Visualisation tools to analyse phonetic confusions for speech perception tests

Chung Ting Justine Hui[1], *Catherine Watson*[2], *Takayuki Arai*[1]

[1] Sophia University, Japan
[2] University of Auckland, New Zealand

justinehui@eagle.sophia.ac.jp, c.watson@auckland.ac.nz, arai@sophia.ac.jp

Abstract

To enhance the intelligibility of synthetic voices, especially for the hard-of-hearing, these voices need to be evaluated in a phonetically systematic way. This enables identification of problematic sounds and confusions that come with them, thus allowing suitable adjustments to the voice. Firstly, modified forms of confusion matrices are introduced to examine and compare phonetic confusions. This is followed by the consonant confusion cloud plots, developed to visualise confusion data by focusing on individual consonants. These tools enable us to identify the distribution and relationship of confusion with the target phone in terms of manner, place and voicing in one glance.

Index Terms: speech perception, phonetic confusion, speech synthesis, confusion matrix, consonant confusion cloud

1. Introduction

Synthetic voices need constant evaluation to ensure they are being perceived accurately. This gives rise to different types of perception tests, especially designed to evaluate the intelligibility of these voices. One of the more utilised tests is the semantically unpredictable sentences, applied in the Blizzard challenges, one of the international recognised benchmark test that invites researchers to present synthetic voices annually and be judged by a public perception survey [1, 2, 3, 4]. While semantically unpredictable sentence perception tests are able to tell us the intelligibility of the voices on an utterance level, it lacks the granularity in focusing on individual phones to allow for possible enhancement on the voices on a phonemic level.

Wolter, on her studies to access synthetic voices on healthcare robots, took a different approach and evaluated the voices using unfamiliar medication names [5, 6]. From the answers produced by the participants, she then analysed the phonetic error in their perceived answers. Taking inspiration from this, a similar perception test has been carried out previously using pseudo medicinal names to test phonemic intelligibility of synthetic voices collected from 160 participants [7].

Part of the test examined 38 English consonants and consonant clusters in non word-final positions as shown in Table 1. Having this granularity to examine how the consonants produced by the synthetic voices are being perceived, we can then make suitable enhancements to the voices to increase their intelligibility. This gives rise to the need for tools for examining the errors the participants are making as well as the relationship of these errors with the target phone.

p pl pɹ t tɹ k kl kɹ b bl bɹ d dɹ g gl gɹ
f fl fɹ θ θɹ s ʃ sm sn st stɹ h v z ʧ ʤ m n w j l ɹ

Table 1: *List of non-word final consonants*

For a start, the confusion matrices introduced by Miller and Nicely in 1955 act as an adequate tool to examine all the consonants and the confusion in the participants' perception [8]. However, when we add in multiple variables, for example, to compare between two different synthetic voices, or two groups of participants, the traditional confusion matrix may not be able to handle the multitudinal nature of the data.

This paper describes such visualisation tools to enable researchers to analyse the perception results; from modifying the traditional confusion matrices to applying the concept of tag clouds to display the phonetic data. We will be using data obtained from a previous perception study [7] and discuss briefly observations made from the plots to highlight the functions of these tools.

2. Background data

While this paper does not focus on how the data was collected, this section will briefly introduce the data to be analysed by the visualisation tools in following sections [7].

In a previous study, perception data were collected from 160 participants to investigate the impact of hearing loss on the intelligibility perception of English consonants. This paper uses the data from one of the intelligibility tests where participants were asked to spell out the pseudo medicinal words as they hear them. 38 consonant and consonant clusters were tested in a series of sentences in the form of "At 9:30am, please take your [medication]", where medication is taken from a list of made-up medicine names consisting of the consonants listed in Table 1.

The participants were separated into groups according to their attributes such as first language English speakers and second language English speakers, hearing impaired (HI) and normal hearing (NH) and participants who are above the age of 60 and under the age of 60. This paper uses the data collected from 160 participants, dividing the participants into those who experience hearing loss (89) and those who do not (71).

Three voices were used to pronounced the stimulus sentences, two synthetic voices and one natural voices. For the purpose of this study, data from all three voices were combined.

2.1. Correct identification analysis

Before we launch into discussing the confusion participants make when they could not identify the consonants correctly, let us have a look at the correct identification rate for each of the consonants in Table 1. Figure 1 shows the list of consonants and their respective accuracy rate in terms of the hearing impaired and the normal hearing participants. We can see that from this simple text plot, how well each consonant is being recognised, and those they are more difficult to understand. Unsurprising,

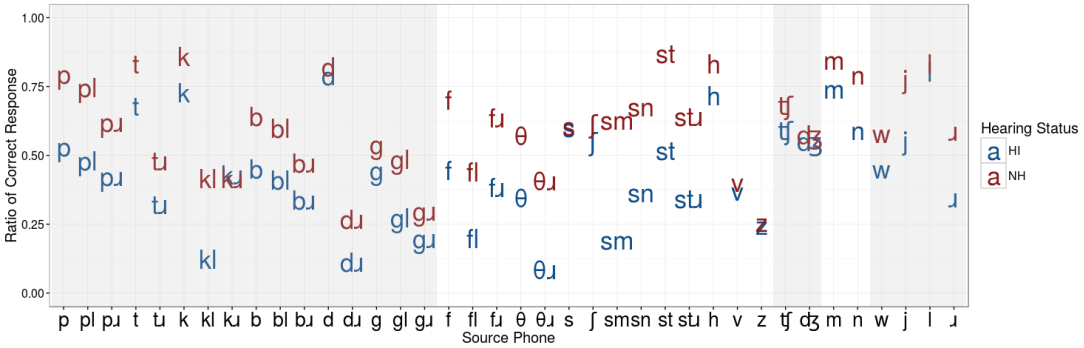

Figure 1: *Consonant identification plot for non word-final consonants*

hearing impaired participants could identify all the consonants less correctly than their normal hearing counterparts. Consonant clusters seem to especially cause distress, where less than 20% of hearing impaired participants could not identify /kl/, /dɹ/ and /θɹ/. On the other hand there are consonants such as /kɹ, /v/ and /z/ where they are similarly difficult for both groups.

3. Visualisation tools to aid analysis

3.1. Aim

While Figure 1 can offer us an insight into what consonants are difficult to perceive for the hearing impaired participants compared to their normal hearing counterparts, we need to be able to identify what the sound they confuse the target phone with and the phonetic relationship between their mistakes and the source phones to ultimately be able to enhance the synthetic voices. For example, if we find that voiceless stops are being mistaken for voiced stops, we might be able to manipulate the voice onset time or the length of the following vowel to make the consonant easier to identify.

This brings us to the need for a set of tools for developers to evaluate the synthetic voice in order to enhance speech signal, especially for the hearing impaired to understand. The following sections will describe these visualisation tools that have been used to display the errors the participants have made in a more systematic way from a phonetic perceptive. Due to restriction on space, the consonant clusters are left out in the plots. For details on the consonant clusters, please see [7].

All figures are generated using R [9], and the visualisation tools were developed from scratch with the use of packages wordcloud [10] and ggplot2 [11].

3.2. Confusion matrices

Modified from the traditional confusion matrices as seen in Miller (1955), Pollack (1960) and Singh (1966) to name a few [8, 12, 13], the confusion matrices in this paper are overlaid with a heat plot where the opaqueness of the tiles represent the number in ratio of participants perceiving a particular phone and the colours represent the two hearing status groups.

The confusion matrices combines the correct identification rate of the consonants and the consonant choices for any particular phones into a bird's eye view of the phoneme error analysis for the intelligibility tests.

Figure 2 and Figure 3 allow us to examine all the consonants identified and confused by the participants, separating the data hearing impaired (HI) and normal hearing (NH) groups. The x-axis of the confusion matrix represents the targeted phones and the y-axis lists out the answers given by the participants. Two arbitrary symbols were chosen to represent vowels, a 'v' enclosed in a circle, and invalid answers, a dash across a circle. Due to the size of the tiles, the ratio has been rounded to 1 decimal point. This means that a '0' tile indicates less than 5% of answers were identified as the specific phone, as opposed to an empty tile, where there were no answer.

Diagonally we can see the a relatively strong response where the participants are identifying the source phone correctly. By combining the numbers and the opaqueness of the tiles, we are then able to identify the problematic phones, and their confusion from the columns.

We can see that while there are some strong candidates identified 80% and above correctly by the hearing impaired participants such as /d/ and /l/, most of the time the diagonal tiles are faint in their colour, indicating a low correct identification rate. We can also observe some clustering of confusions around the sonarants, showing that the confusions for sonorants occur within amongst themselves. On the other hand, the diagonal tiles for the normal hearing results are much more opaque than the hearing impaired results, being in consistent with out results from Figure 1. However, for both groups, it seems that /z/ is a difficult sound to identify in this particular test, with more than 50% of the answers being confused with /s/ instead.

3.3. Combined confusion matrices

Now, what happens if we want to compare the two groups together to show how the differences and similarities in the mistakes the participants make? In the combined confusion matrices, we make use of the colour hue and again opaqueness of the tiles to show both groups on the same confusion matrix, as shown in Figure 4, where again red represents the normal hearing results and blue the hearing impaired results.

A tile that has a purple hue describes both groups having similar levels of difficulty in identifying the phone, a red hue indicates the normal hearing has a stronger presence and a blue hue for the hearing impaired group. From Figure 4, we can approximate the normal hearing having a larger varieties in their confusion, where the tiles away from the expected diagonal line display more 'red' than 'blue'. On the other hand, the hearing impaired group has a higher ratio of invalid answers, shown by the more 'blue' hue along the last row of the matrix.

However, while we can observe all the tested phones in one glance, it is difficult to observe the relationship and distribution between each individual consonant and their confusions. This brings us to the next type of plots, the consonant confusion clouds, where we can focus on individual consonant at a time.

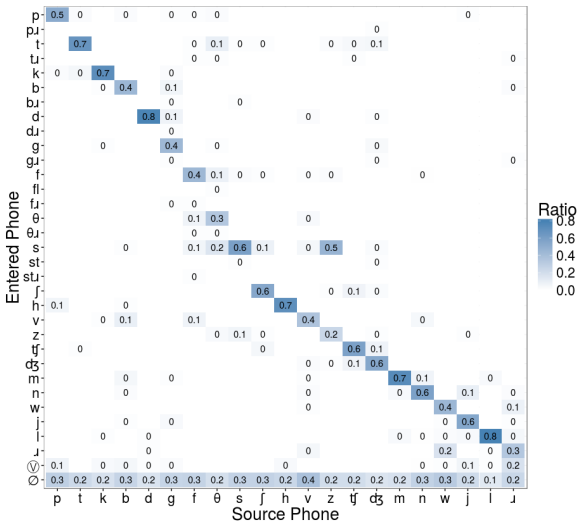

Figure 2: *Confusion Matrix for hearing impaired participants*

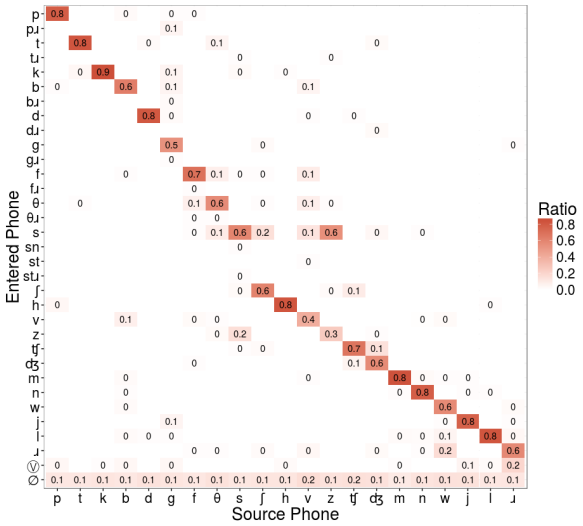

Figure 3: *Confusion Matrix for normal hearing participants*

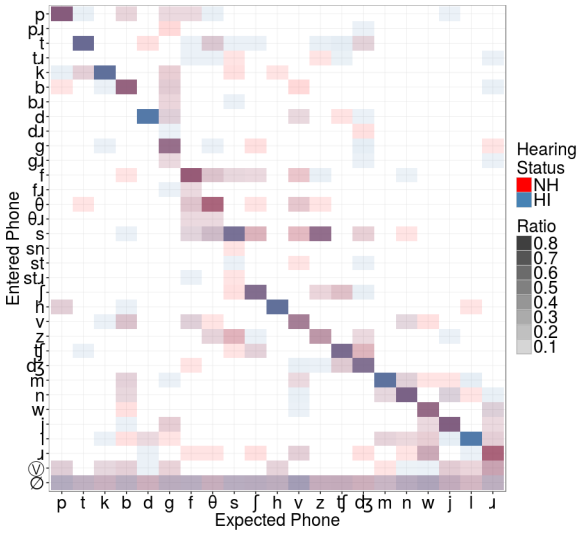

Figure 4: *Confusion Matrix displaying results from both hearing impaired and normal hearing participants*

3.4. Consonant confusion clouds

The consonant cloud plots take inspiration from tag clouds, which are typically used to depict metadata or tags in websites where words are mapped onto a cloud of words with their importance or occurring frequency signified by font size [14]. In the same way, the frequency in which the sounds are identified by the participants governs the font size of phones in the consonant cloud plots.

Unlike the consonant identification plots or the confusion matrices, the cloud plots are not able to indicate the absolute value of how accurately a phoneme is being perceived. Instead, it gives a graphical representation of all the sounds heard mistakenly in an arbitrary consonant grid where the x-axis represents the location of articulation and the y-axis represents the manner of articulation and voiced and voiceless pairs are presented side by side as shown in Figure 5. By having an arbitrary grid reliant on the three characteristics of a consonant: the manner, location and voicing, we may be able to identify a pattern in which the sounds are being confused with. The code to generate these cloud plots has been written from scratch with the help of the R package wordcloud [11].

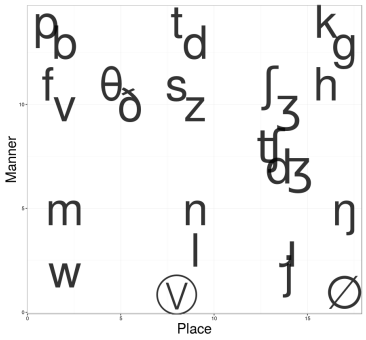

Figure 5: *How the consonants are placed according to their manner and place of articulation on the cloud plots*

Using colours of the font, the cloud plots are able to compare multiple groups, such as how two or more synthetic voices compared to each other, or how different groups of participants compare in their perception for individual consonants.

The font size is determined by an arbitrary inverse square-root relationship with the number of participant to make the less frequent consonants visible. Due to the nature of clutter of the cloud plots, the current set up makes it difficult to include the consonant clusters. International Phonetic Alphabet symbols are used, and again the two symbols used for vowels and invalid answers as in the confusion matrices are applied here.

Let us take a look at Figure 6, showing the answers from the hearing impaired and the normal hearing group perceived for stop consonants. We can see that for /t/, /d/ and /k/, the majority for both groups could identify the correct consonant fairly accurately by the large font of the letters, with little confusions. When we compare the velar stop pair, /g/ seems to trigger many more variation of other phones than /k/ for both groups of participants. While both groups seem to confuse the /g/ sound with /b/ similarly, there is a tendency for the normal hearing to confuse /g/ with /k/, and the hearing impaired participants to confuse /d/ with /g/ than their counterparts.

For all the stops apart from /t/, we can observe the hearing impaired participants producing more invalid answers than their normal hearing counterparts from their sizes of the ”invalid” symbol, suggesting that the hearing impaired had trouble

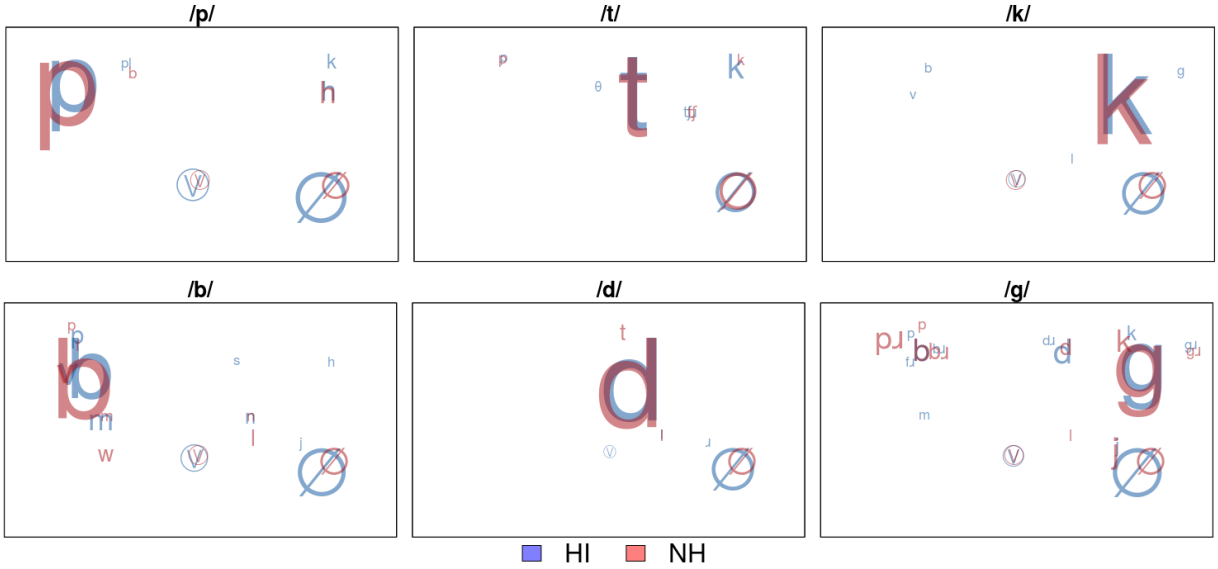

Figure 6: *Consonant identification cloud plots of stops*

hearing the phone at all to take a guess at the question.

We can see from these figures that most confusions only differ in one aspect from the target consonant, that is, the mistake occurs either in identifying the consonant's manner, place of articulation or voicing. For example, we can see in /b/ that the confusions tend to be manner of articulation, with mistaken sounds being /m/, /w/, /v/. On the other hand, /g/ has more places of articulation confusion, such as /d/ and /b/, and the insertion of /ɹ/. Both consonants also have a small percentage of voicing confusion with the consonant being mistaken as their unvoiced counterparts, /p/ and /k/.

These cloud plots allow us to be able to analyse the confusion data with a greater focus than the confusion matrices, helping us to locate the problematic sounds and their confusions leading to possible enhancement for the voices.

4. Conclusions and other applications

Using the data gathered from a previous study, we are able to locate the problematic phones and mistakes participants are making using the confusion matrices and the consonant confusion clouds. These visualisation tools give us a greater level of granularity to examine the phones that the participants may have trouble deciphering. On top of that, the cloud plots focus on each individual phones, allowing us to observe the confusion using phonetic knowledge. Preliminary observation shows that most errors only differ in one aspect of the target consonant, be it manner, place or voicing.

Finally, while these tools were designed to evaluate synthetic voices, the use of these tools does not need to be limited to synthetic voices and can be applied to other perception tests when examining consonant confusions.

5. Acknowledgements

We would like to thank our participants and Triton Hearing for their support in recruitment for this study.

6. References

[1] A. W. Black and K. Tokuda, "The Blizzard Challenge — 2005: Evaluating corpus-based speech synthesis on common datasets," *Interspeech 2005: 6th Annual Conference of the International Speech Communication Association*, pp. 77–80, 2005.

[2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[3] Z.-h. Ling, Y.-j. Wu, Y.-p. Wang, L. Qin, and R.-h. Wang, "USTC System for Blizzard Challenge 2006 an Improved HMM-based Speech Synthesis Method," *Blizzard Challenge Workshop*, 2006.

[4] H. Zen, T. Toda, and M. Nakamura, "Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005," *IEICE transactions on information and systems*, vol. 90, no. 1, pp. 325–333, 2007.

[5] M. Wolters, P. Campbell, C. Deplacido, A. Liddell, D. Owens, and A. Division, "Making Speech Synthesis More Accessible to Older People," in *Sixth ISCA Workshop on Speech Synthesis*, 2007.

[6] M. Wolters, P. Campbell, C. DePlacido, A. Liddell, and D. Owens, "The Effect of Hearing Loss on the Intelligibility of Synthetic Speech," *International Congress of Phonetic Sciences*, vol. 16, no. August, pp. 673–676, 2007.

[7] C. T. J. Hui, "Development and Implementation of a Perception Toolkit to Evaluate the Impact of Synthetic Speech on the Hearing Impaired," Master Thesis, University of Auckland, 2016.

[8] G. A. Miller and P. E. Nicely, "An Analysis of Perceptual Confusions Among Some English Consonants," *Journal of the Acoustical Society of America*, vol. 27, no. 3, pp. 338–352, 1955.

[9] R Core Team, "R: A language and environment for statistical computer," 2015. [Online]. Available: https://www.r-project.org/

[10] I. Fellows, "wordcloud: Word Clouds," 2014. [Online]. Available: http://cran.r-project.org/package=wordcloud

[11] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. [Online]. Available: http://had.co.nz/ggplot2/book

[12] I. Pollack and L. Decker, "Consonant Confusions and the Constant Ratio Rule," *Language and Speech*, vol. 3, pp. 1–6, 1960.

[13] S. Singh and J. W. Black, "Study of Twenty-Six Intervocalic Consonants as Spoken and Recognized by Four Language Groups," *Journal of the Acoustical Society of America*, vol. 39, no. 2, pp. 372–387, 1966.

[14] M. Halvey and M. T. Keane, "An Assessment of Tag Presentation Techniques," 2007. [Online]. Available: http://www2007.org/htmlposters/poster988/