICP and its improvement achieved using data augmentation. ICP can provide valid confident predictions which is lacking in current silent speech recognition applications.

2. Methods

2.1. Data collection

Silent speech data are collected directly from the surface of several specific articulatory muscles of the face and neck (Kenneth, 2010, Schultz and Wand,

2010, Goodier, 2017, Kapur et al., 2018), as shown in Figure 1. It actually contains six channels of the surface electromyography: the zygomaticus major (channels 1,2), the levator anguli oris (channels 1,2), the platysma (channel 3), the extrinsic tongue (channels 4,5), the digastric anterior belly (channel 4) and the lateral pterygoid (channel 6). Channels 1 and 2 are bipolar derivation, whereas channels 3, 4, 5 and 6 are derived unipolarity, with two reference electrodes placed on the mastoid behind ears. Two bipolar derivations are used to help improve common-mode rejection ratio (CMRR) at the front end of some particular muscles and then enhance signal-to-noise ratio (SNR). Note that no voice is produced during data collection, only subtle mouth movement allowed to help the participant concentrate on the experiment.

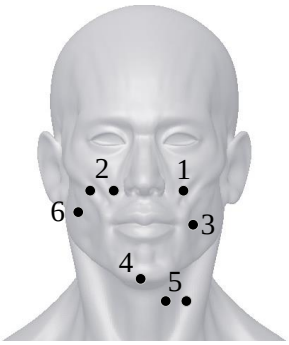


Figure 1. Electrode sites on the face and neck. These specific positions actually form a articulator muscular net to decode the silent speech.

All data are captured by Ag/AgCl electrodes of a device named 'MiCap' (developed in our lab at Zhejiang University). The device features with 24 bits ADC as well as high sensibility at uV level. The amplitude of clean neuromuscular signals always varies between $10uV \sim 5mV$, the summit may sometimes even smaller, so MiCaP can satisfy this measurement. Note that in the frequency domain, the signals mainly consist of low-frequency components with basically less than $300Hz$, therefore the sampling rate is accordingly set to be $1000Hz$. MiCap's power consumption is lower than $30mW$ with all functions on and it transfers data via Bluetooth which further improves its usability. Wireless communication minimises the chances of introducing dynamic interferences into the device. One typical example of the original silent speech signals collected by MiCap is presented in Figure 2. It shows the variations when collecting silent speech data (purely silent). A raw signal always contains many interference components that display high peak-to-peak values and severe drifts, representing the most common scenario.

Young Chinese adult students are invited to take part in our experiments. There are 3 participants that involved in data collection, two males and one female, with an average age of 22 years. The research that associated with human subjects has been approved by Ethics of Human and Animal Protection Committee of Zhejiang University. Participants volunteered to take part in the experiment as a research participant and have signed the informed consent form. All data are only used for analysis and the privacy of the participants is protected.

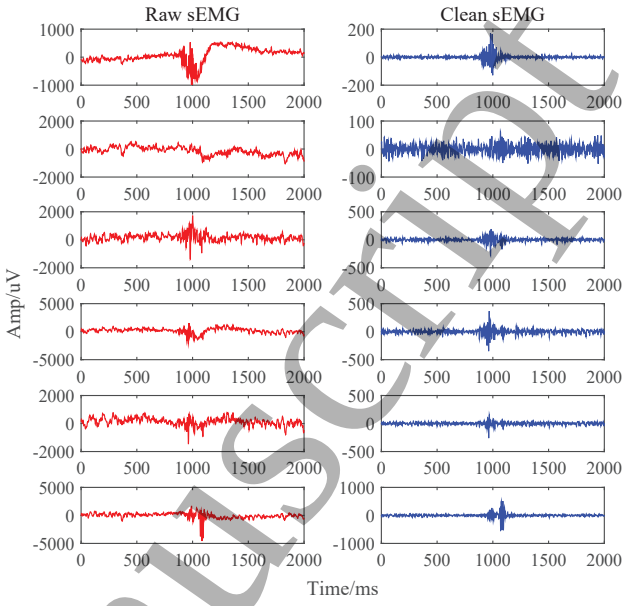


Figure 2. sEMG preprocessing. The left column illustrates an example of raw sEMG signals with 6 channels, and the absolute amplitudes span from $100uV$ to $800uV$. The right column shows the corresponding clean sEMG after preprocessing which suppressed the noise from skin surface and surroundings.

There are 10 surface electrodes in total put on each participant's face and neck for up to two hours. The positions illustrated in Figure 1 are required to be kept as clean as possible. The impedance between each electrode and the according skin surface should not be larger than $5k\Omega$, and Nuprep gel (Weaver and Company, USA) is used at the beginning of experiment. During the data collection, each participant is asked to sit on a chair and stay still, with a short break every 20 minutes. The participant concentrates on a computer screen where isolated commands will be displayed one by one in a defined sequence in order to help us label recorded signals. An LED indicator is used to remind the participant to start silent speech. At the same time, sEMG data are collected by MiCap and saved on the computer.

Ten specific silent speech commands in Chinese are collected for around 40 hours per participant. The commands are '噪', '1', '2', '前', '后', '左', '右', '快', '慢', '停' respectively, which mean 'null', 'one',

Table 1. Data set

Label	'0'	'1'	'2'	'3'	'4'	'5'	'6'	'7'	'8'	'9'
Count	1682	1597	1662	1623	1634	1588	1591	1679	1554	1673

'two', 'forward', 'backward', 'left', 'right', 'accelerate', 'decelerate', 'stop' in English. Note that they are indexed with labels from '0' to '9' for recognitions. The data set has 16283 valid samples in total and its distribution listed in Table 1.

2.2. Preprocessing and feature extraction

Data are sampled at the rate of $1000Hz$ with a 24 bits ADC, as mentioned above. Appropriate preprocessing methods are applied on the data to reduce interferences as much as possible.

In this paper, a pass-band filter of $0.3{\sim}300Hz$ is applied on every sEMG sample as it retains the main information while removing DC and most high frequency components. After that, a comb notch filter of $50Hz$ is used to remove electric power interference, and the baseline artifacts are also removed by Quadratic Variation Reduction (QVR):

$$\boldsymbol{z} = [\boldsymbol{I} - (\boldsymbol{I} + \lambda \boldsymbol{D}^T \boldsymbol{D})^{-1}]\widetilde{\boldsymbol{z}} \tag{1}$$

where $\widetilde{\boldsymbol{z}}$ and $\boldsymbol{z}$ denote the signal before and after using QVR, λ is a constant ($\lambda = 100$), $\boldsymbol{I}$ represents identity matrix and $\boldsymbol{D}$ is a $(n-1) \times n$ matrix:

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \tag{2}$$

where n is the length of $\widetilde{\boldsymbol{z}}$ (Fasano and Villani, 2014). In Equation 1, $(\boldsymbol{I} + \lambda \boldsymbol{D}^T \boldsymbol{D})$ is a symmetric, positive-definite, tridiagonal matrix, which can be solved efficiently (Golub and Van Loan, 1996). Wander components are extracted from sEMG via QVR, which can be seen in Figure 3.

Subfigures on the right of Figure 2 show the processed results after filtering and QVR, which indicate an improvement to the raw data.

Similar to speech recognition, Mel Frequency Cepstral Coefficients (MFCCs) are extracted as features in SSR in this paper. MFCCs reveal signal's spectrum distribution in different frequency intervals, and have proven their effectiveness in silent speech recognition in several related papers (Lee, 2008, Lopez-Larraz et al., 2010, Deng et al., 2012, Meltzner et al., 2018). We investigated several different dimensions of MFCCs for our sEMG channels, such as 60, 80 and 100, and 80 gave the best results. Therefore, each channel is processed to obtain 80 MFCCs under one large window. Then the dimension of the final feature vector is $6 \times 80 = 480$, which will be used for classification.

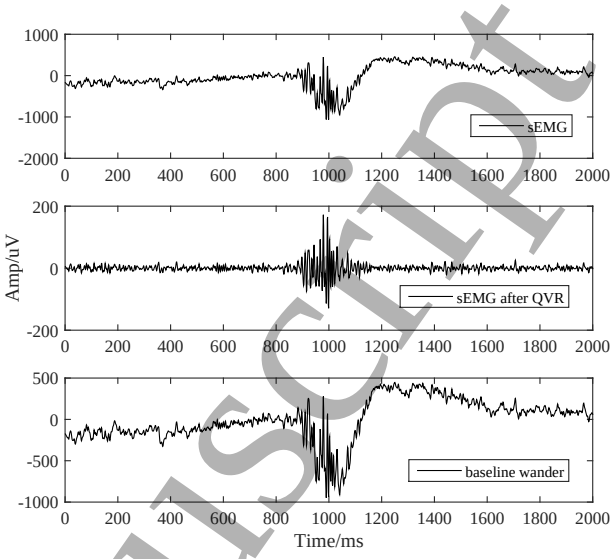


Figure 3. sEMG wander removal. Raw signal comes from channel 1 in Figure 2.

Our entire silent speech recognition process is depicted in Figure 4. The 'feedback' in this figure gives information about the results of silent speech recognition in the forms of voice, visual, or text to the user.

2.3. Random forest model

In this work, we consider the use of a random forest classifier as the underlying algorithm of ICP. Previous research demonstrated that the random forest method does not suffer from overfitting when more trees are added (Devetyarov and Nouretdinov, 2010). Moreover, the random forest method is capable of processing high-dimensional data where each feature carries less information. Another attractive property of the random forest method is that it is robust to noise, missing values and outliers (Breiman, 2001, de Santana et al., 2018).

The following parameters for the random forest construction are used in our experiments: the number of trees is 500, the 'gini' criterion is used to measure the quality of split at nodes, and the depth of trees for forest optimization is 10.

2.4. Inductive conformal prediction

Assuming each example of our data set has the format of $z_i = (x_i, y_i)$, where x_i is the feature vector describing the recorded sEMG signal and y_i is the corresponding label (i.e. command). y_i is unknown for any example

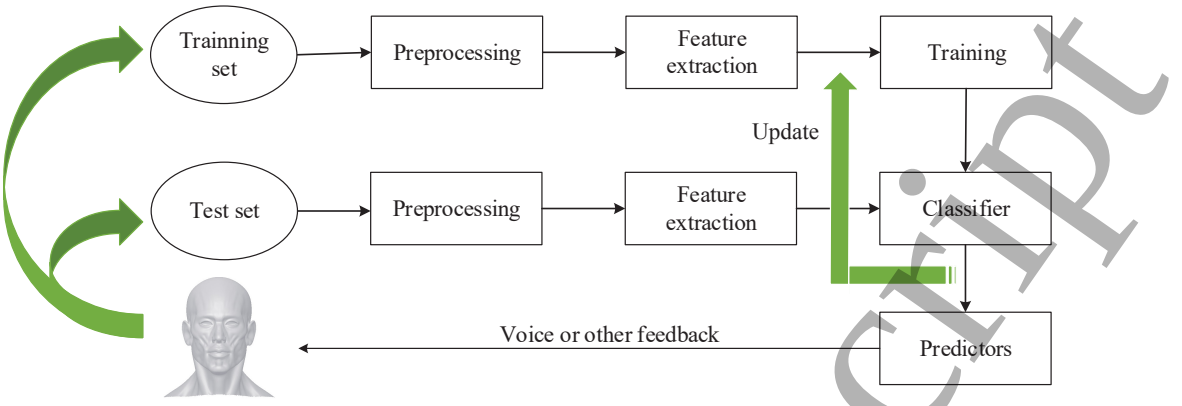


Figure 4. Silent speech recognition process. Note that the test set (new unlabelled samples or examples) can be utilized to supplement into the training set if the new samples can meet a satisfactory reliable level.

in the test set. In this paper, $y_i \in \boldsymbol{Y}$, where $\boldsymbol{Y}$ is the label space and we are dealing with classification problem. In ICP, the training set $\{z_1, z_2, ..., z_l\}$ is randomly divided into two parts: proper training set $\{z_1, z_2, ..., z_m\}$ and calibration set $\{z_{m+1}, z_{m+2}, ..., z_l\}$, where $m < l$. The former is used to build a proper training model using an underlying algorithm and calculate the nonconformity score while the latter computes p-values for all possible labels (Shafer and Vovk, 2008, Devetyarov and Nouretdinov, 2010, Wang et al., 2019).

For a new example x_{l+1}, p-values are computed according to Equation (3) for a possible label $c \in Y$:

$$p^c = \frac{|i = m+1, m+2, ..., l | \alpha_i^c \geq \alpha_{l+1}^c| + 1}{l - m + 1} \quad (3)$$

where $\alpha_i^c = \boldsymbol{A}(\{z_{m+1}, z_{m+2}, ..., z_l\}, z_i)$, $\alpha_{l+1}^c = \boldsymbol{A}(\{z_{m+1}, z_{m+2}, ..., z_l\}, (x_{l+1}, c))$, $\boldsymbol{A}$ is a nonconformity measurable function which is defined by the underlying algorithm, and c is a possible label (Johansson et al., 2013, Toccaceli and Gammerman, 2019). Nonconformity (or strangeness) measure $\boldsymbol{A}$ is a way of scoring how different a new example is from a set of previously observed instances. For the random forest method, the nonconformity measure is defined in Equation (4), where y_i denotes the possible labels in prediction and y the true label.

$$\boldsymbol{A} = 0.5 - \frac{\hat{P}(y_i|x) - max_{y != y_i} \hat{P}(y|x)}{2} \quad (4)$$

With computed p-values, there are two methods to obtain the predicted labels (Vovk, 2012, Wang et al., 2017). The first method is to set a significance level $\epsilon \in [0, 1]$. Given ϵ, then the predicted output is denoted by $\boldsymbol{\Gamma}^\epsilon(z_{m+1}, z_{m+2}, ..., z_l, x_{l+1}) = \{c | p^c > \epsilon\}$. This indicates that the ICP may output a prediction region which can contain empty, one or multiple labels.

The second method is forced prediction, which produces a single prediction label with the highest p-value, i.e. $\boldsymbol{\Gamma}_{force}(z_{m+1}, z_{m+2}, ..., z_l, x_{l+1}) = \{c | max(p^c)\}$. Furthermore, ICP method supplements the forced prediction with two measures, confidence and credibility, to describe the quality of the prediction (Devetyarov and Nouretdinov, 2010, Wang et al., 2019). In terms of p-values assigned to all possible labels, the confidence is a complement to 1 of the second largest p-value and the credibility is the largest p-value.

Intuitively, confidence represents the highest confidence level at which the prediction is certain (i.e. the prediction region consists of at most one label). Credibility indicates how suitable the training data are for classifying the new example. If the credibility is low, this means that any existing hypothesis about the label of the new example is unlikely. The ideal situation ('clean and easy' data set) would have both confidence and credibility close to 1.

2.5. Test time data augmentation using ICP

Conformal prediction has two basic modes: offline and online. Offline mode indicates the training set $\{z_1, z_2, ..., z_l\}$ will not be updated after training. For online mode, however, examples are presented and tested one by one in a sequence. Each example is added to the training set with its true label and then update the model. In this paper, we consider test time data augmentation approach and assume that there are large number of unlabelled examples available in addition to the training and test sets. Achieving acceptable classification accuracy by machine learning algorithms usually requires large amount of labelled data to be used for training the algorithms. However, labeling data typically has to be done manually and therefore it is a time consuming and expensive task by itself in many applications. There are many scenarios where unlabelled data is plentiful and easy to obtain. We propose that these unlabelled examples can be classified by ICP and their corresponding confidence and credibility are then used to decide whether the

unlabelled examples can be added to the training set with their predicted labels in order to improve the performance on the test set. Our approach is similar to active learning, but it does not require the user to provide the true labels for the selected examples (Wang and Kwong, 2014, Matiz and Barner, 2019).

Suppose $\boldsymbol{Y}$ is the label space, j is the size of test subset, and $\boldsymbol{P}$ represents p-values for new examples. The implementation of the test time data augmentation using ICP algorithm is described by Algorithm 1.

Algorithm 1: Test time data augmentation for ICP implementation.

Input: training set $\boldsymbol{Z_{trn}}$, unlabelled set $\boldsymbol{Z_{unlabelled}}$;
Output: augmented training set $\boldsymbol{Z_{trn}^{aug}}$

Fix a significance level: $\epsilon \in (0,1)$;
New set: $\boldsymbol{Z_{new}^{\epsilon}} = \text{null}$;
Prediction size: $\boldsymbol{S_{predidction}^{\epsilon}} = 0$
p-values: $\boldsymbol{P} = 0$
for *each* $j \in \boldsymbol{Z_{unlabelled}}$ **do**
    Test features: $x_j^{\epsilon} \in \boldsymbol{Z_{unlabelled}}$;
    Calculate $\boldsymbol{P}$ for x_j^{ϵ} using ICP: $p_i \in (0,1), i = 0, 1, ..., 9$;
    Predicted labels: $\Gamma_{l,j}^{\epsilon} \subseteq \boldsymbol{Y}$;
    if $\boldsymbol{S_{predidction}^{\epsilon}} \geq 1$ & $p_{max} \geq 3 \times p_{secmax}$:
        Force output: $y_j^{\epsilon} \subseteq \boldsymbol{Y}$;
        Add to new set: $\boldsymbol{Z_{new}^{\epsilon}} \leftarrow (x_j^{\epsilon}, y_j^{\epsilon})$;
    end
end
Update training set $\boldsymbol{Z_{trn}^{aug}} = \boldsymbol{Z_{trn}} \cup \boldsymbol{Z_{new}^{\epsilon}}$

In Algorithm 1, new unlabelled examples will be used to update the training set if the p-values and significance level ϵ satisfy two specific conditions: one is that the prediction size $\boldsymbol{S_{predidction}^{\epsilon}}$ should be lager or equal to one under a specific significance ϵ; the other one should be that the largest p-value is much larger than the second biggest one, $p_{max} \geq 3 \times p_{secmax}$. The latter one generally means obtaining high credibility of predictions since the second largest and other smaller p-values themselves present their invalidity. When the training size increases to infinite, the error rate of ICP will not exceed the given significance level ϵ, where Err_n^{ϵ} is the accumulated number of errors for n examples.

$$\lim_{n\to\infty} sup \frac{Err_n^{\epsilon}}{n} \leq \epsilon \tag{5}$$

This character is regarded as one typical advantage of CP method compared with other machine learning methods.

As described before, the random forest method is used as the underlying algorithm of inductive conformal prediction. The idea of test time data augmentation using ICP is investigated in the following experiment. The silent speech data satisfy the *i.i.d.* assumption and our data are randomly split into three parts: training set, active set (i.e. unlabelled set) and test set. The training set is responsible for building prediction model while the active set is considered as new unlabelled examples which will be predicted by the classifier and then reliable ones are appended into the training set. The test set is used to evaluate the effects of ICP in SSR. The scheme is displayed in Figure 5. For illustration purposes, the active set is randomly divided into 30 equal parts (namely $p1, p2, ..., p30$) to be predicted by ICP and certain examples will be selected to augment the training set. At the end of each part, the updated training data are used to predict the test data. Note that for p_i $(i = 1, 2, ..., 30)$, it means pi will contain all unlabelled examples in $p1 \cup p2 \cup ... \cup pi$, not just examples in part pi.

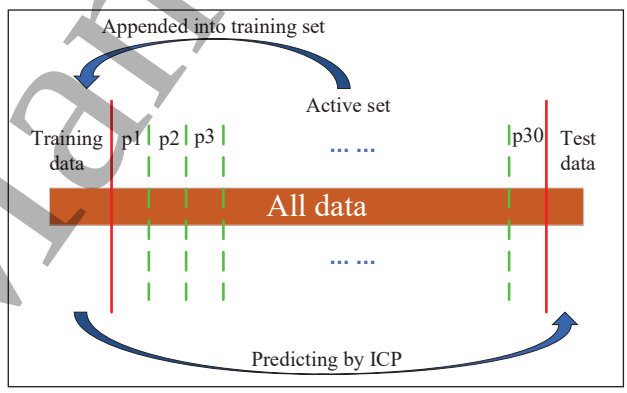


Figure 5. Test time data augmentation scheme. For the 30 parts in active set, data are predicted by ICP with random forest from the start of $p1$ to the end of $p30$.

3. Results

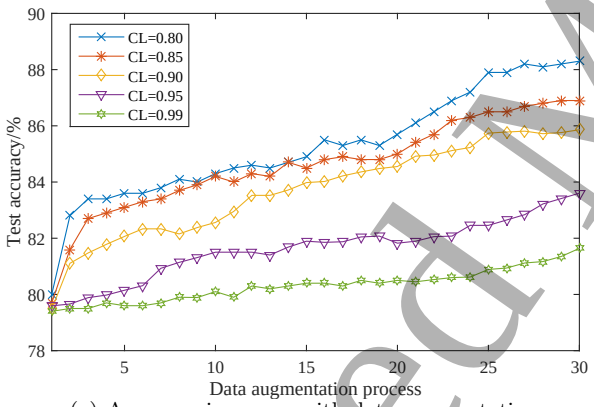
3.1. Improvement by ICP

In our experiment, the training data are randomly divided into a proper training set and a calibration set at the ratio of 7:3. As discussed before, random forest method is chosen as the underlying algorithm. Different significance levels are tried to explore the variations of prediction accuracy, p-values, and changes of the training set. For each significance level, 30 prediction results are obtained as the active set is divided into 30 parts. For each new example, ICP can output p-values, conformal predictions, and their corresponding confidence and credibility values. All results reported below are obtained by averaging 10 different experimental runs.

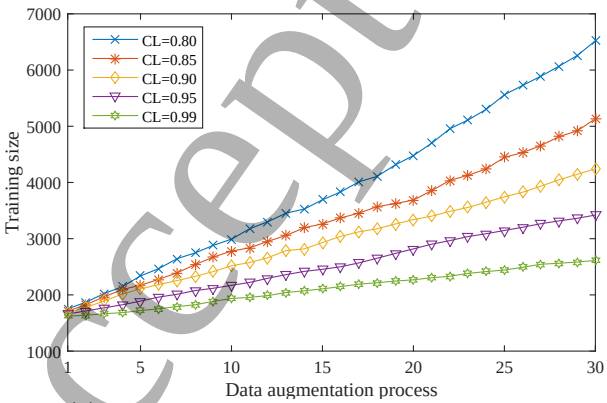
Table 2 shows the predictions of an example of the command 'decelerate' (i.e. label '8') in the test

set with different augmented training models when the significance level is set to $\epsilon = 0.10$. The first column on the left means the ith prediction results after using unlabelled data, $i = 0, 1, 5, 10, 15, 20, 25, 30$. Here, $i = 0$ means the original training set without any augmented unlabelled examples. The size of the training set increases as the number of unlabelled data gets larger. p-values of the prediction of the word 'decelerate' are presented to show each possible label and the forced ICP predicted label (it is the label '8')with the corresponding confidence and credibility. It is clear that the confidence level and credibility of the predicted commend show an increasing trend which indicates more unlabelled data could be used to improve the performance of ICP on the test set.

Figure 6 describes the changes of accuracy on test set along with the increases of training set using our approach. 'CL' in figure denotes confidence level which equals to 1-ϵ. In this experiment, 10% of the whole data are chosen to be the training set (1628 samples), while 5% act as an independent test part (814 samples), and the remaining 65% of the data are used for unlabelled part (10585 samples) in ICP.



(a) Accuracy increases with data augmentation



(b) Training size increases with data augmentation

Figure 6. Changes of test accuracy and training size under different CLs of ICP

The accuracy in Figure 6a shows an slow and zigzag improvement as the training set enlarges. The smaller the significance level is, the lower increment of accuracy will be obtained because fewer new samples could be added to the training set, but the confidence can be further guaranteed. Specifically, for a significance level of $\epsilon = 0.01$, there is only a slight accuracy increase of around 1% because only small number of examples meets the conditions, as shown in Figure 6b. In contrast, a higher significance level may enhance the accuracy on the test set, yet sacrificing the reliability and credibility.

The effects of the training set size are also investigated in this experiment and their results are shown in Figure 7, where 'OriData1', 'OriData2' and 'OriData3' denote the size of original data at 10%, 20%, and 30% of the whole data set, respectively. At the same time, the unlabelled set and test set remain unchanged. Large size of training set is likely to contain plentiful information, thus increasing model quality, therefore only a small improvement is achieved. However, this still helps improve the confidence in predicting new examples, similar to the results presented in Table 2.

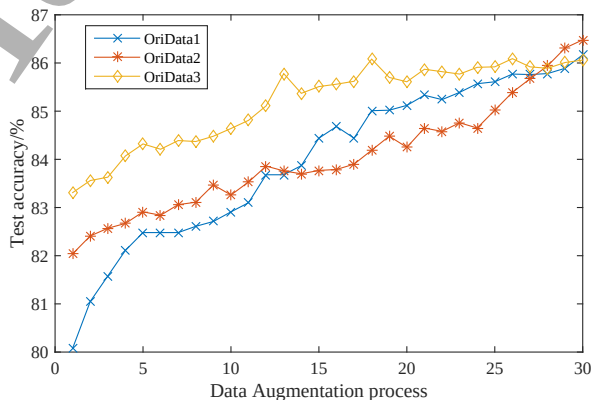


Figure 7. Effects of original training size at confidence of 0.90

3.2. Prediction analysis

An improvement for test accuracy is shown in Figure 6a. The accuracy rate is the average correct prediction for the ten commands in test set. More detailed results are depicted in Figure 8, where the best ICP results in our experiments are presented. The confusion matrix illustrates the rate of being predicted a label against the true label as well as individual accuracy rates for each possible label and a correct prediction. The command with the label '8' achieves the best performance whereas the command with the label '6' has the worst performance. A more prevailing phenomenon is that the commands with the labels '3'

Table 2. Conformal predictors of one word (label '8'). Typically, ICP can output p-values, confidence and credibility for each example, and according which the prediction can be properly evaluated.

No.	p-value for the possible label										Confidence	Credibility
	0	1	2	3	4	5	6	7	8	9		
0	0.014	0.015	0.016	0.016	0.016	0.014	0.016	0.017	0.503	0.015	0.983	0.503
1	0.013	0.013	0.012	0.013	0.014	0.014	0.016	0.015	0.536	0.013	0.985	0.536
5	0.011	0.010	0.011	0.012	0.010	0.011	0.010	0.012	0.568	0.010	0.986	0.568
10	0.009	0.009	0.008	0.008	0.009	0.010	0.008	0.008	0.610	0.009	0.990	0.610
15	0.011	0.010	0.009	0.009	0.009	0.009	0.010	0.010	0.607	0.012	0.989	0.607
20	0.008	0.008	0.007	0.008	0.008	0.006	0.006	0.009	0.656	0.009	0.990	0.656
25	0.004	0.004	0.005	0.005	0.005	0.004	0.006	0.004	0.682	0.004	0.994	0.682
30	0.005	0.005	0.005	0.004	0.005	0.005	0.005	0.006	0.692	0.004	0.994	0.692

and '4' are more likely to be incorrectly predicted as the command with the label '6'. On the other hand, the command with the label '6' seems to be easily recognised as the commands with the labels '2' to '7', especially '4'. This may be due to that the command associated with the label '6' is phonetically similar to the command with the label '4' in Chinese.

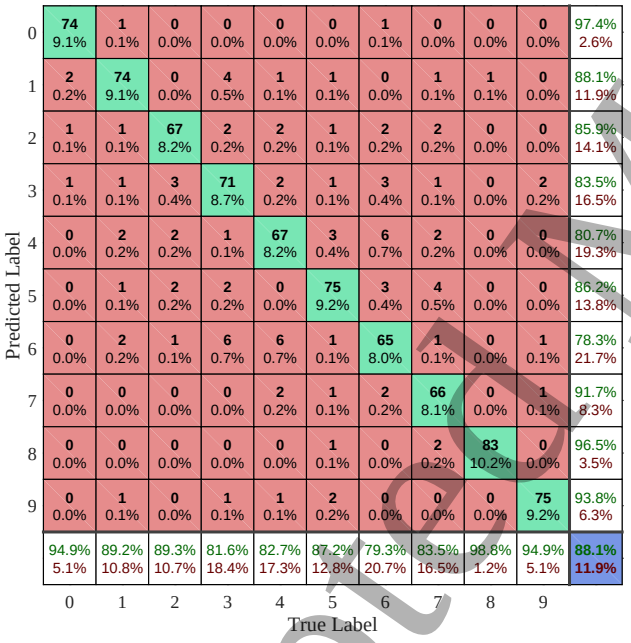


Figure 8. Confusion matrix of force prediction. At the bottom and the most right rows, diagonal cells (in green) indicate accuracy while the others (in red) show error rates. For each cell, the integer represents the number of predictions corresponding to the label on axis, and the number below is the corresponding percentage.

In data augmentation, we obtained best classification results at around 0.88, as shown in Figure 6a and Figure 8. When using all data for training (minus the test set)and then test, the accuracy was 0.85, which is not as good as prediction of ICP with test time data augmentation. We do not know the true labels for new examples. Therefore, without knowing how reliable the predictions are, it is a gamble to add them directly into training set. However, reliable predictions can be obtained by ICP to evaluate how much we can trust the predictions. Only qualified unlabelled data are included in the training set for improvement of future predictions.

3.3. Validity and efficiency of ICP

There are two measures to evaluate the performance of ICP for region prediction: validity and efficiency. Validity means the error rate should not be greater than the specified significance level. The notion of the error rate means ratio of the number of prediction region of examples which do not contain the true label to all the predicted examples (Vovk, 2015). Ideally, the error rate of prediction should equal to the given significance level.

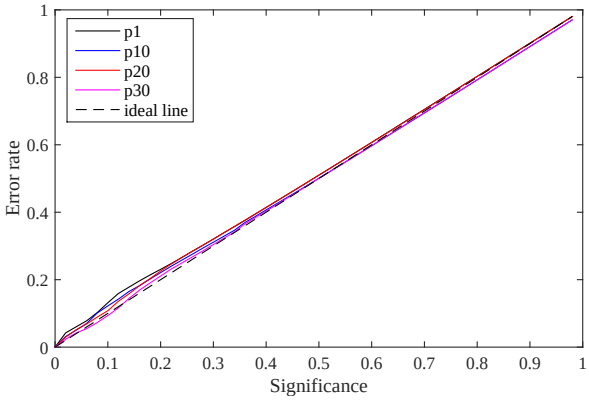


Figure 9. Validity of ICP in SSR

Figure 9 shows four typical plots of significance level and the corresponding prediction error rate on the test set. The plots basically represent the validity in our experiment since small differences exist comparing to other plots. We say that a set of predictors is exactly valid at a significance level ϵ if the probability of a new example makes an error is equal to or just near ϵ. As depicted in the figure, the error rate always follows the changes of significance level, showing the validity. The

lines shown in the figure are near the ideal line and are in minor differences, which verifies the validity of the predictors.

Efficiency is another measure for evaluating ICP, which reflects the informativeness of prediction.One common way to represent efficiency is to consider the size of prediction region in our experiment. The size of prediction region should be as small as possible and size one is the best. The average sizes of prediction region are shown in Figure 10.

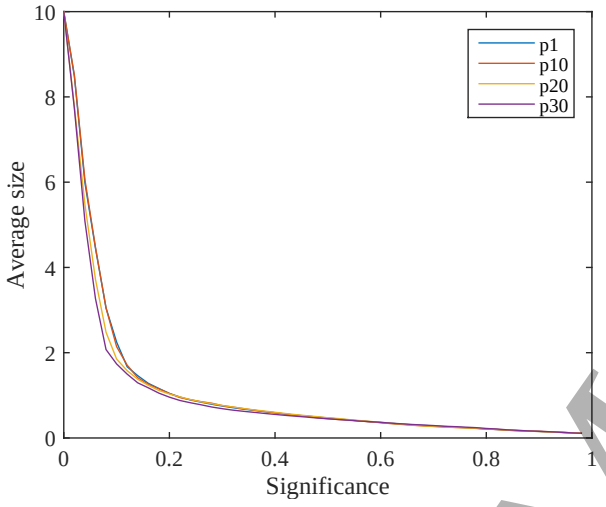


Figure 10. Efficiency of ICP in SSR

Four different subplots for the training set augmented with different size of the unlabelled data are shown in this figure. It is clear that the prediction region size increases quickly with low significance level. There are slightly differences among the four plots that $p30$ and $p20$ are a little better than $p10$ and $p1$. This phenomenon verifies the work in efficiency of test time data augmentation. Also, empty prediction size is possible for high significance level because only low confidence need to be guaranteed.

4. Discussion

4.1. Data recording

The data from surface electromyography play a decisive role in silent speech recognition. We studied the speech mechanism and the related articulatory muscles, carefully choosing six collection and two reference sites on the face and around it. Those positions are considered as a neuromuscular net that covers several accessible and dominant articulatory muscles. The positions shown in Figure 1 are similar to some previous research such as (Wand et al., 2014, Kapur et al., 2018). To capture the potentials on the muscles' surface, a home-made device named 'MiCap' is developed with excellent performance, and 10 monosyllabic logograms are selected for data recording by this device. Previous research applied high-pass filters on sEMG data with high cut-off frequencies of $5Hz$ or higher (Lopez-Larraz et al., 2010, Wand et al., 2014, Meltzner et al., 2018), which reduced drift artifacts but inevitably sacrificed some useful low-frequency components. In this paper, Quadratic Variation Reduction is introduced for baseline wander removal while retaining useful information. This method has been proven a computational complexity of $\mathcal{O}(n)$, and at the same time shows great results in Figure 3.

4.2. Inductive conformal prediction processing

For the unlabelled active silent speech data, test time data augmentation is applied to increase the training set and update the training model. Inductive conformal prediction is used for the whole augmentation process by computing confidence and credibility for every example. The metrics of ICP are naturally suitable for exploring trusted examples from a mass unlabelled dataset, comparing methods based on single or multi criterion in active learning (Settles and Craven, 2008, Wang and Kwong, 2014, Kapoor et al., 2015). In this paper, unlabelled examples of silent speech data are regarded as reliable ones under several conditions described in the third paragraph of Section 2.5. Confidence level can be set beforehand. Figure 6b shows the positive results of test time data augmentation under five different confidence levels. It is not surprising that high confidence level achieves low training data size increase.

The properties of ICP make it as a ideal candidate for predictions with guaranteed error rates in silent speech recognition applications. Previous research in SSR mainly focussed on prediction accuracy, usually just giving a error rate on a specific test set and ignoring how much we can trust it (Chan et al., 2001, Jorgensen et al., 2003, Denby et al., 2010, Schultz et al., 2017, Meltzner et al., 2018). On the other hand, ICP with the underlying algorithm of random forest can output p-values, confidence and credibility for individual examples, as displayed in Table 2. Because ICP outputs a p-value for each possible label, the maximum one is considered to be the credibility level and the corresponding label can be included in the prediction region, thus we know how much it can be trusted if needed. For predictions in our experiment as displayed in Figure 6a, the test accuracy shows a zigzag improvement with confidence levels at 0.80, 0.85, 0.90, 0.95, and 0.99 respectively. When the confidence level (CL) is set to 0.99, it means the prediction confidence should be at least 0.99 before the unlabelled example can be included in the training set. The size of the

training data also affects ICP predictors differently, see Figure 7. Although the increased training set size produces a limited improvement, ICP can give better predictions.

The validity and efficiency of ICP are explored and demonstrated in Figure 9 and Figure 10. Validity indicates how reliable the predictions are. ICP produces provably valid measures of confidence in predictions for individual examples without assuming anything more than that the data are generated independently from the same probability distribution and the framework is general. Figure 9 shows valid predictions when applying ICP in SSR. Since the output of ICP is a prediction set rather than just a single label, the size of the set is important. To make the prediction more efficient, the output prediction region should be as small as possible, but not empty, and a size of one is the best. Hence, there is always a confidence-efficiency trade-off in practical applications because high confidence generally requires the larger prediction region size will in order to maintain validity. As illustrated in Figures 9 and 10, the ICP performed similarly for $p1$, $p10$, $p20$, and $p30$ when validity and efficiency are evaluated.

5. Conclusions

In this paper, a novel method of inductive conformal prediction was introduced to improve the performance of silent speech recognition on a sEMG dataset collected by our own device. Experimental results showed that ICP offers excellent guarantees on prediction reliability for new SSR examples. It outputs p-values which can be used to obtain confidence and credibility; this tells us how much we can trust the prediction results. As a result, a test time data augmentation method was proposed based on ICP to utilize unlabelled data in order to improve prediction performance further on the test set. Our experiments demonstrated that a bigger training set may achieve a smaller improvement in accuracy, but new examples in the test set may be predicted with better confidence and credibility. Test time data augmentation experiments have demonstrated its tremendous value and potential in realistic machine learning applications. In general, ICP for SSR can be applied in risk-sensitive situations to avoid fatal errors.

6. Grants

This work is supported by the Natural Science Foundation of China (Grant No. 61773342) and the Autonomous Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT1914). The views and conclusions in this document are those of the authors and should not be interpreted as representing any of the above official policies.

References

Anumanchipalli, G. K., Chartier, J. and Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences, *Nature* **568**(7753): 493.

Breiman, L. (2001). Random forests, *Machine learning* **45**(1): 5–32.

Chan, A. D., Englehart, K., Hudgins, B. and Lovely, D. F. (2001). Myo-electric signals to augment speech recognition, *Medical and Biological Engineering and Computing* **39**(4): 500–504.

de Santana, F. B., de Souza, A. M. and Poppi, R. J. (2018). Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters, *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **191**: 454–462.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M. and Brumberg, J. S. (2010). Silent speech interfaces, *Speech Communication* **52**(4): 270–287.

Deng, Y., Colby, G., Heaton, J. T. and Meltzner, G. S. (2012). Signal processing advances for the mute semg-based silent speech recognition system, *MILCOM 2012-2012 IEEE Military Communications Conference*, IEEE, pp. 1–6.

Devetyarov, D. and Nouretdinov, I. (2010). Prediction with confidence based on a random forest classifier, *IFIP International Conference on Artificial Intelligence Applications and Innovations*, Springer, pp. 37–44.

Fasano, A. and Villani, V. (2014). Baseline wander removal for bioelectrical signals by quadratic variation reduction, *Signal Processing* **99**(6): 48–57.

Fraiwan, L., Lweesy, K., Al-Nemrawi, A., Addabass, S. and Saifan, R. (2011). Voiceless arabic vowels recognition using facial emg, *Medical & biological engineering & computing* **49**(7): 811–818.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations (3rd ed.)*.

Goodier, J. (2017). The complete human body: The definitive visual guide (revised edition), *Reference Reviews* .

Hakonen, M., Piitulainen, H. and Visala, A. (2015). Current state of digital signal processing in myoelectric interfaces and related applications, *Biomedical Signal Processing and Control* **18**: 334–359.

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G. and Stone, M. (2010). Development of a silent speech interface driven by ultrasound

and optical images of the tongue and lips, *Speech Communication* **52**(4): 288–300.

Ji, Y., Liu, L., Wang, H., Liu, Z., Niu, Z. and Denby, B. (2018). Updating the silent speech challenge benchmark with deep learning, *Speech Communication* **98**: 42–50.

Johansson, U., Boström, H. and Löfström, T. (2013). Conformal prediction using decision trees, *2013 IEEE 13th international conference on data mining*, IEEE, pp. 330–339.

Jorgensen, C. and Dusan, S. (2010). Speech interfaces based upon surface electromyography, *Speech Communication* **52**(4): 354–366.

Jorgensen, C., Lee, D. D. and Agabont, S. (2003). Sub auditory speech recognition based on emg signals, *Proceedings of the International Joint Conference on Neural Networks, 2003.*, Vol. 4, IEEE, pp. 3128–3133.

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F. and Waibel, A. (2006). Towards continuous speech recognition using surface electromyography, *Ninth International Conference on Spoken Language Processing*, pp. 573–576.

Kapoor, A., Grauman, K., Urtasun, R. and Darrell, T. (2015). Active learning with gaussian processes for object categorization.

Kapur, A., Kapur, S. and Maes, P. (2018). Alterego: A personalized wearable silent speech interface, *23rd International Conference on Intelligent User Interfaces*, ACM, pp. 43–53.

Kass, J. S. and Mizrahi, E. M. (2010). *Neurology Secrets E-Book*, Elsevier Health Sciences.

Katirji, B. (2007). *Electromyography in Clinical Practice E-Book: A Case Study Approach*, Elsevier Health Sciences.

Kenneth, S. (2010). Anatomy & physiology: The unity of form and function.

Lee, K.-S. (2008). Emg-based speech recognition using hidden markov models with global control variables, *IEEE Transactions on biomedical engineering* **55**(3): 930–940.

Lei, J., Rinaldo, A. and Wasserman, L. (2015). A conformal prediction approach to explore functional data, *Annals of Mathematics and Artificial Intelligence* **74**(1-2): 29–43.

Löfström, T., Johansson, U. and Boström, H. (2013). Effective utilization of data in inductive conformal prediction using ensembles of neural networks, *The 2013 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1–8.

Lopez-Larraz, E., Mozos, O. M., Antelis, J. M. and Minguez, J. (2010). Syllable-based speech recognition using emg, *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, IEEE, pp. 4699–4702.

Manabe, H., Hiraiwa, A. and Sugimura, T. (2003). Unvoiced speech recognition using emg-mime speech recognition, *CHI'03 extended abstracts on Human factors in computing systems*, ACM, pp. 794–795.

Matiz, S. and Barner, K. E. (2019). Inductive conformal predictor for convolutional neural networks: Applications to active learning for image classification, *Pattern Recognition* **90**: 172–182.

Meltzner, G. S., Heaton, J. T., Deng, Y., De Luca, G., Roy, S. H. and Kline, J. C. (2018). Development of semg sensors and algorithms for silent speech recognition, *Journal of neural engineering* **15**(4): 046031.

Nguyen, K. A. and Luo, Z. (2015). Reliable indoor location prediction using conformal prediction, *Annals of Mathematics and Artificial Intelligence* **74**(1-2): 133–153.

Papadopoulos, H., Vovk, V. and Gammermam, A. (2007). Conformal prediction with neural networks, *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Vol. 2, IEEE, pp. 388–395.

Preston, D. C. and Shapiro, B. E. (2012). *Electromyography and Neuromuscular Disorders E-Book: Clinical-Electrophysiologic Correlations (Expert Consult-Online and Print)*, Elsevier Health Sciences.

Schultz, T. and Wand, M. (2010). Modeling coarticulation in emg-based continuous speech recognition, *Speech Communication* **52**(4): 341–353.

Schultz, T., Wand, M., Hueber, T., Krusienski, D. J., Herff, C. and Brumberg, J. S. (2017). Biosignal-based spoken communication: A survey, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **25**(12): 2257–2271.

Settles, B. and Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks, *Conference on Empirical Methods in Natural Language Processing*.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction, *Journal of Machine Learning Research* **9**(Mar): 371–421.

Sun, J., Carlsson, L., Ahlberg, E., Norinder, U., Engkvist, O. and Chen, H. (2017). Applying mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets, *Journal of chemical information and modeling* **57**(7): 1591–1598.

Toccaceli, P. and Gammerman, A. (2019). Combination of inductive mondrian conformal predictors, *Machine Learning* **108**(3): 489–510.

Vovk, V. (2012). Conditional validity of inductive conformal predictors, *Asian conference on machine learning*, pp. 475–490.

Vovk, V. (2015). Cross-conformal predictors, *Annals of Mathematics and Artificial Intelligence* **74**(1-2): 9–28.

Vovk, V., Gammerman, A. and Shafer, G. (2005). *Algorithmic learning in a random world*, Springer Science & Business Media.

Wand, M., Janke, M. and Schultz, T. (2014). Tackling speaking mode varieties in emg-based speech recognition, *IEEE Transactions on Biomedical Engineering* **61**(10): 2515–2526.

Wang, R. and Kwong, S. (2014). Active learning with multi-criteria decision making systems, *Pattern Recognition* **47**(9): 3106–3119.

Wang, Y., Wang, Z., Diao, J., Sun, X., Luo, Z. and Li, G. (2019). Discrimination of different species of dendrobium with an electronic nose using aggregated conformal predictor, *Sensors* **19**(4): 964.

Wang, Z., Sun, X., Miao, J., Wang, Y., Luo, Z. and Li, G. (2017). Conformal prediction based on k-nearest neighbors for discrimination of ginsengs by a home-made electronic nose, *Sensors* **17**(8): 1869.