HU UNIVERSITY OF APPLIED SCIENCES UTRECHT
INSTITUTE OF INFORMATION & COMMUNICATION TECHNOLOGY

Finding Differences Between Slips of the Tongue in a Conversation of Individuals with Aphasia and or Apraxia of Speech and Non-Brain-Damaged Individuals Using Kaldi ASR

By:
Gaynora van Dommelen (1778287)

Supervisor:
R. Ossewaarde (roelant.ossewaarde@hu.nl)
H. Aldewereld (huib.aldewereld@hu.nl)

A RESEARCH REPORT TO SHOW THE STEPS TAKEN DURING THE SEMESTER AS WELL AS THE RESULTS AND CONCLUSIONS FOR EACH OF THE DIFFERENT APPROACHES USED FOR THIS RESEARCH PROJECT.

RESEARCH SEMESTER
HBO-ICT ARTIFICIAL INTELLIGENCE, UTRECHT

December 2022

ABSTRACT

Neurodegenerative diseases could possibly be diagnosed earlier by getting more insight into the speech errors made by people with these conditions. This study aims to prove whether the Kaldi ASR toolkit could be used to recognize these speech errors. The expectation is that the acoustic and language model probability scores for respectively the phone and word transcriptions, can be used to determine speech errors in audio. Because of the setup framework of Kaldi, retrieving the acoustic and language model probabilities was deemed out of the scope of Kaldi and instead Kaldi's own confidence scores were used. From the results of the time aligned transcription confidence scores, the scores always returned 1. This leads to the conclusion that for this dataset and model implementation, the confidence scores are not usable for determining speech errors, mainly due to the limited available data.

Keywords: Automatic Speech Recognition (ASR), Aphasia, Apraxia of Speech (AoS), Kaldi ASR

TABLE OF CONTENTS

INTRODUCTION

The prevalence of aphasia in the Netherlands is around 30,000. And every year approximately 10,000 people acquire this condition after a stroke (El Hachioui, 2012; Fierens, 2019).

Although there has been no consensus on the exact definition of aphasia, one of the most current definitions of aphasia as defined by American Speech-Language-Hearing Association (n.d.) is "Aphasia is an acquired neurogenic language disorder resulting from an injury to the brain—most typically, the left hemisphere." The primary areas impaired by aphasia are: spoken language expression, spoken language comprehension, written expression and reading comprehension (American Speech-Language-Hearing Association, n.d.).

The current way for diagnosing aphasia and speech disorders such as apraxia of speech (AoS) is done through different language tests carried out by speech-language pathologists and linguists. Following the guidelines made by Berns et al. (2015, p. 24-26) for diagnosing aphasia in the Netherlands, diagnosing aphasia is done by applying different language tests such as the Semantic Association Test (SAT; Visch-Brink, Stronks, & Denes, 2005) and the Comprehensive Aphasia Test (CAT-NL; Braak & Heethuis, 2014). These tests combine the characteristics of aphasic speech and test these characteristics (mostly) independently to determine the presence of aphasia as well as the severity of the condition. However, the assessment solely relies on the subjective evaluation of the practitioner who must meet the utmost requirements on not only the clinical knowledge of aphasia but must also have the corresponding linguistic background. To objectively assess the speech of a patient, automating these speech analysis could lead to faster, more accessible and objective diagnosis.

With the upcoming techniques in speech analysis, various studies have researched the use of automatic speech recognition (ASR) for extracting various speech features and using these to recognise aphasia. For instance, Qin, Lee, Kong, and Law (2016) found that the ASR output can be used for determining speech and language impairments for the Cantonese language. This study concentrated on duration-related features, such as: number of pauses, number of speech chunks, number of syllables etc. Difficulties when transcribing emerged when encountering speech errors, which resulted in higher scores. Another study done by Le, Licata, and Provost (2018) on English

Aphasic speech, used both information density as mentioned by Qin et al. (2016) as well as dysfluency, lexicon diversity etc. as features for determining aphasic speech. They found that these could help in determining the severity of aphasia.

Where previously done studies primarily focus on extracting multiple speech characteristics of aphasic speech at the same time, this study forms a substudy of recognising individuals with aphasia and rather focuses on finding speech error features.

Slips of the tongue, also known as speech errors, are very normal in everyday conversation, but are especially apparent in individuals with aphasia (Ogar, Slama, Dronkers, Amici, & Gorno-Tempini, 2005). Subdividing speech errors to different categories has also been a big cause for discussion since this largely depends on what the cause for a speech error could be. To study this, numerous researches have created large datasets of speech errors (Fromkin, 1971; Garrett, 1975). In this paper, the categorization of (Carroll, 1986, p. 195) of the 8 most basic found speech errors wil be followed.

That there are multiple ways to differentiate between the types of errors that are made, is later proven by Fromkin (1971). Fromkin (1971) took the linguistic approach when looking at the same errors since "[...] particular errors shed light on the underlying units of linguistic performance, and on the production of speech" (p. 29). She goes on to show that most of the speech errors in the collected data substituted to the type of speech errors in the categorization of Carroll (1986) can be divided into phone, lexicon and morpheme speech errors. But, the most frequent speech errors found in the data are segments of the size of a phone (Fromkin, 1971, p. 30).

This study focuses on recognizing the speech error of type blend where a mix of two closely related words are uttered from the categorization of Carroll (1986) such as 'to be spanked/paddled' $\longrightarrow$ 'to be spaddled' and the wrong use of consonants since people with neurogenic diseases tend to make errors with consonants rather than vowels (Jonkers, Feiken, & Stuive, 2017) by looking at the probability scores of the transcriped phones.

In this paper first a short introduction will be given on the most used methods for creating an ASR model, followed by the used setup for creating the ASR model on the digits 0 to 9. Then, a description of the steps that were to be taken for determining the usefullness of probability scores in recognizing speech errors is explained. Following the methodology and design of the study the results will be shown and explained. To conclude the study a conclusion on the use of probability scores for recognizing speech errors will be given as

well as ideas for future research.

METHODS

To familiarize the reader with the general knowledge of building a simple Automatic Speech Recognition model, a general setup as well as commonly used tools are explained. Thereafter, the build of the created ASR model for this study is described followed by the plan that would be used for setting up a baseline in recognizing speech errors using the made ASR model.

2.1 Automatic Speech Recognition

Jurafsky and Martin (2008) describe that a simple ASR model at least consists of a Feature extraction phase followed by a Language and Acoustic model which input will be translated using a Decoder.

Feature Extraction Also known as signal processing state. Samples the acoustic waveform into frames. These frames are then transformated into spectral features which are represented by vectors.

Acoustic Modelling Also known as the phone recognition phase. Given a set of linguistic units (ngrams, phones, etc.) computes the likelihood of an observed spectral feature.

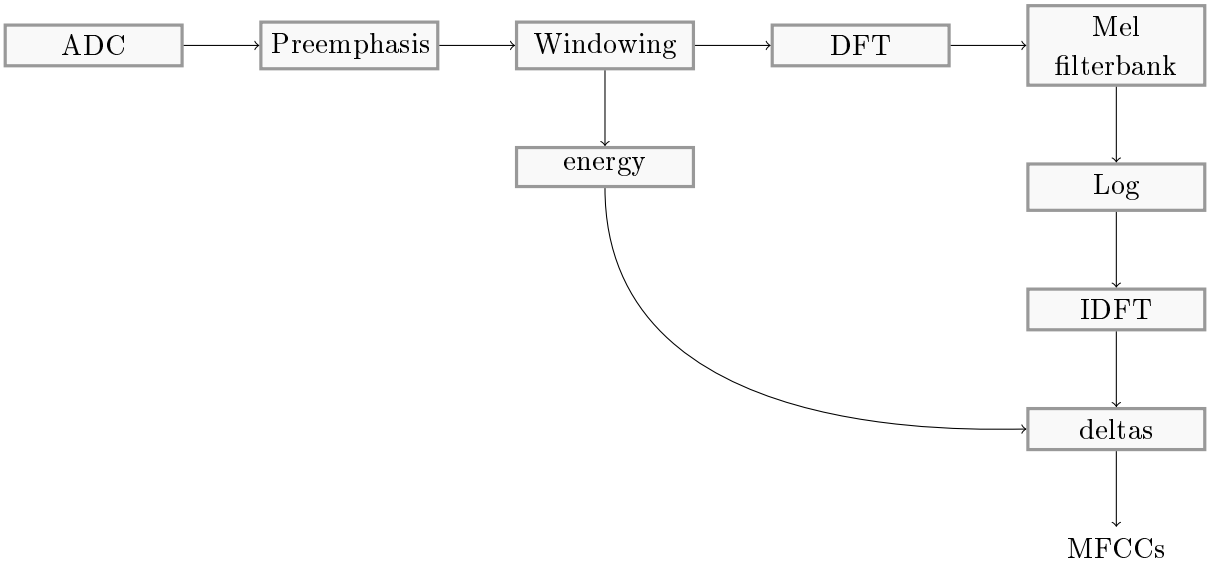
Language Modelling Given an input string s and language L, expresses the likelihood that s is a sentence of the language L.

Decoding Outputs the most likely sequence of words using the Acoustic Model combined with the Language Model.

2.1.1 Feature Extraction

For this study one of the most well-known feature extraction methods for speech audio was used: mel frequency cepstral coefficients (MFCCs), shown in 2.1 (Sharma, Umapathy, & Krishnan, 2020; Alim & Rashid, 2018). The extraction of MFCC features uses the Mel filterbank to model the unequal sensitivity of the human hearing for different frequencies.

Figure 2.1: The workflow of extracting MFCC feature vectors from a speech signal. (Jurafsky & Martin, 2008, p. 330)



Quantization

Quantization or analog-to-digital conversion, A/D conversion in short, forms the first step in the feature extraction process. A/D conversion compresses the waveform while maintaining the most characteristic features of the audio signal, effectively diminishing the computational complexity. This is done by sampling a signal by measuring its amplitude at disrete time intervals. Then the waveform is quantized by representing each real-valued number as an integer. A quantized waveform is often referred to as

$$x[n]$$

, where n is an index over time (Jurafsky & Martin, 2008, p.329).

Pre-emphasis

Pre-emphasis filters are ways to boost higher modulating frequencies. Most often a high-pass filter (HPF) is used to boost high frequency energy. Increasing the high frequency is done because there is more energy in the lower frequencies in the spectrum for voiced segments like vowels. Boosting the

energy in these lower frequencies makes it easier for the acoustic model to retain information from the formants of phones which in turn improves the accuracy of phone detection (Jurafsky & Martin, 2008, p. 330).

Windowing

A speech audio signal is not constant over time: non-stationary. This means that taking the spectrum, for example, of a full audio signal would mean laying frequencies of different timestamps on top of eachother. This would make the spectrum so generalized that the original signals' characteristics would not be recognisable. Therefore, to make the signal closely resemble a constant or stationary signal, small slightly overlapping windows of the original signal are retrieved. Then, for each window a frame is extracted from the spectral features. The extraction is done by multiplying the value of a signal at time n: $s[n]$, by the value of the window at time n: $w[n]$, such that the extracted signal is now:

$$y[n] = w[n]s[n]$$

However, because cutting the signal by rectangular windows creates cut-off edges, abruptly stopping a signal, other forms of taking windows were made, such as the Hamming and Hanning window (Jurafsky & Martin, 2008; Taylor, 2009).

Fast Fourier Transformation

Extracting spectral information from the windowed signal $w[n]$ is done by retaining the amount of energy contained within the signal at various frequency bands. To extract this information for a sampled signal is called fast fourier transformation (FFT). The FFT takes a sampled signal and returns for each of the windowed samples in $w[n]$ a complex number $X[k]$ which represents the magnitude and phase of the frequency.

Mel Filter Bank

Though the FFT returns the amount of energy for each of the frequency bands, it does not resemble human hearing. Human hearing is less sensitive when it comes to the higher frequencies: above 1000 Hz. This is important,

because modelling human hearing improves the speech recognition performance (Jurafsky & Martin, 2008; Taylor, 2009). A mel is a unit of speech where paired sounds perceptually equidistant are seperated by an equal number of mels. Following the human hearing the mel scale is linear below 1000 Hz and logarithmic above 1000 Hz. Using the raw acoustic frequency, the mel frequency m can be derived by:

$$mel(f) = 1127ln(1 + \frac{f}{700})$$

(Jurafsky & Martin, 2008, p. 333).

Inverse Fast Fourier Transformation

The name cepstrum reverses the first part of the word spectrum. The cepstrum is retrieved by taking the inverse FFT of the log magnitude of the FFT of a signal, such that the cepstrum for a windowed frame of speech $x[n]$ is:

$$c[n] = \sum_{n=0}^{N-1} log(|\sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn}|)e^{-j\frac{2\pi}{N}kn}$$

The cepstrum seperates the source: fundamental frequency, and filter: position of the vocal tract of a signal. The filter returns the shape of the vocal tract, which helps understanding which phone was produced; improving phone recognition. Commonly used are the first 12 cepstral features solely representing the vocal tract filter (Jurafsky & Martin, 2008, p. 334-335; Taylor, 2009, p. 364-365).

Deltas

Another feature usually added to the feature vector besides the 12 cepstral coefficients of a frame is the energy retrieved from said frame. The energy of a frame highly correlates with the identity of a phone and thus helps in detecting phones. This is calculated by taking the sum of the power of the samples in the frame over time, such that the energy of a signal x in a window between time sample t_1 and t_2 is given by:

$$Energy = \sum_{t=t_1}^{t_2} x^2[t]$$

(Jurafsky & Martin, 2008, p. 336)

2.1.2 Acoustic Model

Modelling the acoustic likelihood for the feature vectors retrieved during feature extraction is the next step in creating an ASR model. For this study the acoustic model consists of a Hidden Markov Model where each states' acoustic likelihood is computed using a Gaussian Mixture Model: HMM-GMM, because the size of the dataset as well as the corpus makes using more complex models such as neural networks unnecessary. Therefore, for the explanation of the general workflow of an acoustic model will be fully focused on the HMM-GMM.

HMM

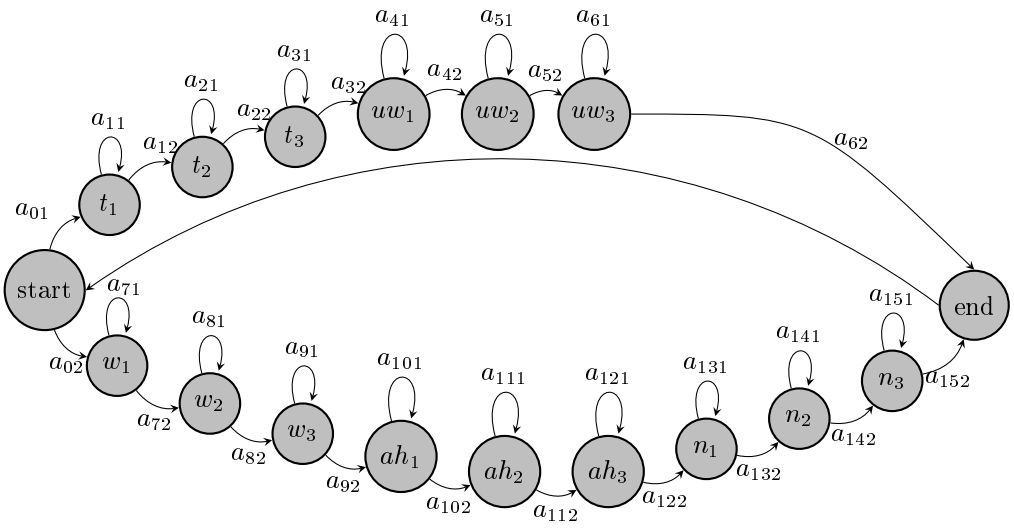
The most commonly used acoustic model is the left-to-right HMM, that only allows for state transitions to itself (self-loop) or to a single succeeding state (Jurafsky & Martin, 2008, p. 327). It models the likelihood of succeeding language units, most often phones and consists of the following components:

- a set of states: $Q = \{q_1, ..., q_N\}$
- a transition probabilty matrix $A = [a_1 \ ... \ a_n]$, where $a_i = \begin{bmatrix} v_{1i} \\ ... \\ v_{mi} \end{bmatrix}$
- a set of observations: $O = \{o_1, .., o_N\}$, retrieved from a vocabulary $V = \{v_1, ..., v_V\}$
- a set of observation likelihoods (emission probabilities): $B = b_i(o_t)$. Each element of this set forms the probability that observation o_t following a state i.
- a start: q_0 and end: q_{end} state

Here the transition probabilites: A along with the states: Q forms the lexicon library.

Because phones are not homogeneous: can last multiple frames and are not acoustically identical, phones usually each have three emitting states (2.2) (Jurafsky & Martin, 2008, p. 327-328).:

Figure 2.2: HMM model for the words "one [w ah n]" and "two [t uw]".



GMM

To capture the wide variability within a speech signal the observation probabilities are directly computed on the feature vectors' real-valued, continuous input. This form of computation over time is also known as a probability density function or pdf in short. Gaussian Mixture Models (GMM) are most oftenly used for computing acoustic likelihoods. A GMM consists of multiple multivariate Gaussians that in turn consist of multiple Gaussian distributions. Where a single Gaussian distribution assigns only one acoustic likelihood for a single cepstral feature, the multivariate Gaussians compute the acoustic likelihood of multiple feature vectors. But, to estimate the probability that a HMM state j generates the value of a feature vector, the assumption that the possible values of this observation feature vector o_t will always be normally distributed is too large an assumption. Therefore, a multitude of a weighted mixture of multivariate Gaussians is used to approach the non-normal distribution (Jurafsky & Martin, 2008, p. 340-346).

2.1.3 Language Model

A language model is also often called N-gram language model, because of the number of words used as input for the model. Large corpus data most

oftenly use trigrams or quadgrams, while small-vocabulary data mostly use bigrams or unigrams (Jurafsky & Martin, 2008, p. 348). Suppose a language model for the English language was made on trigrams, then the possibility of the sentence 'I am hungry' to be a part of that language is calculated by:
P(<s> i am hungry </s>) = P(i|<s> <s>)P(am|<s> i)
P(hungry|i am)P(</s>|am hungry).

2.1.4 Decoding

To compute the most probable string of words given a set of acoustic observations, a form of the Bayes' rule can be used. Here the best sequence of words is the one that maximizes the product of the language model prior and an acoustic likelihood:

$$\hat{W} = \underset{W \in \mathcal{L}}{argmax} \overbrace{P(O|W)}^{likelihood} \overbrace{P(W)}^{prior}$$

(Jurafsky & Martin, 2008, p. 349)

2.2 Speech Error Recognition Design

For a basic baseline model on proving the usefullness of acoustic model (AM) probabilities on recognising speech errors on phone level would include assigning the acoustic probability to each transcribed phone. By classifying each phone as an error (0) or no error (1) the distinction between speech errors and normal speech can be made on phone level. Then, by choosing a threshold for the acoustic probability to cut off speech errors from normal speech, the accuraccy can be computed.

To compare the speech error recognition made on phone versus word level another basic baseline model should be made using the language model (LM) probabilities. The steps would be the same. Each word gets assigned a 0 in case of a speech error and a 1 for normal speech. Then, a threshold should be chosen to distinguish between speech errors and normal speech and the accuracy for the baseline model can be computed.

RESULTS

To evaluate the ASR model, the Character, Word and Sentence Error Rate, CER, WER and SER respectively, are used. The CER is given by:

$$WER = \frac{Insertions + Substitutions + Deletions}{Total\ \#\ of\ Characters\ in\ Correct\ Transcript} \cdot 100\%$$

Here an insertion is seen as an extra character appearing in the transcription where it does not appear in the actual text. A substitution is the act of transcribing a different character from the original. And, contrary to insertion, deletion is the act of not transcribing a certain character. The WER is defined as:

$$WER = \frac{Insertions + Substitutions + Deletions}{Total\ Words\ in\ Correct\ Transcript} \cdot 100\%$$

The WER score shows the similarity between the transcribed string and the actual string. An insertion is counted when an extra word was transcribed where it, according to the original alignment, should not be. Contrary to insertion, deletion does not transcribe words where in the original alignment a word does exist. Substitution is the act of transcribing the wrong word. The SER score on the other hand, shows how many sentences from the test set had at least one error and is given by:

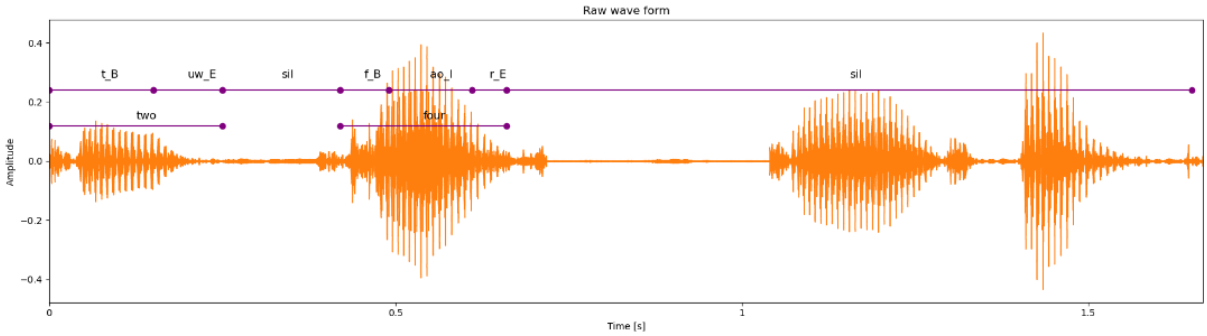
$$SER = \frac{\#\ of\ sentences\ with\ at\ least\ one\ word\ error}{total\ \#\ of\ sentences}$$

Table 3.1 shows the performance scores for the monophone and first triphone pass training. The TRI1 training retrieves better scores on both the WER and CER. However, looking at the SER scores shows that with each transcription there are mistakes.

Table 3.1:
Comparison of MONO and TRI1 training in CER, WER and SER scores

	CER	WER	SER
MONO	9.59	10.55	100
TRI1	8.94	9.28	100

Figure 3.1: Shows the time aligned transcription, made by the ASR of the sentence “3 4 5”.



This becomes clearer, by looking at the raw waveform and time aligned transcriptions as in figure 3.1. Immediately noticeable is the cutoff sound at the beginning, which is also found in the other test files. Expectations are that this is caused by the cutting of a large audio file containing all the digits for the test speakers based on a constant time range. This was done to follow along with the instructions left by the makers of the FSDD dataset to add data. This causes some digits to be abruptly cutoff causing the ASR to not properly recognise these sounds. The first time alignment shows the transcriped phones found by the ASR, while the second alignment below that of the phones, shows the trasncribed words by the ASR. From the phone transcription we can also see the beginning of a state denoted by ”_B”, the transitioning states following the starting state denoted with ”_I” and the ending of a word found, denoted using ”_E”. For this file the word alignment is:

Actual:	THREE	four	FIVE
ASR:	TWO	four	****
Error:	S		D

Which gives

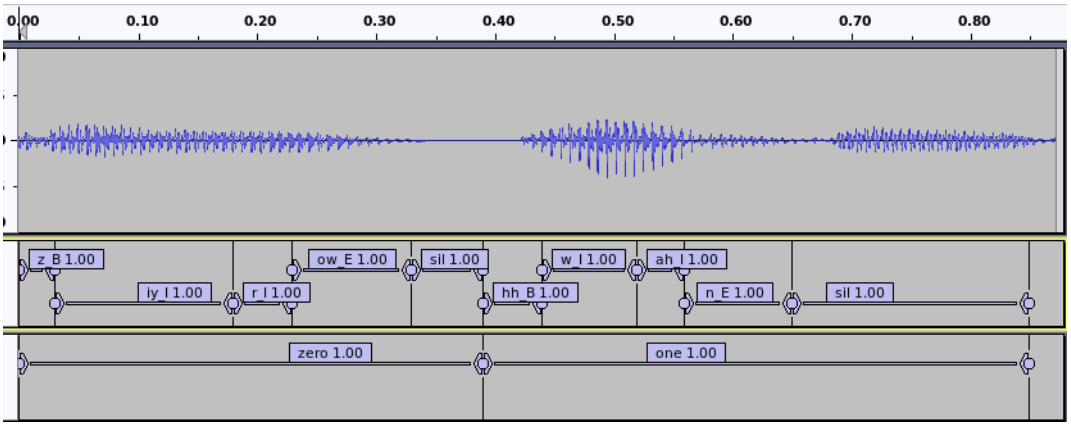
$$WER = \frac{0+1+1}{3} \cdot 100\% \approx 66.7\%$$

for this file.

Retrieving the acoustic and language model scores for each transcribed phone, was not deemed possible for the current setup, because the Kaldi framework does not directly store this information.

Otherwise obtained confidence scores calculated using the lattice depth, word-categories, unigrams and the number of phones in a word (*Fwd: Confidence calibration, observations,*, n.d.) where then used for the time aligned phone and word transcriptions to determine the certainty of transcription. Figure 3.2 made using Audacity shows the audio with directly underneath the raw wave form of the audio the time alignment of the phones found followed by the confidence score. Then, underneath the phone time aligned transcription is the word aligned transcription followed by the confidence score. From analysis on the test files, it was found that these always returned 1. These scores were retrieved by aligning the phone lattices, retrieving the best scoring paths and then calculating the confidence of time transcribed phones. The reason for the confidence scores to always return 1 is due to the lack of competing alternative paths.

Figure 3.2: Shows the time aligment of the transcribed phones and words for the TRI1 training with their confidence scores.



CONCLUSION AND FUTURE WORK

This paper explores the approach for recognizing the speech errors of type blend and the wrong use of consonants by retrieving the confidence scores of transcribed phones and words.

Though the original intention was to use the actual acoustic and language model scores frame by frame of transcribed phones or words, this approach was later disregarded. The reason for this was that Kaldi doesn't save these scores, because frame by frame scores generally don't say much about the confidence of transcribed words or phones (*frame-by-frame acoustic scores in kaldi*, n.d.).

Because the dataset is very small, the confidence scores do not provide any useful information about how using these scores could help in recognizing speech errors. The small dataset causes lattices to only contain one path which causes the computed confidence scores to always return 1 since there are no competing optimal paths.

To prove the usefullness of the confidence scores in relation to recognizing speech errors a bigger dataset is needed. Another approach in proving the usefullness of the confidence scores includes using a pre-trained model on the Dutch language in Kaldi and retrieving the confidence scores on aphasic speech.

Bibliography

Alim, S. A., & Rashid, N. K. A. (2018). *Some commonly used speech feature extraction algorithms.* IntechOpen London, UK:.

American Speech-Language-Hearing Association. (n.d.). *Aphasia.* Retrieved from https://www.asha.org/practice-portal/clinical-topics/aphasia/

Berns, P., Jünger, N., Boxum, E., Nouwens, F., van de Staaij, M., van Wessel, S., ... CBO (2015). *Logopedische richtlijn 'diagnostiek en behandeling van afasie bij volwassenen'.* Woerden: Nederlandse Vereniging voor Logopedie en Foniatrie.

Braak, M., & Heethuis, A. (2014). *Cat-nl kennismaking met de comprehensive aphasia test - nederlandstalige bewerking white paper i* (A. Kooij, Ed.). Pearson Assessment and Information B.V.

Carroll, D. (1986). *Psychology of language.* Brooks/Cole Publishing Company. Retrieved from https://learnclax.com/utilities/view-doc.php?book-id=4472

El Hachioui, H. (2012). Aphasia after stroke: The speak study.

Fierens, L. (2019). Differentiatie bij personen met ernstige afasie. een onderzoek naar de betrouwbaarheid en klinische meerwaarde van globamix, een nieuw dynamisch assessment.

frame-by-frame acoustic scores in kaldi. (n.d.). Retrieved from https://groups.google.com/g/kaldi-help/c/Hu6dC0Mp00Q

Fromkin, V. (1971). The non-anomalous nature of anomalous utterances. *Language*, *47*(1), 27–52. Retrieved 2022-10-14, from http://www.jstor.org/stable/412187

Fwd: Confidence calibration, observations,. (n.d.). Retrieved from https://groups.google.com/g/kaldi-developers/c/FyTd2Wmgy-k

Garrett, M. (1975). The analysis of sentence production. In G. H. Bower

(Ed.), (Vol. 9, p. 133-177). Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/S0079742108602704 doi: https://doi.org/10.1016/S0079-7421(08)60270-4

Jonkers, R., Feiken, J., & Stuive, I. (2017). Diagnosing apraxia of speech on the basis of eight distinctive signs. *Canadian Journal of Speech-Language Pathology and Audiology*, *41*(3), 303–319.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing* (S. Russell & P. Norvig, Eds.). Pearson Prentice Hall.

Le, D., Licata, K., & Provost, E. M. (2018). Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, *100*, 1–12.

Ogar, J., Slama, H., Dronkers, N., Amici, S., & Gorno-Tempini, M. L. (2005). Apraxia of speech: An overview. *Neurocase*, *11*(6), 427-432. Retrieved from https://doi.org/10.1080/13554790500263529 (PMID: 16393756) doi: 10.1080/13554790500263529

Qin, Y., Lee, T., Kong, A. P. H., & Law, S. P. (2016). Towards automatic assessment of aphasia speech using automatic speech recognition techniques. In *2016 10th international symposium on chinese spoken language processing (iscslp)* (pp. 1–4).

Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, *158*, 107020.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Visch-Brink, E., Stronks, D., & Denes, G. (2005). *De semantische associatie test*. Harcourt Test Publishers. Retrieved from https://books.google.nl/books?id=XodZtwAACAAJ