**Supporting Information-A: Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) and Deep Neural Networks (DNN)**

HMM is a state transition model, trained separately for each class, and each of the states in an HMM generates acoustic signals $o\_{t}$ (*t* is time), whose features exclusively depend on the state. In other words, each state has its own output probability $P(o|c\_{i})$ and each HMM is trained with training data so that the trained HMM generates the training data with a higher probability. Here, it is usually assumed that $P(o|c)$ follows a Gaussian distribution (i.e., bell curve shape) or a mixture of Gaussian distributions. For ASR, with a given acoustic observation *o*, posterior probability $P(c|o)$ has to be calculated. For this calculation, $P(o|c)$ is referred to because $P(c|o)$ is proportional to $P(o|c)P(c)$.

On the other hand, DNN is trained so that it can output $P(c|o)$ directly, where DNN functions as a speech classifier and it shows how probably input observation *o* belongs to class *c*. Here, no statistical assumption, such as Gaussian assumption, is needed. With DNN, each state of the HMM can have its own DNN classifier, and GMM-HMM has been overcome by DNN-HMM with respect to their performances. By using the DNN acoustic model, any speech input $o\_{t}$ can be converted to its set of class posteriors {$P(c\_{i}|o\_{t})$}, $1\leq i\leq I$, where *I* is the total number of the speech classes. By using {$P(c\_{i}|o\_{t})$} to represent $o\_{t}$, speech sequence {$o\_{t}\}$ is converted to a temporal sequence of probability distributions or probability vectors. This probabilistic representation is often called as phone posteriograms.

**Supporting Information-B: Calculation of Posterior-Based Phonemes**

Notably, the number of phonemes used in posteriorgrams is not equivalent to the number of linguistically-driven phonemes. In automated speech recognition (ASR), a phoneme is often divided into three states (i.e., beginning, intermediate and ending states). In addition, a phoneme is often defined as per surrounding phonemic contexts. For instance, phonemes /x/ can be treated as different categories when they are found in various contexts (e.g., /axb/ vs. /cxd/). If the number of the linguistically-driven phonemes is N, then, the number of the context-dependent phonemes is *N*3. Furthermore, since a phoneme is divided into three states, the number of context-dependent phoneme states is 3\**N*3. If *N* = 50, 3\**N*3 = 375,000, which is the logically-driven number of the phoneme states, and is the logically-driven dimension of a posterior probability vector in the posteriorgram. However, this number is too huge, and by clustering the phoneme states down into a few thousands, the posteriogram with its dimension being 2,000 to 3,000 are widely used in the current ASR. When applying the posteriorgram to L2 speech assessment, the dimension can be further reduced by bottom-up clustering with Ward’s method, where the distance matrix between any pair of the states is used (see Kashiwagi et al., 2016).

**Supporting Information-C: Descriptive Statistics of Automated Measures**

*Descriptive Statistics of Automated Measures: Study 2*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *M* | *SD* | 95% *CI* |
|  | Low | Upper |
| A. Temporal quantity |  |  |  |  |
| Articulation rate | 3.406 | 0.424 | 3.279 | 3.534 |
| Pause ratio | 0.383 | 0.108 | 0.350 | 0.415 |
| B. Phonological quality |  |  |  |  |
| Maximum posterior probabilities | 0.802 | 0.022 | 0.795 | 0.809 |
| Posterior gaps to natives | 0.086 | 0.021 | 0.079 | 0.093 |
| C. Prosodic quality |  |  |  |  |
| Pitch variability  | 56.564 | 29.449 | 47.716 | 65.412 |
| Intensity variability | 0.655 | 0.142 | 0.612 | 0.698 |

*Descriptive Statistics of Automated Measures: Study 3*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *M* | *SD* | 95% *CI* |
|  | Low | Upper |
| A. Temporal quantity |  |  |  |  |
| Articulation rate | 4.117 | 0.621 | 3.930 | 4.304 |
| Pause ratio | 0.426 | 0.159 | 0.378 | 0.474 |
| B. Phonological quality |  |  |  |  |
| Maximum posterior probabilities | 0.771 | 0.029 | 0.762 | 0.779 |
| Posterior gaps to natives | 0.075 | 0.022 | 0.069 | 0.082 |
| C. Prosodic quality |  |  |  |  |
| Pitch variability  | 43.040 | 18.482 | 37.487 | 48.592 |
| Intensity variability | 0.591 | 0.119 | 0.556 | 0.628 |

**Supporting Information-D: Oral Interview Materials**

Describe the hardest and toughest challenge in your life.

**Your story should start with the following words:**

**One of the hardest/toughest challenges in my life was \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* Discussion points
* When? How old and where were you?
* Why did you encounter this challenge?
* Why was it so challenging?
* Did anybody (e.g., friends, parents) help you?
* Rounding off questions
* What did you learn from this experience?
* Would you like to go through the same experience again?