Term paper for Acoustic Phonetics WS20/21 Wang and Lee (2015) Summary and Terminology

Jingwen LI and Kuan TANG [1]

[1] Seminar für Sprachwissenschaft, Eberhard Karls Universität Tübingen *jingwen.li/kuan.tang@student.uni-tuebingen.de*

28. Mar 2021

This term paper aims to introduce the main concepts and methods of Wang and Lee (2015). The content and authorship of this paper as follows: In the first section, Kuan summaries the whole paper in general, and make clear that the research questions, methodology, and experimental results in Wang and Lee (2015). In section 2, Jingwen introduces the new frameworks for both Supervised EP detection and Unsupervised EP discovery respectively through simple and vivid examples to explain the core issues of the (un)supervised acoustic patterns. On this basis, Kuan gives more details about the concepts of Clustering and Metric and the applications of both in the paper in section 3, and then put forward two questions after reading the full text in the end.

1. Summary

How to apply the Speech Processing Technology in Second Language Learning (SLL)? How to generate informative feedback for Second Language Learners (SLLs) to improve their pronunciation based on Computer-assisted language learning (CALL) and Computer-aided pronunciation training (CAPT)[2]? How to offer SLLs not only quantitative measures of language proficiency, but also specific types of errors the they have made? These are three trends which have led to substantial efforts toward CALL to meet the strong demand of SLL. At least two different but closely related Pronunciation error patterns (EPs)[3] tasks in CAPT: First, to derive the EP dictionary for a given L1-L2 pair, or a given L2 but non-specific L1; Second, to verify whether a voice segment produced by a learner is correct, or if it belongs to a specific EP based on the EP dictionary.

[2] Computer-aided pronunciation training (CAPT) aims to analyze the produced utterance to offer feedback to the language learner in the form of quantitative or qualitative evaluations of the pronunciation proficiency.

[3] Pronunciation error patterns (EPs) are patterns of mispronunciation frequently produced by language learners, and are usually different for different pairs of target and native languages.

Wang and Lee (2015) propose two new novel frameworks for both Supervised Error Patterns Detection and Unsupervised EP Discovery. For supervised EP detection, they use hierarchical multi-layer perceptrons (MLPs) as the EP classifiers to be integrated with the baseline using HMM/GMM in a two-pass Viterbi decoding architecture. As for unsupervised EP discovery, they use the Hierarchical Agglomerative Clustering (HAC) algorithm to explore sub-segmental variation within phoneme segments and produce fixed-length segment-level feature vectors to distinguish different EPs.

They tested K-means [4] and the Gaussian mixture model with the

[4] Assuming a known number of EPs

minimum description length principle [5] for EP discovery. And they also propose to use the universal phoneme posteriorgram (UPP), derived from an MLP trained on corpora of mixed languages, as frame-level features in both supervised detection and unsupervised discovery of EPs.

[5] Estimating an unknown number of EPs

Preliminary experiments offered very encouraging results. As the experimental results shown, comparing with other models, the new framework enhances the power of EP diagnosis, and using UPP achieves the best performance and is useful in analyzing the mispronunciation produced by language learners.

2. *Theoretical Framework*

2.1 *Error Patterns*

Error patterns (EPs) in pronunciation training are frequently encountered mispronunciations that have intrinsic commonalities. In other words, mispronunciations of the same error pattern usually display a high level of intra-group similarity, although they can still be very close to the canonical pronunciation or other EPs.
The discovery and detection of EPs can help give more informative and targeted feedbacks in pronunciation training. For example, language learners of different L1s may all pronounce the word wrong (wrong as in deviatation from the canonical pronunciation), but their mispronunciations can be incorrect in different ways. It would be more helpful if the CAPT system can provide feedbacks that are specific enough to show the area of improvement. EPs answer the question "In what way is the given pronunciation wrong?" instead of simply stating it as an incorrect pronunciation.

2.1.1. *Supervised EP detection*

2.1.2. *Unsupervised EP discovery*

2.2 *From Perceptrons to Multilayer Perceptrons*

A Perceptron is a single-neuron model that acts as a **binary linear classifier**. It is inspired by the biological neuron and can be seen as an abstraction of it. The perceptron receives a real-valued feature vector as input, it also has a weight vector associated with the input vector, which is to be trained. A weighted sum is calculated and passed into an activation function before the output layer. There are varies kinds of activation functions, they help normalise the weighted sum to a range between 0 and 1, or, in some cases, -1 and 1. Like the biological neuron needs an action potential above the threshold of -55 mA to "fire", the perceptron neuron uses the activation function and decides whether to pass on the output or not (acting like a gate), or, in the case of a binary classifier, it decides if the input belongs to a class or not. In the case of the example below, a step function is used where 0 is the threshold, everything above 0 is mapped to 1 and 0 otherwise.

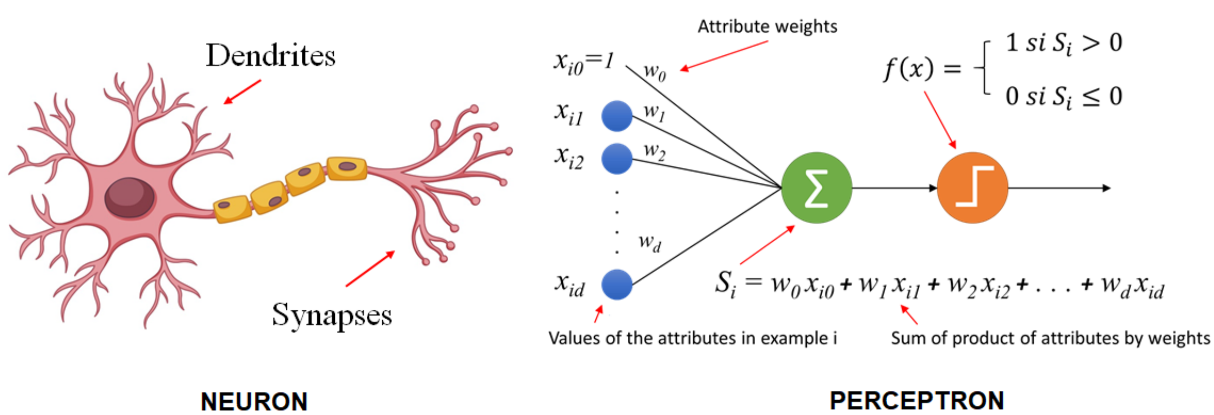

Figure 1: biological neuron vs. perceptron, adopted from Blog Inteligencia Futura (2019)

As suggested by the name, multilayer perceptrons (MLPs) have more than one (usually fully connected) layers. The most simple schematic MLPs illustrated in Figure 2 have one input layer, a hidden layer and an output layer. There can be more than one hidden layers, but usually one is enough. MLPs are a class of feedforward artificial neural network (ANN), with "feedforward" meaning single-way, acyclic movement, just like biological neurons.
MLPs use backpropagation (BP) as training method, which compares the results to the desired value and adjust the weights that contributed to the error most in each layer using the chain rule. The adjusted weights are used in the next training cycle. The training process terminates when the loss function is minimised.

2.3 *Universal Phoneme Posteriorgram*

Universal Phoneme Posteriorgram (UPP) is used as the fundamental frame-level feature of supervised EP discovery in Wang and Lee (2015). The motivation behind this is has to do with the nature of EPs in language learning. Usually EPs are caused by articulator mech-

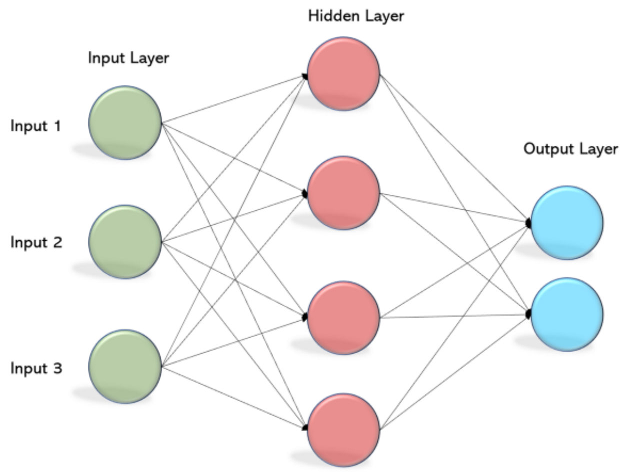

Figure 2: A schematic multilayer perceptron model, adopted from Mohanty (2019)

anisms or acoustic phenomena present in the target language but missing from learners' native languages as suggested by Wang and Lee (2015). However, the discrepancies between mispronunced instances and canonical pronounciations can be small but still obvious to the native ear. Therefore if one simply compare the pronunciations from learners of mixed languages in the same acoustic space, for example the acoustic space spanned by MFCCs, the pronunciations are usually too close to be distinguished by clustering algorithms. What UPP extraction does is to use an MLP as a posterior probability estimator to span the pronunciation instances by phoneme posteriors. Each frame of MFCC feature is transformed into a vector of posterior probabilities of all predefined phonemes in the training corpora of mixed languages.

2.4 Hidden Markov Model

2.5 Gaussian Mixture Model

3. Methods

3.1 Clustering

Before starting about Clustering, we should understand first what is a cluster. [6] Simply speaking, **Cluster** is the collection of data. The data objects in one cluster are similar to one another within the same group (class or category), and are different from the objects in the other clusters. The closer the objects in a cluster, the more likely they belong to the same cluster.

[6] This part is based on Seema Singh (2018) "An Introduction To Clustering". Here is the Web link `https://medium.datadriveninvestor.com/an-introduction-to-clustering-61f6930e3e0b`

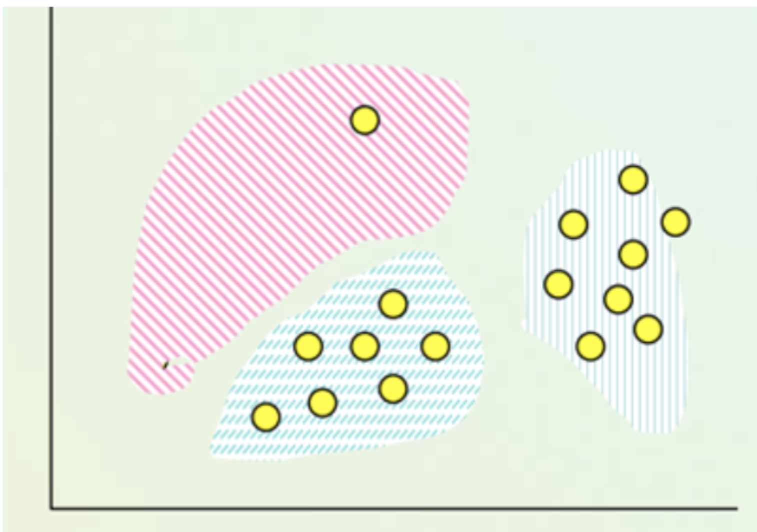

Figure 3: clustering, adopted from `https://medium.datadriveninvestor.com/an-introduction-to-clustering-61f6930e3e0b`

Based on the definition of the Cluster, as an unsupervised learning technique in which there is predefined classes and prior information, Clustering means how the data should be grouped or labeled into separate classes. It could also be considered as Exploratory Data Analysis (EDA) [7] process which help us to discover hidden patterns of interest or structure in data. Clustering can work as a standalone tool to get the insights about the data distribution or as a preprocessing step in other algorithms.

[7] Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns to spot anomalies to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

In Wang and Lee (2015), they uses Hierarchical Agglomerative Clustering (HAC) to define segment-level features. According to Chhabra and Mohapatra (2020), Hierarchical Agglomerative Clustering (HAC) refers to a class of greedy unsupervised learning algorithms that seek to build a hierarchy between data points while clustering them in a bottom-up fashion. The HAC algorithm used in Wang and Lee (2015) automatically arranges the frames in a speech segment into a tree-structured hierarchy based on the merging order of sub-segments, in which similar adjacent frames are clustered together in lower layers, while relatively dissimilar adjacent clusters are merged in higher layers.

3.2 Metrics

There are many different metrics for evaluating clustering algorithms. Metrics are measures of quantitative assessment commonly used for comparing, and tracking performance or production. Metrics can be used in a variety of scenarios. Wang and Lee (2015) define the pairwise true acceptance (TA′), true rejection (TR′), false acceptance (FA′), and false rejection (FR′) for clustering tasks based on all instance pairs. Let us explain these one by one.

The True Accept rate (TA′) is a statistic used to measure biometric performance when performing the verification task. It is the percentage of times a system (correctly) verifies a true claim of identity. While the True Reject rate (TR′) is a statistic used to measure biometric performance when performing the verification task. It refers to the percentage of times a system (correctly) rejects a false claim of identity.

False Acceptance (FA′) is the percentage of identification instances in which unauthorised persons are incorrectly accepted, and it occurs when an unauthorized subject is accepted as valid. While the False Rejection (FR′) is the percentage of identification instances in which authorised persons are incorrectly rejected. As the number of false acceptances (FA′) goes down, the number of false rejections (FR′) will go up and vice versa.

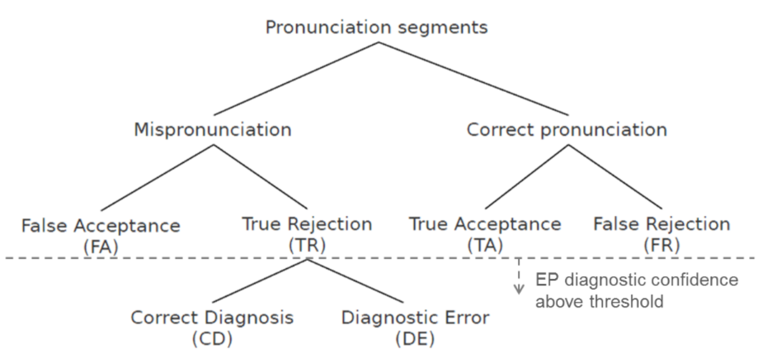

Figure 4: Hierarchical structure of the metrics used in EP detection adopted from Wang and Lee (2015)

4. Post-reading questions

In both supervised EP detection and unsupervised EP discovery, Wang and Lee (2015) propose to utilized the universal phoneme posteriorgram (UPP), derived from a multi-layer perceptron (MLP) trained on corpora of mixed languages, but if this methods can be applied to other dataset? and how to explain the differences of performance between the two tasks?

References

Blog Inteligencia Futura (2019). Deep neural networks, or Perceptron vs dogs and cats. [Online; accessed March 25, 2021].

Chhabra, A. and Mohapatra, P. (2020). Fair algorithms for hierarchical agglomerative clustering. *arXiv preprint arXiv:2005.03197*.

Mohanty, A. (2019). Multi layer Perceptron (MLP) Models on Real World Banking Data. [Online; accessed March 25, 2021].

Wang, Y.-B. and Lee, L.-s. (2015). Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):564–579.