UNIVERSITY OF BIRMINGHAM

THE EFFECTS OF CHILD LANGUAGE DEVELOPMENT ON THE PERFORMANCE OF AUTOMATIC SPEECH RECOGNITION

By

EVANGELIA FRINGI

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Electronic, Electrical and Systems Engineering
College of Engineering and Physical Sciences
University of Birmingham
September 2018

Abstract

In comparison to adults', children's ASR appears to be more challenging and yields inferior results. It has been suggested that for this issue to be addressed, linguistic understanding of children's speech development needs to be employed to either provide a solution or an explanation. The present work aims to explore the influence of phonological effects associated with language acquisition (PEALA) in children's ASR and investigate whether they can be detected in systematic patterns of ASR phone confusion errors or they can be evidenced in systematic patterns of acoustic feature structure. Findings from speech development research are used as the framework upon which a set of predictable error patterns is defined and guides the analysis of the experimental results reported. Several ASR experiments are conducted involving both children's and adults' speech. ASR phone confusion matrices are extracted and analysed according to a statistical significance test, proposed for the purposes of this work. A mathematical model is introduced to interpret the emerging results. Additionally, bottleneck features and i-vectors representing the acoustic features in one of the systems developed, are extracted and visualised using linear discriminant analysis (LDA). A qualitative analysis is conducted with reference to patterns that can be predicted through PEALA.

To yiayia Vaggelio

Acknowledgements

I would like to extend my sincerest and wholehearted gratitude to my supervisor professor Martin Russell, for the extraordinary kindness and generosity with which he shared his vast knowledge and expertise, inspiring every step of my way, constantly offering moral support on this arduous process and never running out of patience.

I would also like to thank Dr Jill Lehman for her thoughtful guidance and for the invaluable opportunity of two DRP internships which she offered with warmth and hospitality.

Many thanks are due to all the brilliant colleagues of the Speech Group, dr Peter Jancovic, dr Philip Weber, dr Linxue Bai, Chloe Seivwright, Xizi Wei and especially Mengjie Qian for always abounding in adeas and solutions to each of my concerns.

I am deeply grateful to my parents, Maria Papaioannou and Antonis Fringis for all their love and support and for instilling the belief in me from very early that I can achieve anything I choose in life.

Many thanks to all of my friends for all their support, and especially to all of my friends from the tenth floor of the Muirhead Tower who really came through offering me a writing home during the last turbulent weeks of writing my thesis.

Finally, Mattias Hjort thank you for absolutely everything!

Contents

List of Abbreviations

ASR	**A**utomatic **S**peech **R**ecognition
HMM	**H**idden **M**arkov **M**odels
PDF	**P**robability **D**istribution **F**unction
GMM	**G**aussian **M**ixture **M**odel
DNN	**D**eep **N**eural **N**etwork
BNF	**B**ottleneck **F**eature
PEALA	**P**honological **E**ffects **A**ssociated with **L**anguage **A**cquisition
DRP	**D**isney **R**esearch **P**ittsburgh
WER	**W**ord **E**rror **R**ate
MAP	**M**aximum **A** **P**osteriori
MLLR	**M**aximum **L**ikelihood **L**inear **R**egression
fMLLR	**f**eature **M**aximum **L**ikelihood **L**inear
MFCC	**M**el **F**requency **C**epstral **C**oefficient
DFT	**D**iscrete **F**ourier **T**ransform
ANN	**A**rtificial **N**eural **N**etwork
MLP	**M**ulti **L**ayer **P**erceptron
UBM	**U**niversal **B**ackground **M**odel
BP	**B**ack **P**ropagation
PCA	**P**rinciple **C**omponent **A**nalysis
LDA	**L**inear **D**iscriminant **A**nalysis
VTLN	**V**ocal **T**ract **L**ength **N**ormalisation
PN	**P**itch **N**ormalisation
SRN	**S**peech **R**ate **N**ormalisation
IPA	**I**nternational **P**honetic **A**lphabet

A large proportion of the work presented here has been published in four conference papers, which contribute considerably to the content of the present thesis. They are listed below:

- E. Fringi, J. Lehman and M. Russell, "Evidence of phonological processes in automatic recognition of children's speech", in INTERSPEECH 2015, 2015, pp. 1621-1624.
- E. Fringi, J. Lehman and M. Russell, "Analysis of phone errors in computer recognition of children's speech", in SLaTE - Workshop on Speech and Language Technology for Education 2015, 2015, pp. 101-105.
- E. Fringi, J. Lehman and M. Russell, "The role of phonological processes and acoustic confusability in phone errors in children's ASR", in WOCCI 2016 - Workshop on Child Computer Interaction, 2016, pp. 10-15.
- E. Fringi and M. Russell, "Analysis of phone errors attributable to phonological effects associated with language acquisition through bottleneck feature visualisations", in INTERSPEECH 2018, 2018, pp. 2573-2577.

In each of the publications E. Fringi is the main author, who carried out the described research under the supervision of M. Russell. J. Lehman represents the Disney Research Lab, which provided several of the data sets used in the experimental processes involved in the studies.

Chapter 1

Introduction

Children are significant potential users of speech and language technology with applications in education (Golonka et al., 2014,Wang, Waple, and Kawahara, 2009), entertainment (Al Moubayed and Lehman, 2015, Lehman and Al Moubayed, 2015) and speech therapy (Kitzing, Maier, and Åhlander, 2009). Speech offers children hands-free access to educational software without a need for keyboard skills. Furthermore, in applications such as interactive pronunciation tuition (Russell et al., 2000) and reading tutors (Mostow et al., 1994), automatic speech recognition (ASR) is not just another way to communicate with a computer, but the key enabling technology. Like computer assisted language learning, these applications make additional demands of the underlying speech technology, such as the ability to judge the quality of a child's pronunciation.

In comparison with adults', children's ASR appears to be more challenging and generally yielding inferior results. This aspect of ASR attracted researchers' attention after a publication which pointed out that age matched training data are of great importance in order to reach high accuracy results, especially in the case of children speakers, but even with matched training (i.e. training on children's speech) error rates for children's speech are typically greater than for adults' speech (Wilpon and Jacobsen, 1996). Because the vast majority of systems were trained on adults' speech, a large amount of research has been conducted since, aiming to adapt adult-trained systems to successfully recognise children's data (for example (Sinha and Ghai, 2009)), or alternatively to develop techniques which can smooth out the acoustic differences

between adults' and children's speech (for example (Gerosa, Giuliani, and Brugnara, 2007)). As a result, children's ASR has presented signs of improvement, but not to such an extent as to elicit accuracies comparable to those of systems trained and tested on adult speech. It has been suggested that for this issue to be addressed, linguistic understanding of children's speech development needs to be employed in order to either provide a solution or an explanation (Russell and D'Arcy, 2007), however such an approach has yet to be pursued, except through attempts to modify the pronunciation dictionary, which have nevertheless given disappointing results (Shivakumar et al., 2014a).

The present work aims to explore the influence of phonological effects associated with language acquisition (PEALA) in children's ASR and investigate whether they can be detected in systematic patterns of ASR phone confusion errors or they can be evidenced in systematic patterns of acoustic feature structure. Research from phoneticians and speech therapists is presented in the context of ASR in order to define a set of errors that can be predicted from PEALA. A measure of statistical significance is introduced, which offers a direct comparison between the presence of PEALA related errors in children's and in adults' speech. This technique is applied to the confusion matrices for a number of ASR experiments on children's speech using GMM-HMM and DNN-HMM systems and a range of children's speech corpora. The latter include corpora of read speech recorded in a quiet environment, read speech recorded in more realistic environments, prompted and spontaneous speech. The results reveal a relationship between the proportion of errors that are attributable to PEALA and ASR performance, and a simple model is proposed to explain this relationship. Moreover, state of the art methods for the visualisation of acoustic features of children's speech, such as bottleneck features (BNFs) and i-vectors, are implemented affording meaningful insights into the development of acoustic structure in children's ASR.

The thesis is structured into seven parts, the first one being the present introductory chapter, which outlines the technical background of the project. Chapter 2 features a comprehensive literature review of speech development and language acquisition research, aiming to establish a framework which will be later used as the point of

reference in the analysis and interpretation of the thesis' results. Chapter 3 contains the description of the speech corpora used, comprising both children's and adults' data and including two newly collected corpora which were provided for the purposes of this research by the Disney Research lab in Pittsburgh (DRP). Chapter 4 contains the description of all the ASR systems developed for the present work, along with an analysis of their results. Chapter 5 is one of the focal points of the thesis, analysing the output of the recognisers described previously in search of PEALA related systematicity and predictability within the patterns of ASR phone confusion errors. It involves the development of a simple statistical test to detect PEALA-related phone substitution errors in ASR confusion matrices, the analysis of the ASR phone confusion results, as well as the development of a simple mathematical model to explain the dependency of the ability to detect PEALA-related errors and ASR error rate. Chapter 6 continues the pursuit of PEALA related systematicity and predictability, this time in the patterns observed in visualisations of the acoustic features of our ASR systems, obtained through bottleneck feature and i-vector extraction. Finally, Chapter 7 lists the conclusions drawn from each part of the present research.

1.1 Automatic Speech Recognition (ASR)

Automatic speech recognition is the computational process through which a speech signal is transcribed into its corresponding text. It involves three main procedures: signal processing, model generation and decoding.

During signal processing, speech is converted from a continuous signal to a sequence of discrete acoustic feature vectors which serve as speech representations in the computational domain. During model generation, part of the resulting vectors is used as training data in the creation of an acoustic model, while their corresponding transcriptions are used in the creation of a language model. The acoustic model consists of statistical rules which express the correspondence between feature vectors and their linguistic transcriptions, while the language model determines how probable a sequence of linguistic units (phonemes or words) is, given the statistical information

provided in the training set. An N-gram language model describes this probability for a particular unit taking into account N-1 units preceding it. The acoustic and the language model along with the rest of the extracted feature vectors, which make up the test set, are used in the decoding process. According to this, each sequence of features observed in the test set is mapped to the most probable sequence of linguistic units (phonemes in phone-level ASR or words in word-level ASR) dictated by the models. Thus, the aim of the decoding is to find a phone or word sequence $\hat{W}$ that will maximise the probability $P(W|X)$ for a given sequence of feature vectors, X:

$$\hat{W} = argmax P(W|X) \tag{1.1}$$

Application of Baye's rule to the probability in question,

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \tag{1.2}$$

demonstrates that it is dependent on $P(X|W)$ which is determined by the acoustic model, on P(W) which is determined by the language model, and on P(X) which is constant and independent to the choice of W (therefore can be ignored). It becomes apparent then, that ASR accuracy depends directly on the quality of the acoustic and language models (Holmes and Holmes, 2001).

A scoring system is in place to evaluate the accuracy of the decoding, allowing for errors such as deletions, insertions, or substitutions (Young et al., 1997) in the calculation of the word (or phone) error rate (WER). The equations below show how the scoring measures are calculated.

$$\%Correct = \frac{N - D - S}{N} \times 100\% \tag{1.3}$$

$$\%Accuracy = \frac{N - D - S - I}{N} \times 100\% \tag{1.4}$$

$$\%WER = 100\% - \%Accuracy \tag{1.5}$$

An enduringly popular method to carry out ASR is through the use of Hidden Markov Models (HMMs). An HMM is a Markovian system of states defined by state probabilities, state transition probabilities and an initial state distribution. It is Markovian because any state at time t depends only on the previous state at time $t-1$ and not on the whole sequence of preceding states, and it is hidden because its state at each time t is unknown. In the case of ASR, words are represented by HMMs as sequences of phones, using a pronunciation dictionary, and each of the phones is represented as a sequence of states. The aim of the training process is to create an optimum model λ which will locally maximize the probability of each observation sequence of speech features in the training set given this model: $P(X|\lambda)$ (Rabiner, 1989). Therefore, if $X = \{X_1, \cdots X_N\}$ is a set of sequences of feature vectors, where each sequence X_n corresponds to a spoken utterance in the training set, the objective is to (locally) maximize

$$P(X|\lambda) = \prod_{n=1}^{N} P(X_n|\lambda) \tag{1.6}$$

This can be addressed in different ways. In the present thesis two methods are employed and described in the following sections: Gaussian Mixture Model HMMs (GMM-HMM based ASR) and Deep Neural Network HMMs (DNN-HMM based ASR). In both cases the decoding process is succeeded through the Viterbi algorithm, a technique based on dynamic programming methods, which computes the single best state sequence to maximise $P(W|X)$ (Rabiner, 1989).

1.1.1 GMM-HMM based ASR

GMM-based HMMs have been the most commonly used method in ASR research for several decades (Liporace, 1981, Juang, 1985). A GMM of M components is a weighted sum of M Gaussian probability density functions (PDFs):

$$\sum_{m=1}^{M} w_m p_m(x) \tag{1.7}$$

(where $0 < w_m < 1, \sum_{m=1}^{M} w_m = 1$ and each p_m is a Gaussian PDF) which is associated with an HMM state and used to calculate the probability of each observed feature vector relative to that state. An optimisation of each GMM's parameters is required in order to develop an efficient model. This is achieved using the Baum-Welch algorithm. According to this, a first estimation λ_0 of the model is made, and then an iterative procedure arrives to a model λ_n which satisfies the relation $P(X|\lambda_n) \geq P(X|\lambda_{n-1})$.

Phone-level HMMs are typically classified as monophone or triphone, depending on the inclusion or not of contextual information. In a monophone system, each phone is typically represented by a 3-state HMM which corresponds to a single phone in the pronunciation dictionary. In a triphone system, phones are represented by 3-state HMMs, which contain information about the corresponding phoneme as well as the preceding and following phonemes. For example, the phone [t] in the word 'water' would be expressed as /t/ in a monophone system and as /ao-t+er/ in a triphone system. The introduction of triphones significantly increases the number of parameters and hence increases the amount of training data that is needed. One solution is to use phonetic decision tress to identify triphone states that are sufficiently similar, so that different triphone HMMs can share the same states. Such systems are referred to as tied state systems. Figure 1.1 shows an example of conventional and tied state triphones. The use of phonetic decision trees differentiates between contextually equivalent sets of HMM states by employing a binary tree method according to which questions are applied to each phoneme in a word or phrase, establishing the properties of its adjacent phonemes. This way less complicated models can be developed requiring less training data (Young, Odell, and Woodland, 1994).

To enhance the performance of a system, especially when there is not a lot of training data available, there are a few adaptation techniques which are widely applied. These techniques might be applied to adapt a speaker-independent system to a specific speaker, using a small amount of training data from that speaker.For example, maximum a posteriori (MAP) estimation for HMMs, estimates the parameter vector λ

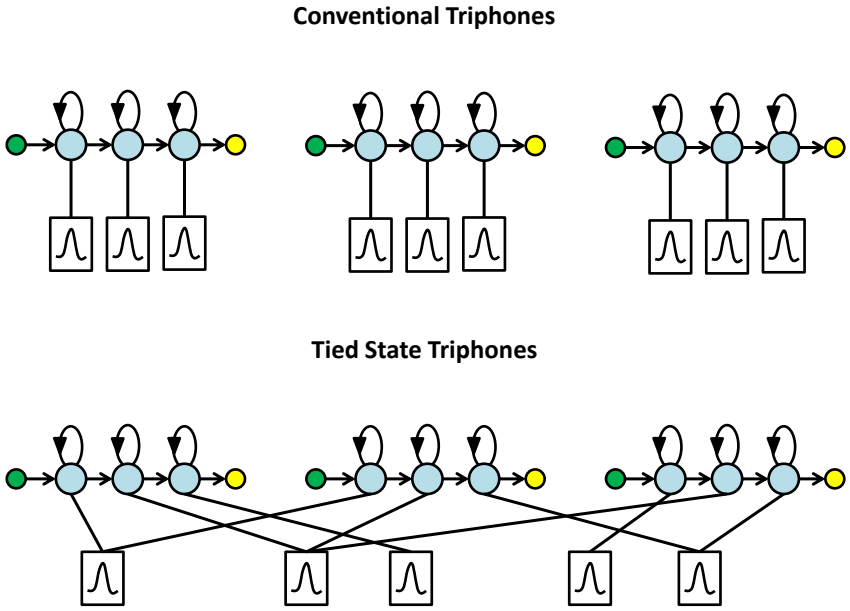


FIGURE 1.1: *Triphone HMMs*

to maximize the Bayesian relationship mentioned previously.

$$P(\lambda|X) = \frac{P(X|\lambda)P(\lambda)}{P(X)} \tag{1.8}$$

Whereas in ML training the objective is to find parameters λ that maximise $P(X|\lambda)$, equation (1.8) shows that in MAP adaptation the objective is to maximize the product $P(X|\lambda) \times P(\lambda)$, where $P(\lambda)$ is the *prior* probability of the parameter set λ. In a typical application of MAP adaptation to speaker adaptation, the robustly-trained speaker-independent model is used as the prior distribution. It has been shown that this way a recognizer can perform much better on limited training data (Gauvain and Lee, 1994). Another useful technique is that of maximum likelihood linear regression (MLLR). According to MLLR a linear transformation is applied on the Gaussian mean parameters of an HMM system, leading to improved speaker adaptation results (Leggetter and Woodland, 1995). A variant of MLLR is fMLLR, or feature MLLR. As in MLLR a linear transformation is derived that maximises the probability of the adaptation data

relative to the model. However, unlike in MLLR where the transformation is applied to the model parameters, in fMLLR it is applied to the feature vectors. In this way, for example, a set of adaptation data from different speakers can be transformed relative to the same GMM-HMM, resulting in a more homogeneous training set and the GMM-HMM can be discarded if it is no longer required.

As far as speech representation methods are concerned, Mel frequency cepstral coefficient (MFCC) feature extraction, a method that encompasses the perceptual particularity of human hearing toward frequencies, has been shown to be optimal (Davis and Mermelstein, 1980) for GMM-HMM systems and has been in the mainstream of ASR for years. According to this process, a pre-emphasis filter is initially applied on the speech signal to make up for an attenuation of upper spectrum frequencies (6dB per octave). Then, the signal undergoes a frame blocking procedure multiplied by a Hanning or Hamming window in order to be segmented into 20-25 ms frames which overlap every 10 ms. Next, a Discrete Fourier Transform (DFT) is applied on each of those frames leading to its frequency domain representation. The next step is the Mel filterbank processing, a transformation that maps the signal from its physical frequency scale (Hz) to a scale which corresponds to its perceived frequency (Mel). During this, the Fourier transformed signal passes through a set of triangular band-pass filters. Finally the MFCCs are extracted by applying a Discrete Cosine Transform (DCT) to the logarithm of the Mel frequency coefficients produced before. From the resulting features, only the first 13 are typically used in speech recognition together with their first and second derivatives constituting a set of 39-dimensional feature vectors (Holmes and Holmes, 2001).

1.1.2 DNN-HMM based ASR

A deep neural network (DNN) is a type of artificial neural network (ANN) which has become prominent in the realm of machine learning applications, such as ASR. ANNs are computational structures attempting to simulate the function of biological neural activity towards information processing. A basic such structure is the multi-layer

perceptron (MLP), which involves a feed-forward network consisting of multiple layers, equipped with an activation function (threshold, sigmoid, or tanh), and produces non-linear input-to-output mappings (Rummelhart, Hinton, and Williams, 1986). All layers between the input and the output layer are referred to as hidden layers, and all units within hidden layers are referred to as hidden units or nodes. Each layer is connected with the preceding and the following layer through feed-forward links between the nodes (Bishop, 1995). ANNs with a minimum of two hidden layers are classified as deep neural networks (DNNs). MLPs with over two hidden layers, are the kind of DNNs employed in ASR. Figure 1.2 shows an example of an MLP with two hidden layers.

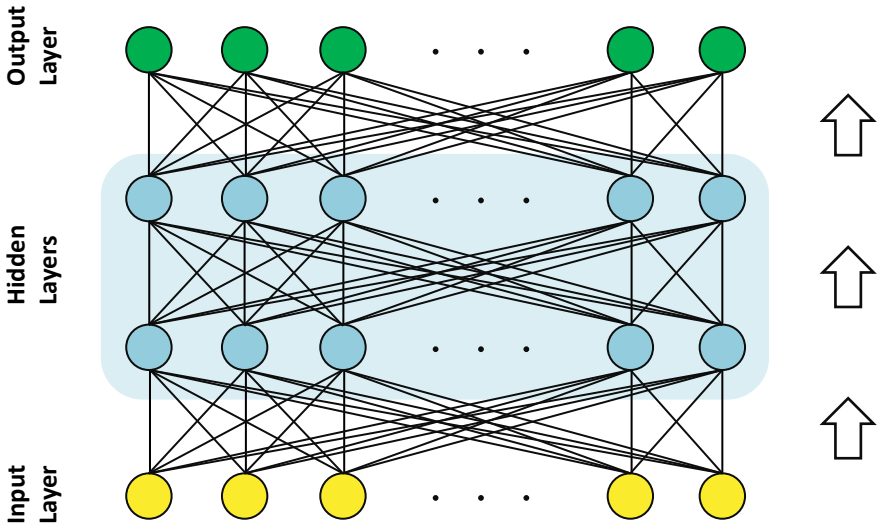


FIGURE 1.2: *An MLP with two hidden layers*

In the context of ASR, an l-layer MLP is defined by $l-2$ hidden layers between the input and the output ones, $l-1$ weight matrices W_m ($m = 1, ..., l-1$) (where the $(i,j)^{th}$ element of W_m is the weight associated with the i^{th} unit in layer m and the j^{th} unit in layer $m+1$, considering the input layer to be layer 1 and the output layer to be layer l), $l-1$ bias vectors b_m, along with an activation function a. The linear output of

each layer m ($m \geq 0$) is given by the following formula:

$$o_m = o_{m-1}W_{m-1} + b_{m-1} \tag{1.9}$$

The number of input units is equal to the dimension of the input feature vectors and the number of output units is equal to the number of tied states. In the context of DNNs the tied states are often referred to as 'senones'. In this case the outputs of the DNN are required to be the probabilities of the input feature vectors given each of the senones. The softmax activation function is used to convert the outputs of the DNN into the posterior probabilities of each of the senones given the input feature vector and these are converted into state-dependent feature vector probabilities using Bayes' rule:

$$a_{l-1} = Softmax(o_{l-1}) \tag{1.10}$$

where

$$Softmax(z) : \sigma(z)_j \equiv \frac{exp(z_j)}{\sum_{k=1}^{K} exp(z_k)} \text{ for } j = 1, \cdots, K \tag{1.11}$$

This incorporates discriminative training, since the requirement that the DNN outputs are posterior probabilities, and hence sum to one, so it means that the probability of a particular senone can only be large if the probabilities of the remaining senones are low.

Training of an MLP is accomplished through a supervised process called error back propagation (BP). According to BP, a training set comprising inputs along with their corresponding outputs is provided to the system. Each input is applied to the input layer and propagated through the network to the output layer where the error, the difference between the actual output and the target output, is calculated. Granted that the activation functions are differentiable, the derivatives of the error with respect to the weights and the biases can be used to optimise these parameters using gradient descent, resulting in a local minimum of the error function (Bishop, 1995).

In the case of DNN-HMM training the inputs to the DNN are feature vectors in

context and the outputs are the corresponding senones. Thus, for each training sequence an alignment is required that maps each feature vector in the sequence onto the corresponding senone. If there are N senones then the DNN has N output units and the target for a feature vector x_t is a 'one-hot' N-dimensional vector p_t such that $p_t(n) = 1$ if x_t corresponds to the n^{th} senone, and 0 otherwise. In the standard procedure a triphone GNMM-HMM system is first produced and the senones (tied states) are determined using the phone-decision-tree method described above. Once the GMM-HMM system has been fully optimised, the Viterbi decoder is used to compute the optimal alignment between each training sequence of acoustic vectors and the senone set. These alignments are passed to the DNN-HMM training process. It is evident that the quality of DNN-HMM training depends critically on the quality of this alignment (Yu and Deng, 2014).

As far as feature representations are concerned, DNNs facilitate the extraction of an alternative to MFCCs, which is implemented in one of the recognisers developed in this thesis, namely bottleneck features (BNFs). A bottleneck neural network is a DNN with a compressed hidden layer of lower dimensions (bottlenck layer). The information required in the DNN to construct the output for a particular input feature vector is forced through the bottleneck and hence the bottleneck provides a low-dimensional encoding of that information. BNFs consist of the output of such a layer, and offer the advantage of low dimensionality in the feature representation.

1.1.3 Recent progress on DNN-HMM based ASR

In recent years there has been affluence of studies exploring alternative types of neural networks. As a result, algorithms such as convolutional neural networks (CNN), recurrent neural networks (RNN) and long short-term memory networks (LSTM) have become state of the art in ASR offering robust acoustic modelling techniques.

Convolutional neural networks are commonly used in image processing, but have also been successfully implemented in ASR (Palaz, Magimai-Doss, and Collobert, 2015, Palaz, Doss, and Collobert, 2015, Huang, Li, and Gong, 2015). They train into detecting specific types of local patterns in the input data, which are stored in feature

maps. Downsampling of these feature maps produces a single output from a specific region of the data which is then fed into a standard network (Huang, Li, and Gong, 2015).

Recurrent neural networks accommodate a loop structure which allows information from the entire history of inputs fed to the system in the past, to persist and shape the contents of the output vector at any given moment. This way, RNNs are more advantageous than typical DNNs which process information at a fixed-length sliding window of frames. However, during back propagation training, RNNs may end up with gradients that either increase (tending to infinity) or decrease (tending to zero) exponentially causing computational problems. As a solution, long short-term memory RNNs have been introduced to offer stability during back propagation. They employ input, output and forget gates, which control the flow of information in and out of the system over arbitrary time intervals (Li et al., 2015). LSTM-RNN systems have been broadly used to build improved acoustic models which have produced promising resuts (Sak et al., 2015a, Sak et al., 2015b, Weninger et al., 2015, Zeyer et al., 2017).

Another approach towards more efficient DNNs, is through end-to-end speech recognition, which was introduced in 2014 (Graves and Jaitly, 2014). This method questions the necessity of GMM-HMMs and attempts to re-think the training process in a less conventional way. Aiming to save computational time, it skips the GMM-HMM-based pre-training that provides the DNN with time aligned phone sequences in traditional systems. Instead, it uses a single network to automatically match the raw input signal to the appropriate sequence of phones, without applying linguistic knowledge (Hori et al., 2017). End-to-end speech recognition systems combined with RNNs (Miao, Gowayyed, and Metze, 2015), LSTM (Soltau, Liao, and Sak, 2016) or CNN models (Zhang et al., 2017) have recently offered enhanced ASR performance with considerably reduced decoding time. Moreover, applications in emotion recognition have been found to outperform the traditional approaches (Trigeorgis et al., 2016).

1.2 Feature Visualisation Techniques

In addition to the quantitative analyses of children's speech included in this thesis it is useful to visualize the data for more subjective analyses. This section describes the different methods that have been used to visualise children's realisations of individual speech sounds and how these realisations change with age, and to visualise the relationships between groups of children of different ages.

1.2.1 Bottleneck Features (BNFs)

DNN bottleneck features have already been described in section 2.1.2. They provide a relatively low dimensional representation that captures the information used in the DNN to map acoustic feature vectors onto senone posterior probabilities. It has been shown elsewhere that 2 dimensional projections of BNFs provide a visualisation of speech in which the relative locations of speech sounds reflect their phonetic properties and the relationships between them (Bai et al., 2015, Weber et al., 2016a).

1.2.2 iVectors

I-vectors are vector representations of speakers, or groups of speakers, that were developed for detection problems such as automatic speaker identification (SID) and language identification. Early approaches to SID used Gaussian Mixture Models (GMMs). First, a single speaker-independent GMM is trained on data from all available speakers. This is the 'Universal Background Model' (UBM). Speaker-dependent GMMs are obtained from the UBM using MAP-adaptation with a small amount of speaker data. This method is referred to as GMM-UBM (Reynolds, 1992; Reynolds and Rose, 1995).

GMM-UBM methods were superseded by 'supervector' methods in which the means of a speaker-dependent GMM are stacked to create a speaker-dependent supervector. For example, for an M component GMM and D dimensional feature vectors the dimension of the supervector is $D \times M$. SID is achieved using vector space classification methods such as Support Vector Machines (SVM). This method is referred to as GMM-SVM (Campbell et al., 2006).

A problem with supervectors is their high dimension. For example, for a 512 component GMM and 20 dimensional feature vectors, the supervectors are 10,240 dimensional. Therefore reducing the dimension of these supervectors is a priority. Principal Components Analysis (PCA) is not truly applicable because the covariance matrix of the supervector data cannot be reliably computed (the number of training vectors is typically less than the number of dimensions). A solution to this problem is to replace the supervector with an i-vector (Kenny, Boulianne, and Dumouchel, 2005). In the i-vector approach a low dimensional (lower than the dimension of the supervector space) 'Total Variability space' V is defined together with a linear mapping $T : V \rightarrow S$, where S is the supervector space. A supervector s is written as $s = Tw + \epsilon$ where w is the i-vector representation of s, chosen so that the GMM corresponding to the supervector Tw maximises the probability of the training data across the training set. The term ϵ is an error term. I-vectors are widely used in SID (for example, Dehak et al., 2011).

The i-vector training process is an iterative algorithm that calculates the mapping T and the i-vectors w. It uses posterior probabilities identical to those used in the E-M algorithm. In the context of i-vectors these are referred to as the 'sufficient statistics'. The most recent i-vector systems dispense with GMMs, use senones from an ASR system instead of GMM components, and use the senone posterior probabilities from a DNN trained for ASR as the sufficient statistics. The resulting i-vectors are sometimes referred to as DNN i-vectors.

1.2.3 Dimensionality Reduction

Dimensionality reduction is the process of transforming a set of variables to an equivalent set of lower dimensions. Two of the most popular linear transformation techniques for achieving this are: principal component analysis (PCA) and linear discriminant analysis (LDA). PCA is aimed to project a dataset onto a lower dimensional subspace whilst preserving the information represented by the dataset. Its target is to identify patterns within the given data and locate the axes along which the set's variance is maximized. It is an unsupervised algorithm since it does not take into account

any labels of data classification. LDA on the other hand, aims for the directions maximizing its class discriminability. It is a supervised technique which uses class labelling information in order to ensure optimal class separation.

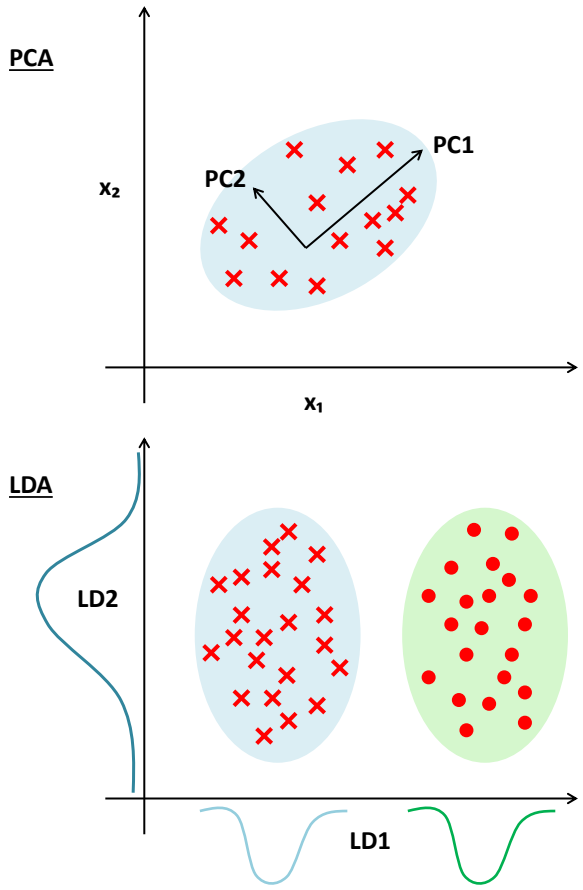


FIGURE 1.3: *Examples of PCA and LDA taken from https://sebastianraschka.com/faq/docs/lda-vs-pca.html*

Since the visualisation of acoustic features relies highly on the discrimination between features of different phonetic labels, the most appropriate method for dimensionality reduction is LDA. Assuming the dataset in question consists of d-dimensional features and needs to be projected onto a k-dimensional subspace, this can be implemented in five steps [1]. Firstly, the d-dimensional mean vectors for the various classes

[1]https://sebastianraschka.com/Articles/2014_python_lda.html

need to be computed.

$$m_i = \frac{1}{n_i} \sum_{x \in D_i}^{n} x_k \tag{1.12}$$

Then, follows the computation of the two scatter matrices, the within-class scatter matrix (S_W) and the between-class scatter matrix (S_B).

$$S_W = \sum_{i=1}^{c} S_i = \sum_{i=1}^{c} \sum_{x \in D_i}^{n} (x - m_i)(x - m_i)^T \tag{1.13}$$

$$S_B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T \tag{1.14}$$

where $m =$ the overall mean and $N_i =$ the size of each class. Next, the eigenvectors $(e_1, ..., e_d)$ and the eigenvalues $(\lambda_1, ..., \lambda_d)$ for the product of scatter matrices $S_W^{-1} S_B$ need to be computed by solving the equation $Av = \lambda v$, where $A = S_W^{-1} S_B$, $v = eigenvector$ and $\lambda = eigenvalue$. Once these are obtained, the eigenvectors have to be sorted by decreasing eigenvalue and the k first of them have to be accumulated into forming a $d \times k$ matrix W. Finally, W can be used to transform a $n \times d$ set X into a $n \times k$ subspace $Y = XW$.

Chapter 2

Phonological Effects of Speech Development

This chapter builds up the motivation for the research presented in the thesis. It contains an extensive review of developments in children's ASR and how these point to further investigation of the linguistic aspects of children's speech in order to obtain a better grasp of the reasons behind children's ASR lower performance. Additionally, an exhaustive review of research in speech development is presented, providing the basis for formulating the thesis' research questions.

2.1 Children's ASR

A special case of ASR is this on children's speech, because it appears to be much more challenging than that on adults'. The majority of research on the subject has been conducted using adult-trained systems, which prove to be inappropriate for processing children's speech. In 1996 Wilpon and Jacobsen investigated the impact of training data in ASR across a great span of ages including children, adolescents, adults and the elderly. They found that when trained with a combined dataset of all age groups, the word error rate (WER) of their recognizer was 170% higher for children than for adults, and when trained on children data alone this percentage was reduced to 122% (Wilpon and Jacobsen, 1996). It thus became apparent that ASR performance depends

on the choice of training data, but not exclusively. There must be more factors keeping it from reaching 100% accordance with the adult systems' performance.

This was confirmed in a study which straightforwardly compared the performance of an adult trained recognizer to that of a child trained recognizer when tested on both adults' and children's speech (Elenius and Blomberg, 2004). The results showed that the adult system reached almost perfect accuracy on adult testing data (97%), but was severely impaired when tested on children data (51%). Similarly, the child system performed very well on children speech (87%) and poorly on adults (61%). In spite of the relatively better performance of the child recognizer on children speech, there was still a large gap distinguishing it from the almost excellent performance of the adult system (Elenius and Blomberg, 2004).

Bandwidth restrictions also have a detrimental effect on ASR accuracy for children's speech compared with adults' speech. The effect of speech bandwidth reduction on human and computer recognition of children's speech is studied in (Russell, D'Arcy, and Qun, 2007). The results suggest that machine speech recognition WERs for telephone bandwidth (4kHz) speech are likely to be between 30% and 95% greater for children's speech than for adult's speech, even if the children's ASR system is trained on speech from children of a similar age.

These studies attribute the divergence between adults' and children's ASR to the fact that children's articulators, such as the vocal tract, are smaller than those of adults and not yet fully developed. A series of studies focusing on the acoustic properties of children's speech and the way they differ from those of adults (which will be presented further in Chapter 3), indicates that acoustic variability of children's speech is a major factor in the mismatch between children's and adults' recogniser performance (Lee, Potamianos, and Narayanan, 1999, Lee, Potamianos, and Narayanan, 2014, Gerosa et al., 2006).

In order to deal with the observed contrast between the acoustics of adults' and children's speech, several compensating techniques have been applied. The most salient of those techniques is vocal tract length normalization (VTLN). The main idea behind this method is to estimate a so called warping factor which approximates the

ratio between the length of each speaker's vocal tract and the length of a standard hypothetical vocal tract used as a point of reference. In practice this is realized by computing the warping factor $\check{\alpha}$ specific to speaker i which is implemented in the feature extraction process and warps the feature vectors so that $\check{\alpha}_i = argmax_\alpha P(X_i^\alpha|\lambda, W_i)$, where $X_i^\alpha = X_{i,1}^\alpha, X_{i,2}^\alpha, \ldots, X_{i,N_i}^\alpha$ is the set of features corresponding to all the utterances by speaker i warped by α and $W_i = W_{i,1}, W_{i,2}, \ldots, W_{i,N_i}$ is their set of matching transcriptions. VTLN has been reported to decrease WER by 20% (Lee and Rose, 1998).

Another important normalizing technique is pitch normalization (PN). As well as VTLN, this method is also implemented in the feature extraction process by altering the bandwidths of the first few filters in the Mel filterbank. The high fundamental frequency in children's speech results in widely-space harmonics in the spectrum and the spacing between these harmonics is potentially greater than the bandwidth of the low frequency filters in the mel-scale filterbank. Thus the positions of the harmonics relative to the filters is an additional source of variability in conventional feature vector representations of children's speech. Due to the use of a mel filterbank on the high fundamental frequency typical of children, it has been suggested that the main pitch related issues in high frequency signals occur in the area below 1 kHz. Therefore, the bandwidth of all filters with center frequencies below 1 kHz is raised to approximately 200Hz or 300Hz. This way the recognizer performs better on high fundamental frequency speech, with WER improvement of 9% or 16% (depending on the task difficulty) on children speech. This method is considered to be efficient both in adult trained ASR systems and children trained ones (Ghai, 2011).

A minor but still valuable improvement has been attributed to the use of speech rate normalization (SRN). According to SRN, the length of each utterance has to be adjusted as to be comparable to the length of a "standard" utterance length, set by the training data. The parts of the speech feature extraction that get most afflicted by the variability in speech rate are considered to be the first and second derivatives of the MFCCs. To solve this, the frame grid which is used in the computation of the deltas and the deltadeltas, is modified to accommodate for fast or slow speech rate (Pfau, Faltlhauser, and Ruske, 2000).

In light of these adaptation processes, several important studies have been undertaken seeking to bridge the performance gap between children and adults ASR. In Narayanan and Potamianos (Narayanan and Potamianos, 2002, Potamianos and Narayanan, 2003), an adult trained and a children trained recognizer were tested in parallel on children speech, either with or without speech normalization techniques. The WER of the baseline adult trained system dropped from 15.9% to 6.7% when trained on children data and this percentage declined further to 4.9% when normalization methods were involved (Narayanan and Potamianos, 2002, Potamianos and Narayanan, 2003). Similarly, another study exploited a combination of acoustic modelling techniques under both matched and unmatched training conditions, resulting in a baseline WER of 20% for adult HMM acoustic models and 11.6% for child HMM acoustic models, which decreased to 14.8% and 10.5% respectively after normalization (Gerosa, Giuliani, and Brugnara, 2007). Such results illustrate perfectly the gradual progression from a poorly performing ASR system to its improved versions; there is however still a lot of room for improvement for the recognition of the younger speakers. A study analogous to the previously mentioned compared the improvement induced by VTLN on a system trained and tested on adult speech with that induced on a system trained and tested on children speech. It turned out that error rate for the adult system was reduced by 10.5% due to VTLN, while the error rate for the children system only allowed for a 5.3% reduction (Giuliani and Gerosa, 2003). Therefore it is becoming clear that even when its training and testing data are matching in respect to age, and even when the most advanced speech normalization methods are applied, an ASR system does not yield as good results on children speech as it does on adults'.

Apart from the discrepancies in acoustic components, it has also been hypothesized that it is a general linguistic variability in children's speech that impedes children's ASR. The constant phonological development that children's speech is undergoing creates disfluencies and hesitation phenomena in younger speakers, which eventually recede with age. This aspect of the problem has not been looked into thoroughly, as ASR research has been mainly focusing on making the best out of the acoustic adaptation techniques (Potamianos and Narayanan, 2007). In addition, the

emergence of new state-of-the-art technologies (Chapter 1.1.3) combined with access to vastly large data sets by companies such as Google, shift the focus from understanding how the different aspects of children's speech variability affect ASR, to building high performing large vocabulary systems whose data demographics nevertheless, are publicly unavailable. In a recent paper published by Google labs (Liao et al., 2015), high word recognition accuracies were attained (9.4% WER in the best instance) by two LSTM recognisers (one using a recurrent neural network and one using a convolutional DNN) built on thousands of hours of children's speech collected through a Google application. The raw training material was anonymised, therefore a child speech DNN classifier had to be employed in order to isolate children's speech training data. This method even though efficient as far as word error rates are concerned, does not offer any insight on how the typical difficulties of children's ASR were tackled. Especially since there was no definite confirmation that all the training data came indeed from child speakers, or even if it did, there was no information on the distribution of speakers' ages, therefore it could well be claimed that the sample was shifted towards older children who have ceased to exhibit developmental phenomena or do so at a minimal degree.

The present thesis aims to investigate how linguistic variability in children's speech affects ASR for different age groups. In the next section, work from linguists and speech therapists is presented aiming to examine children's speech from a phonological perspective.

2.2 Early Speech Development

Prominence of speech in human growth and evolution has been pointed out by researchers not only due to the fact that it enables advanced communication, but also because it shapes individuals' mental processes and influences their understanding of the world around them. It is through word comprehension that children manage to isolate objects' functions, assign meanings and make generalisations which eventually lead to categorical and abstract thinking (Luria and Yudovich, 1959). Speech

development is a complex and lengthy natural procedure, which involves an intricate combination of physical and cognitive mechanisms. Production as well as perception of speech, progress in an interconnected manner from the first weeks of infancy.

Evidence shows that since the first months of life, infants have the ability to categorically discriminate among speech sound contrasts they are presented with, such as the phonemic distinction between voiced and voiceless consonants (Eimas et al., 1971 Jusczyk and Derrah, 1987). This ability is not specific to native language, but extends to sound contrasts found in non-native languages that infants have not been exposed to in their environment and adults fail to identify (Vihman, 2014, Werker et al., 1981, Werker and Tees, 2002). However, this sensitivity to universal phone contrasts ceases after about six months, and infants' discrimination performance is no longer superior to adults' (Kuhl et al., 2006, Werker and Tees, 1983). This so-called 'perceptual narrowing' signifies a shift from implicit to explicit learning and has been interpreted as a consequence of different brain areas maturing within different time frames (Vihman, 2017). During the first six months, learning relies on the neocortex, which does not require conscious attention, but gradually collects procedural information. At the second half of the first year though, the prefrontal lobes and hippocampus start contributing to the learning experience, the former by focusing or inhibiting attention when necessary and the latter by binding and preserving the experienced information into memory. This way speech perception is reorganised and directed towards the specificities of native language.

Perceptual narrowing is of great significance for speech development research because it offers a basis for phonemic contrast classification. According to Burnham, all human language contrasts lie on a robust to fragile continuum depending on their universality and psycholinguistic salience. Robust contrasts are psychoacoustically strong and thus represented widely across the different languages of the world, while fragile contrasts have a weak psychoacoustic basis and are less common. The developmental loss of contrast perception after the sixth month of infancy is thought to occur in two stages, first between 6 and 12 months as a result of no exposure to non-native

fragile contrasts and then between 4 and 8 years as a result of no experience with non-native robust contrasts (Burnham, 1986). The distinction between 'exposure to' and 'experience with' is noteworthy, since it implies that exposure involves perception alone and experience is a combination of perception and production. It follows then that children's own attempts to speech production play a reinforcing role in language acquisition.

Within the first six weeks of life, infants produce reflexive and vegetative sounds, which slowly integrate into cooing, an early type of structured vocalization in a 'consonant-vowel' form (Stark, 1978). After the first six or seven months cooing grows into babbling, around the first twelve months first words make their appearance, and after 3 years of age first sentences emerge (Lust, 2006). A substantial vocabulary expansion comes out at a high pace around the age of two, with what is termed 'vocabulary spurt'. During this phenomenon, at least 8 new words per week are added in the child's repertoire, which increases rapidly from 20-40 to 500 words or more (Barrett, 1995). Nonetheless, the quality of these early productions is characterised by high variability, which is ubiquitous throughout the whole duration of phonological development both phonetically and phonemically.

2.3 Acoustic and Linguistic Variability

The connection between children's physiological development and the acoustic properties of their spoken utterances was explored by Lee at al. in 1999 and in 2014 with two papers focusing on the developmental changes of temporal and spectral features of American English monophthongs and diphthongs respectively (Lee, Potamianos, and Narayanan, 1999, Lee, Potamianos, and Narayanan, 2014). It was found that with decreasing age there is an increase in both within and between subject variability of speech duration, frequency and spectral envelope. It was also noted that the transition between the initial and final part of a diphthong starts earlier for older speakers. Along the same lines Gerosa et al. looked into the average duration of vowels and consonants, the consonant-vowel transitions as well as the temporal and spectral

intra-speaker variability across ages. Their results exhibit the expected age correlation trend, with a decrease of phone durations (25% for consonants and 41% for vowels) between the ages 7 and 17 and a reduction in the intra-speaker temporal and spectral variability as a function of age. Interestingly enough, it was also observed that consonant-vowel transitions last longer for adults than for younger children, supporting the view that co-articulation is not well established in younger ages (Gerosa et al., 2006).

From a linguistic perspective, phonemes are acquired gradually after numerous unsuccessful production attempts varying among words (inter-word variability) or within different productions of the same word (intra-word variability)(Ferguson and Farwell, 1975). This can be explained as a consequence of a not yet mature motor control system (Kent, 1992) combined with an unsettled underlying phonological representation of words (Sosa and Stoel-Gammon, 2006). However, variability as a property of typically developing children should not be confused with variability as an indicator of therapy requiring speech ailment. According to an evaluation of normative data collected from children between 3 and 7 years old, normal variability occurs with consistency in typically developing children, while children with speech sound disorders are inconsistent in their variable productions (Holm, Crossbie, and Dodd, 2007). Differentiating between variability and inconsistency is important for diagnostic reasons and also because consistency of typical mispronunciations facilitates their categorization.

2.4 Age of Acquisition

Evidently, not all sounds are equally variable or remain variable equally long. Words that appear more frequently in a language or groups of words that share phonological similarities have been found to facilitate less variable pronunciations (Sosa and Stoel-Gammon, 2012). Research suggests that phones commonly appearing in prelinguistic vocalisations are also frequent in early word productions and are the first ones to reach adult level articulation (Stoel-Gammon, 1985). Pinpointing the exact age when each

phoneme can be considered as acquired has been the focus of linguists and psychologists for several decades, leading to a number of studies offering inventories of phonological acquisition. For the most part, there is a consensus in literature that across languages vowels are acquired very early, usually by the age of three, while consonants take longer and exhibit varying levels of difficulty, particularly when they occur in clusters (Priester, Post, and Goorhuis-Brouwer, 2011, Gangji, Pascoe, and Smouse, 2015, Saaristo-Helin, Kunnari, and Savinainen-Makkonen, 2011). Stops and nasals are reported to be the first consonants to be mastered by children around the world, while fricatives and liquids prove to be more challenging and are acquired last. Apart from occasional exceptions (for example /d/, which is among the first consonants to be acquired in most languages, but due to its minimal presence in the Finnish language it is acquired last in Finnish (Saaristo-Helin, Kunnari, and Savinainen-Makkonen, 2011) there appear to be universal trends rendering specific speech sounds less challenging or perhaps (to quote Burnham) more psychoacoustically robust than others.

In English, research indicates that vowel acquisition is already completed by the third year, consonants such as /m/, /n/, /p/, /b/, /d/ and /k/ are the first ones to be picked up and /dh/, /th/, /s/ and /r/ are among the last ones to be mastered. Consonant clusters involving /s/ and /r/ are especially challenging, with /fr/, /spl/, /pr/, /thr/ and /spr/ posing the longest lasting difficulty, continuing beyond the age of ten (McLeod and Arciuli, 2009). Table 2.1 presents a summary of consonant acquisition inventories compiled from five normative studies in the English language, either American or British. The phonetic alphabet used is based on the CMU phoneme set, which is created for American English, however it can be applied in this purpose as phonemic representation of consonants is the same in American and British English. It becomes apparent that there are certain discrepancies between studies, regarding the estimated age of acquisition for each sound, however these can be ascribed to differences in the criteria used in each case to define acquisition. In (Templin, 1957) and (Arlt and Goodban, 1976) the minimum percentage of subjects to correctly pronounce a phoneme for it to be considered as acquired, was set to 75%, with each phoneme in question appearing in initial, medial or final word position. In (Olmsted, 1971) the

minimum percentage was set to 50% and the phonemes' word positions were either initial or final. Finally, (Smit et al., 1990) and (Dodd et al., 2003) applied a 90% criterion, the former using words where the target phonemes appeared either in initial or final positions and the latter including medial positioning as well. As a consequence, acquisition ages extend to older groups for the studies with higher percentage criteria. The order in which different types of consonants are acquired though, follows the general trends described previously. By the age of nine the latest, phonological acquisition is thought to be completed.

TABLE 2.1: *Phonological Acquisition Inventory.*

Age/Study	**Templin (1957)**	**Olmsted (1971)**	**Arlt & Goodban (1976)**	**Smit et al. (1990)**	**Dodd (2003)**
< 2;0		/n//f//hh/ /p//b//k/ /m/ /g/			
2;0 - 2;6			/m//n//ng/ /p//b//t/ /d//k//g/ /f//w//hh/		
2;6 - 3;0	/m//n//ng/ /p/ /f/ /w//hh/	/t//d//v/ /s//z/ /w//y/			
3;0 - 3;6	/y/		/v/	/m//n//hh/ /w//b//p/ /d//k//f/	/p//b//t/ /d//k//g/ /m//n/ /ng//f//v/ /s//z//hh/
3;6 - 4;0	/b//d/ /k//g/ /r/	/ng/ /ch/ /jh//r//l/	/ch//jh//l/ /s//z//zh/	/t/ /g/	/w//l//y/
4;0 - 4;6	/ch/ /sh/	/th//dh/ /sh//zh/	/sh/		/ch/ /zh//jh/
4;6 <	/t//jh//v/ /l//th//dh/ /s//z//zh/		/th//dh/ /r/		
5;0 - 5;6				/y/	/sh/
5;6 - 6;0				/l/ /f/ /v/	
6;0 - 6;6					/r/
6;6 - 7;0				/dh//jh/ /ch//sh/	
7;0 <					/th//dh/
8;0				/th//r/	
7;0 - 9;0				/ng//s//z/	

2.5 Phonological Processes

The constant phonological development that children are undergoing throughout the speech acquisition period, creates disfluencies and hesitation phenomena which eventually recede with age (Potamianos and Narayanan, 2007). It has been hypothesized that part of the maturational procedure is the development of an internal representation system, which will aid the systematic organization of phonological rules emerging from the perception of adult word productions and create a type of word template which will guide their reproduction (Vihman, 1996). This way there are four distinct word forms included in the language acquisition discourse; the adult pronounced form, the child's perceived form, the child's underlying form and the child's spoken form. The child's underlying form combines productive and perceptual elements and is used to differentiate between cases where the adult form is perceived differently from the way it is reproduced (Ingram, 1974). For example, two different words may elicit identical reproductions, however there may be evidence that their perceived forms are distinct. The use of underlying representation models offers an approximation of the evolving stages of language maturation. According to relevant research, during those stages many sounds might be omitted, assimilated or substituted and until their grammatical mapping gets settled, several distortions of the target adult sound will occur (Lust, 2006). Therefore, apart from missing the adult target acoustically through pitch or utterance duration, children also mispronounce or alter words due to lack of full conceptual grasp of it. Such mispronunciations follow specific patterns and are identified by speech experts with the term phonological processes.

Phonological processes were first termed by Stampe in 1969 (Stampe, 1969) and according to his definition they are consistent simplifications of speech used by children in their attempt to imitate the adult target sound. They are universal, since they can be employed by any child regardless of their native language, and they are structured on a hierarchy of more or less basic types with varying levels of persistence, until their

TABLE 2.2: *Categorization of Phonological Processes.*

Substitution Processes		
Stopping	Fricatives, and occasionally other sounds, are replaced with a consonant	sea → tea
Fronting	Velar and palatal consonants tend to be replaced with alveolar ones	goat → doat
Gliding	A glide /w/ or /y/ is substituted for a liquid sound, i.e. /l/ or /r/	lap → yap, ready → weady
Vocalization	A vowel replaces a syllabic consonant	apple → appo
Vowel Neutralization	Vowels tend to be changed into oral and often centralized vowels, i.e. /aa/ or /ah/	hug → hag
Deaffrication	Modification of the affrication feature	cheese → sheeze
Fricative Simplification	A fricative, /th/ is substituted for another fricative, i.e. /f/	three → free
Assimilatory Processes		
Voicing	Consonants tend to be voiced when preceding a vowel, and devoiced at the end of a syllable	tiny → diny, bed → bet
Consonant Harmony	In $C_1VC_2(X)$ contexts, consonants tend to assimilate to each other in certain predictable ways	duck → guck, tub → bub
Progressive Vowel Assimilation	An unstressed vowel will assimilate to a preceding (or following) stressed vowel	flower → flawa, hammer → hamma
Syllable Structure Processes		
Cluster Reduction	A consonant cluster is reduced to a single consonant	clown → cown, dress → dess
Deletion of Final Consonants	A CVC syllable is reduced to CV by deleting the final consonant	bike → bi, more → mo
Deletion of Unstressed Syllables	An unstressed syllable is deleted, especially if it precedes a stressed syllable	banana → nana, potato → tato
Reduplication	In a multi-syllabic word, the initial CV syllable is repeated	water → wawa

eventual and gradual fadeout. Ingram offers an extensive account of the different observed processes, which is summarised in Table 2.2. His analysis is based on a direct comparison of children's mispronunciations against the adult word model, pointing out the mismatch between them. This way, phonological processes are divided into

three main categories with several subcategories each: substitution processes, assimilatory processes and syllable structure processes (Ingram, 1979). Table 2.2 also features two substitution processes not mentioned by Ingram, namely deaffrication (Dodd et al., 2003) and fricative simplification (Cohen and Anderson, 2011). The examples used to outline each error pattern involve one type of process per word, however it is possible and quite common that children's mispronunciations can combine two or more processes in a single word.

Similarly to phonological acquisition age norms, chronology of phonological processes has been the subject of normative research seeking to determine the ages at which each of them begins to disappear. A process is defined as present in a specific age group, if more than 10% of the children in that group exhibit this process. The detail level in which processes are categorized varies among studies, but basic ones such as stopping, fronting or gliding are investigated by most. Figure 2.1 features a comparison of results from three studies on phonological process development (Grunwell, 1981, Dodd et al., 2003, Cohen and Anderson, 2011). Overall it appears that there is minor disagreement in the cases of stopping, fronting, cluster reduction and gliding, however this can be attributed to the considerably different sample sizes used in each study, with (Grunwell, 1981) being under 10, (Cohen and Anderson, 2011) being 94 and (Dodd et al., 2003) being 684. Nevertheless, the combined conclusion from this comparison is that no phonological process is expected to continue beyond the age of six.

2.6 Adult level speech competence

As suggested by the literature mentioned so far, in English, speech acquisition is finalised between seven and nine years of age, with phonological processes already being vanished by the sixth year. This is a widely established linguistic norm, which has proven to be very effective as a guide for diagnosing developmental speech disorders. However, completion of speech sound acquisition does not necessarily amount to adult level speech competence. There is evidence from several domains of speech

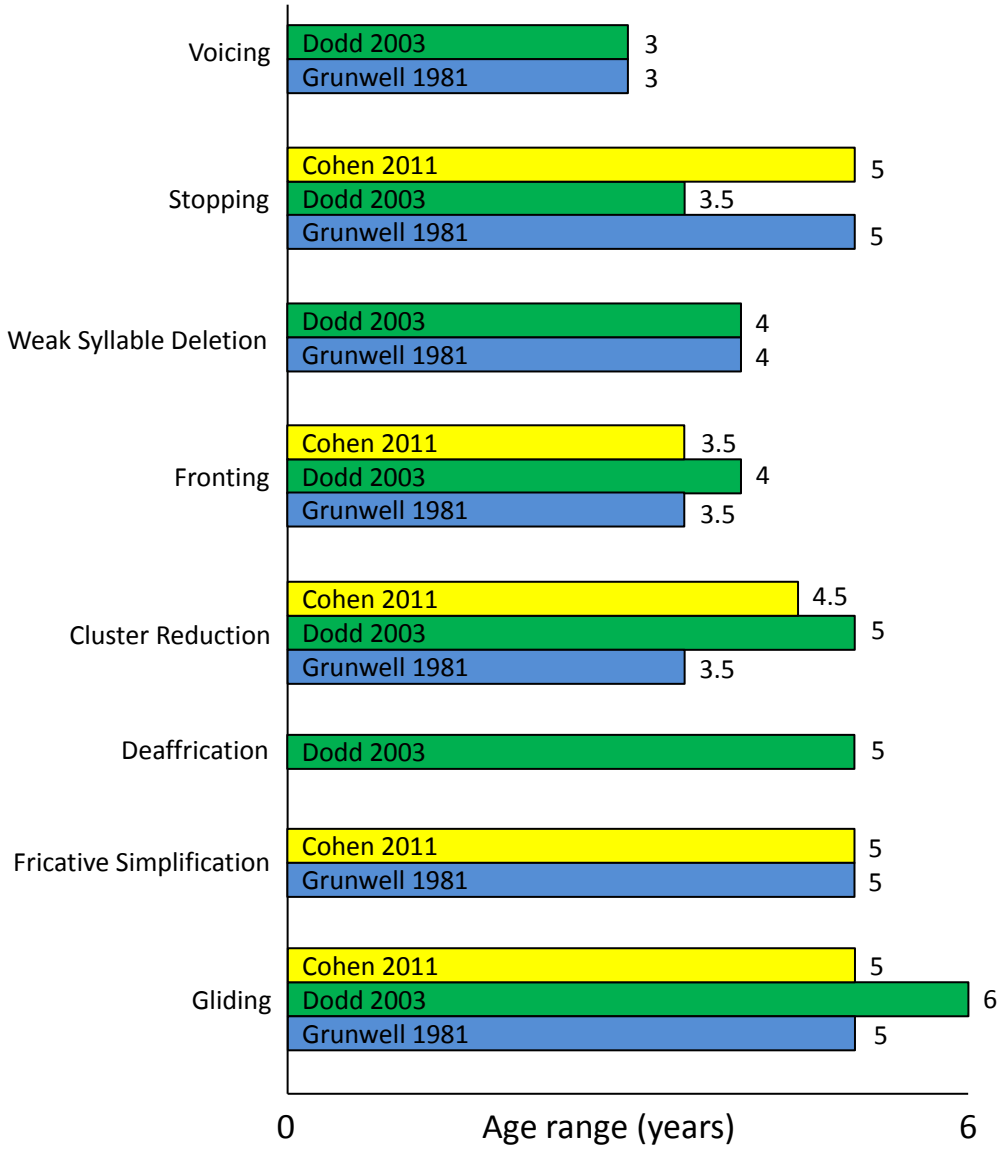


FIGURE 2.1: *Chronology of phonological processes barchart. Yellow bars represent (Grunwell, 1981, green bars represent Dodd et al., 2003, and blue bars represent Cohen and Anderson, 2011).*

development research indicating that even after acquisition, speech mechanisms continue their maturing processes well into adolescence. This is manifested both in terms

of speech perception, as well as in terms of articulatory motor control.

In order to achieve adult phonological competence children have to go beyond their innate acoustic discrimination skills mentioned before (Eimas et al., 1971 Jusczyk and Derrah, 1987), and learn to organise sound patterns consistently into phonemic categories which will enable linguistic decision making (Simon and Fourcin, 1978). Phonemic categorization is thus a key aspect of adult speech. In (Romeo, Hazan, and Pettinato, 2013) children's (9-14 year-olds) and adults' productions were compared with respect to within category dispresion, between category distance and discriminability of phonemic categories for two common contrasts in English language, /s/-/sh/ and /p/-/b/. With the exception of discriminability, which was found to reach adult-like values between 13 and 14 years, both other measures showed that children's data differed largely from adults' even for the oldest children in the sample. Interestingly enough, the different children's age subgroups did not exhibit an age effect and did not predict category distance or dispersion, suggesting that narrowing of variability into adult-like category structure probably takes place between 14 and 18 years. This is in accordance with earlier work by Flege and Eefting (Flege and Eefting, 1986), who examined the /t/-/d/ contrast and found that 9-year-old native English speakers realised /d/ with lead voice onset time significantly more often than English adults. In combination with their finding that the perceptual boundary of the same contrast took place at significantly longer values for adults than for 9, 11, 13 or 17-year-olds, the case was strongly made for speech sound category maturation to continue through adolescence.

In support of this theory are results from (Hazan and Barrett, 2000), where children aged between 6 and 12, as well as adults, were tested on a perceptual categorisation task across four phonemic contrasts: /k/-/g/, /d/-/g/, /s/-/z/, /s/-/sh/. Not only was there a strongly significant age group effect with categorisation consistency increasing between 6 and 12 years and between 12 years and adulthood, but also children were not as consistent as adults when faced with stimuli that had limited acoustic cue information. This indicates that children as old as 12 have not yet reached adult level flexibility in their perceptual strategies. In fact, evidence from (Johnson, 2000)

confirms this claim and extends the age of maturation above 15 years. According to their findings, when exposed to reverberation or noise, children's consonant identification ability becomes adult-like approximately at 14 years, while when exposed to reverberation and noise combined, their identification ability remains below adult level at the age of 15, leading to the conclusion that maturation is finalised in the late teenage years.

At the same time, evidence from kinematic studies of oral articulators reveals protracted development of speech motor control and coordination during adolescence (Walsh and Smith, 2002, Smith and Zelaznik, 2004). More specifically, in (Walsh and Smith, 2002), motion tracking of upper lip, lower lip and jaw was conducted for adolescents (12-, 14-, and 16-year-olds) and young adults, offering a comparison of spatiotemporal consistency for the resulting trajectories. Data from all three adolescent age groups differed significantly from those of the adult group, manifesting higher variability in articulatory trajectories, longer duration of speech segments, lower velocities and smaller movement amplitudes than the adults. A similar paradigm was applied over a larger age span (4-22 years), investigating the development of speech coordinative processes (Smith and Zelaznik, 2004). An age effect was again observed, with younger speakers' functional synergies being significantly less consistent than that of older speakers. Adult level performance was again obtained in late adolescence. Such findings suggest that still at age 16, speech motor control development is not yet completed.

As far as neuromuscular maturation of the speech apparatus is concerned, acoustic studies can provide insightful information. For example, fundamental frequency values can be used to outline the development of laryngeal adjustments for vowels, formant frequency values can convey an image of the development of the positioning of the articulators for vowels and voice onset times for stops can describe articulatory-laryngeal coordination. In (Kent, 1976), a comprehensive literature review on acoustic studies is analysed with reference to their anatomical and neuromuscular implications. The main conclusion drawn from this analysis is that adult-like stability of speech motor control is obtained a few years after phonological acquisition, namely in

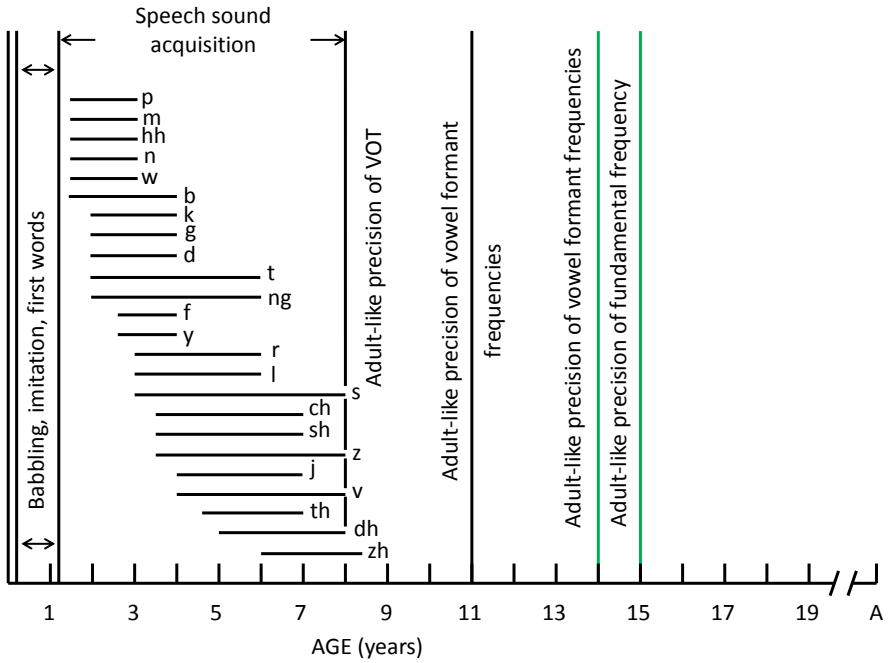


FIGURE 2.2: *Chronological profile of various aspects of speech maturation. The horizontal lines represent consonant acquisition ages as found in (Sander, 1972). They begin at the median age of articulation for each sound and end at the age at which that sound is acquired by 90% of the children. The black vertical lines in the centre of the figure represent the ages at which children attain adult-like precision in the control of voice onset time (VOT) for stops and the control of vowel formant frequency precision, both as found in (Eguchi and Hirsh, 1969). The green vertical lines in the mid-right section of the figure represent the ages at which children acquire adult-like precision of formant frequencies according to (Lee, Potamianos, and Narayanan, 1999). 'A' at the end of the X-axis stands for 'adult'.*

the beginning of puberty, at approximately 12 years. More recently, a thorough analysis of temporal and spectral characteristics of children's speech was conducted using a substantially sized sample (Lee, Potamianos, and Narayanan, 1999), and similarly pointed to puberty for adult-like competency. As reported, vowel duration variability reaches adult level around age 12, /s/ duration variability does so at the age of 13 and sentence duration variability does so after the age of 14. Fundamental frequency hits adult level at 15 years, while all formant frequencies and spectral envelope at 14

years. Figure 3.2 presents a chronological profile of various aspects of speech maturation as featured in (Kent, 1976), with the addition of two green vertical lines representing the ages of adult-like precision of formant frequencies according to (Lee, Potamianos, and Narayanan, 1999). This illustrates clearly that maturational processes are not exhausted in the phase of phonological acquisition, but continue their progression throughout adolescence. Most essentially, it demonstrates that speech sound acquisition is one among several separate stages of child speech development, until adult-like precision is reached. This way, the distinction between speech acquisition and adult speech competence becomes clear.

Taking into account this information, we hypothesize that even after speech sound acquisition, even after phonological processes have faded away, there is still a quality difference between children's and adults' speech, which bears some vestige of phonological effects associated with language acquisition (PEALA). This vestige, even though indiscriminable to a human listener might still be detected by ASR systems and cause phone confusions. The term 'phonological effects associated with language acquisition' (PEALA) is a novel term introduced in this thesis to describe all the normative patterns of mispronunciation or deviation from the adult target sound that can be encountered throughout speech maturation, before adult level competency is reached. They follow the motifs of language acquisition phenomena such as phonological processes, but unlike them, PEALA are not confined by acquisition age norms, instead they continue on a decreasing trajectory until speech maturation is complete. It is hypothesized that the presence of these effects even though subtle, is sufficient to interfere with the training and testing processes of an ASR system. Emerging from this hypothesis, the two research questions of the thesis are: Can PEALA be detected in systematic patterns of ASR phone confusion errors? and, Can PEALA be evidenced in systematic patterns of acoustic feature structure?

Unlike other work done in the field aiming to correlate developmental changes to ASR errors (Hämäläinen et al., 2014), or indirectly investigate the effect of developmental changes with pronunciation modelling (Shivakumar et al., 2014b), this thesis approaches the topic systematically and presents a comprehensive analysis of a wide

range of developmental factors that could have an impact in children's speech (and hence children's ASR). This way, not only are all the possible and expected developmental speech patterns defined, but also the motivation of our hypotheses is explained and supported by a solid body of literature.

The method chosen to address the two research questions is by defining a set of possible ASR phone substitutions that can be predicted from PEALA and investigating their systematicity and in the acoustic models and phone confusion errors of the ASR systems built for the purpose of this thesis. These are introduced in Table 2.3 and will be the central point of reference in the analysis of the experimental results reported in the present study. The choice of these particular substitution pairs has been made based on the substitution patterns defined by phonological processes and has been informed by the various examples featured in relevant studies (Grunwell, 1981, Dodd et al., 2003, Cohen and Anderson, 2011). Their effect in ASR is expected to be most apparent in the youngest age groups and gradually fade as speakers' ages increase. The measures chosen to identify the potential effect of PEALA in ASR involve the use of confusion matrices and a statistical significance test, which are further described in Chapter 5.

TABLE 2.3: *Predictable Substitutions based on PEALA.*

Voicing	Stopping	Fronting
$/p/ \rightarrow /b/$	$/s/ \rightarrow /t/, /v/ \rightarrow /b/$	$/k/ \rightarrow /t/$
$/t/ \rightarrow /d/$	$/f/ \rightarrow /p/, /th/ \rightarrow /p/$	$/g/ \rightarrow /d/$
$/k/ \rightarrow /g/$	$/jh/ \rightarrow /d/, /v/ \rightarrow /p/$	$/g/ \rightarrow /t/$
$/s/ \rightarrow /z/$	$/ch/ \rightarrow /t/, /dh/ \rightarrow /d/$	$/sh/ \rightarrow /s/$
	$/sh/ \rightarrow /t/, /s/ \rightarrow /th/$	
Deaffrication	**Fricative Simplification**	**Gliding**
$/ch/ \rightarrow /sh/$	$/th/ \rightarrow /f/$	$/r/ \rightarrow /w/$
$/jh/ \rightarrow /zh/$		$/r/ \rightarrow /l/$
$/ch/ \rightarrow /k/$		$/l/ \rightarrow /w/$
$/zh/ \rightarrow /z/$		$/l/ \rightarrow /y/$

Chapter 3

Speech Corpora

This chapter features detailed descriptions of the speech corpora used in the automatic speech recognition experiments (ASR). These involve both children's and adults' data from native speakers of American and British English. All the speakers included have no recorded history of speech impairment and particularly in the case of children speakers, they were all marked as typically developing never having received any speech therapy. The two first corpora presented next, WT and Copycat, were especially collected and provided for the purpose of this project by Disney Research in Pittsburgh (DRP).

The 40 phone set of the CMU pronunciation dictionary was used for the American English speech annotations and the 45 phone set of the BEEP pronunciation dictionary was used in the transcriptions of British English data. TIMIT adults' speech corpus was initially transcribed with the use of a 61 phone set, which for the presentwork was mapped down to the 40 phone set of the CMU dictionary. Appendix A illustrates how the involved phone sets relate to each other.

3.1 Children's Speech Corpora

3.1.1 WT

The WT corpus was collected by Disney Research and features speech from 60 students attending the Winchester Thurston kindergarten and primary school in Pittsburgh, Pennsylvania. It is divided in five age groups, from 5- to 9-year-olds. All

recordings took place at school premises, the majority in typical classroom environment, with the exception of the kindergarteners' (5-year-olds') recordings, which were carried out in a quiet classroom environment. This disparity in the background quality between the 5-year-olds' data and the rest of the groups' data is reflected in the ASR experiments as reported in chapter 5, with the youngest group displaying slightly better performance than expected.

Speech was elicited with the use of an interactive animation app, which was designed specifically for this purpose. It consisted of 15 surveys of 3 multiple choice questions each, which were presented to the children on an ipad, through animation characters prompting them to repeat their preferred choice for each question in order to create a story narrative. All possible responses were purposefully selected to make up a phonologically balanced set of utterances. Figure 3.1 shows an example of the animations used. Sound was recorded through the built-in microphone of the ipad.

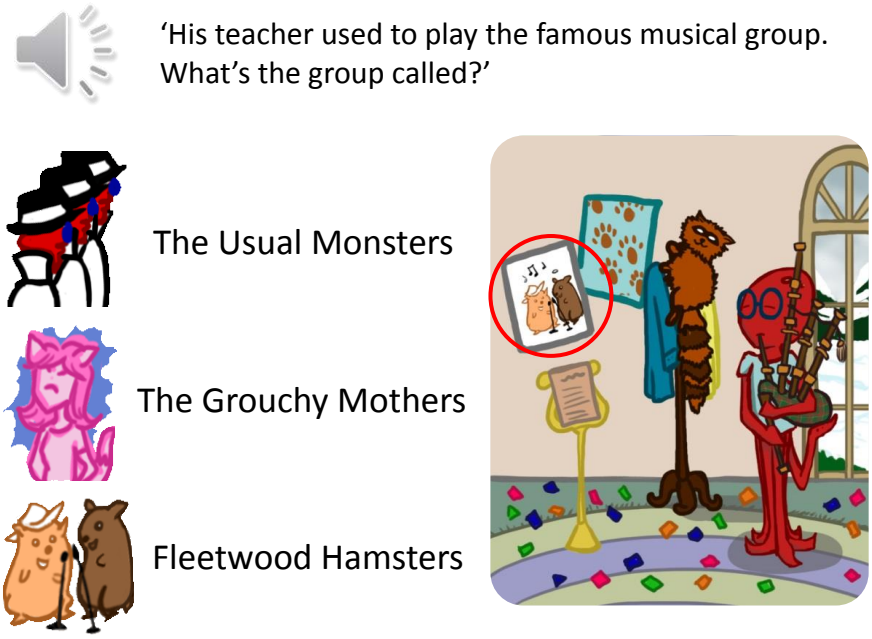


FIGURE 3.1: *WT animation example. A question prompts the children to choose a band name for the main character, three options are made available, and children have to repeat their preferred option after the prompt message.*

After collection, the recordings were transcribed manually both at the word and at

the phone level according to the 40 phone set of the CMU pronunciation dictionary. The annotators had no formal training in phonetics before this task. After removing responses that did not contain one of the given alternatives, the final set consisted of approximately 2337 phonologically balanced utterances, each extending between one and six words. The total duration of the data set was only 57 minutes, therefore its application in ASR has been conducted through 14 cross validation experiments, where data from one survey were used as the test set and the rest of the data as the training set (data from survey 3 were not used due to discrepancies in the transcriptions).

TABLE 3.1: *WT corpus.*

Age Group	**# Speakers**	**# Utterances**	**Duration**
5-year-olds	10	422	9:21 mins
6-year-olds	16	478	12:06 mins
7-year-olds	14	502	12:25 mins
8-year-olds	13	498	12:37 mins
9-year-olds	17	437	10:35 mins

3.1.2 Copycat

Copycat is another children's speech corpus collected by Disney Research and was provided in the present study with the intention to complement the limited sized WT. A total of 61 Pennsylvanian students, belonging in the same age range as those in WT, were recruited to the Disney Research lab in Pittsburgh where the recordings took place in a quiet environment with the use of a microphone.

The speech material was a subset of WT consisting of 17 phonologically balanced sentences featuring key word combinations from the multiple choices of the WT surveys. Children were prompted to repeat each sentence after the experimenter with the help of animation stimuli similar to those of WT. A few indicative examples are presented in figure 3.2.

The data were transcribed manually at the word level, again by annotators who

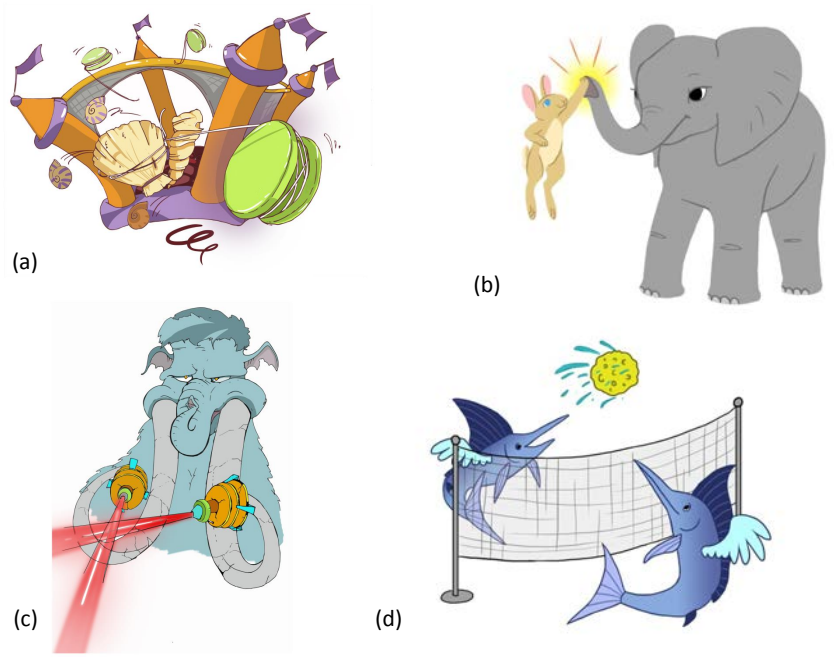


FIGURE 3.2: *Copycat animation examples. (a) Seashells and yoyos in a bouncing castle. (b) A rabbit's foot and an elephant trunk. (c) A space ray gray behemoth's tusk. (d) Flying swordfish and a wet sponge ball.*

had not received any formal training in phonetics. Phone level transcriptions were extracted, through the application of forced alignment to the obtained word level transcriptions, with the use of the 40 phone set of the CMU pronunciation dictionary. The resulting data set comprised 1349 utterances, each extending between three and eight words, with total duration of 63 minutes. Likewise WT, Copycat was used in ASR through 3-fold cross validation experiments.

TABLE 3.2: *Copycat corpus.*

Age Group	**# Speakers**	**# Utterances**	**Duration**
5-year-olds	11	282	12:13 mins
6-year-olds	10	226	10:45 mins
7-year-olds	18	358	17:26 mins
8-year-olds	12	256	12:00 mins
9-year-olds	10	227	10:42 mins

3.1.3 PSR

The PSR (Primary School Reading) corpus contains speech from 11 five- and six-year-old primary school students from Worcester, England (Russell et al., 2000). It was recorded during the development of a computer-based pronunciation tutor for primary school children. All recordings took place in a quiet mobile classroom through a head-mounted microphone. Children were asked to repeat single words from an approximately 1000-word long vocabulary. The selection of the words featured in the PSR vocabulary was completed in collaboration with teachers, language experts and speech therapists, in order to be deemed appropriate for five- and six-year-olds.

The corpus is divided in two subsets, PSR1 and PSR2. PSR1 comprises data from 5 children who were judged to be fluent by their teachers, while PSR2 comprises data from 6 children whose speech varied from good to poor. Manual word level annotations were created, marking poorly pronounced words where necessary. For the present work, 4924 utterances from PSR1 and 814 utterances from PSR2 were automatically transcribed using forced alignment on the word level annotations, according to the 45 phone set of the BEEP pronunciation dictionary.

3.1.4 CSLU

The CSLU Kids' speech corpus contains spontaneous and read speech collected at the Northwest Regional School District near Portland, Oregon (Shobaki, Hosom, and Cole, 2000). It includes data from over 1100 speakers spanning from five- to fifteen-year-olds. The speech material consists of 205 isolated words, 100 sentences and 10 numeric strings as well as a few minutes of spontaneous speech per speaker.

The data collection was carried out with the use of the CSLU toolkit [1] and a head-mounted microphone. Each utterance was prompted on a computer screen both in text form and through an animated 3D character. Each repetition was played back to

[1] http://cslu.ogi.edu/toolkit/

TABLE 3.3: *PSR corpus.*

	PSR1		
Speaker ID	**pronunciation judgement**	**# Utterances**	**Duration**
PSR1_A	Very Good	1069	13.31 mins
PSR1_B	Very Good	1042	13.25 mins
PSR1_C	Very Good	982	13.41 mins
PSR1_D	Very Good	918	13.59 mins
PSR1_E	Very Good	913	11.12 mins
Total		4924	64.68 mins
	PSR2		
Speaker ID	**pronunciation judgement**	**# Utterances**	**Duration**
PSR2_A	Very Good	211	3.69 mins
PSR2_B	Very Good	59	1.12 mins
PSR2_C	Average	62	1.17 mins
PSR2_D	Average	41	0.79 mins
PSR2_E	Poor	55	0.94 mins
PSR2_F	Very Poor	386	5.96 mins
Total		814	13.67 mins

the speaker and the experimenter and if judged necessary, it was re-recorded. Elicitation of spontaneous speech was achieved through a series of questions that the experimenter directed towards the speaker.

After data collection, the recordings were further processed by two independent assessors in order to verify their quality. A three-point scale resulted from this procedure: '1 - good' utterances (where only the target word or sentence is said clearly and intelligibly with no significant background noise or extraneous speech), '2 - questionable' utterances (which were intelligible but accompanied by other sounds as well) and '3 - bad' utterances (where the target word was missing or pronounced unintelligibly). Word level orthographic transcriptions were produced and provided with the corpus.

For the purposes of the present work, only read utterances from the first category ('1 - good') were used, amounting to 47,532 utterances in total. They were randomly divided in the test and train set, with 451 speakers in the former and 665 speakers in the latter set. Table 3.4 offers a detailed distribution of the data into age groups. Phone transcriptions were produced automatically, according to the 40 phone set of

TABLE 3.4: *CSLU corpus.*

	Train Set		
Age Group	**# Speakers**	**# Utterances**	**Duration**
5-year-olds	52	1472	87.4 mins
6-year-olds	53	2780	184.9 mins
7-year-olds	68	3404	223.1 mins
8-year-olds	69	4038	275.1 mins
9-year-olds	54	3065	190.3 mins
10-year-olds	59	3404	176.9 mins
11-year-olds	67	2190	83.5 mins
12-year-olds	58	1884	71.1 mins
13-year-olds	59	1969	74.4 mins
14-year-olds	64	2209	84.7 mins
15-year-olds	62	2033	77.6 mins
Total	665	28448	25.48 hours
	Test Set		
Age Group	**# Speakers**	**# Utterances**	**Duration**
5-year-olds	36	994	63.7 mins
6-year-olds	36	1885	124.3 mins
7-year-olds	46	2398	158 mins
8-year-olds	46	2648	176.4 mins
9-year-olds	37	1987	120.8 mins
10-year-olds	40	2297	116.9 mins
11-year-olds	45	1359	50.4 mins
12-year-olds	39	1301	47.7 mins
13-year-olds	40	1310	48.4 mins
14-year-olds	44	1546	57.9 mins
15-year-olds	42	1359	49.9 mins
Total	451	19084	16.9 hours

the CMU pronunciation dictionary, using forced alignment applied to the word level transcriptions that are provided with the data.

3.2 Adults' Speech Corpora

3.2.1 TIMIT

TIMIT (Texas Instruments/Massachusetts Institute of Technology) is a phonologically balanced American English speech corpus covering all 8 major dialect regions of the

U.S.A. (Garofolo et al., 1993). It contains data from 630 speakers producing 10 utterances each, thus amounting to 6300 utterances and approximately 5 hours of speech in total. For the purposes of the present study, a substantial subset of TIMIT was utilised, comprising a total of 4288 utterances produced by 536 different speakers and spanning over 3.6 hours.

The TIMIT speech material is divided in 3 categories of sentence types: 'SX' (450), 'SI' (1890), and 'SA' (2) sentences. 'SX' sentences, are hand-designed in order to be phonetically compact and inclusive of all possible phone pairs, especially those with particularly interesting or challenging phonetic contexts. 'SI' sentences, are phonetically diverse and hand-picked from existing sources, such as stage play dialogues. Finally, 'SA' sentences, are specifically designed to evoke possible dialect variations among the different speakers. The data subset used in this study only features sentences from the 'SX' and 'SI' categories.

All data were recorded in a noise isolated recording booth through a headset mounted microphone. The sentences were elicited through on-screen prompts that the speakers were instructed to read out loud in a 'natural' voice. In the event of any detected mispronunciations in a sentence, it was discarded and re-recorded.

Word and phone level transcriptions were produced manually and then automatically time aligned. Any errors in the resulting phone boundaries were hand-corrected by experienced phoneticians. The phone label set used in this procedure consists of 61 phonemes, but in order to fit in and be comparable with the rest of the corpora employed in this study, it was mapped down to the CMU 40 phone set.

3.2.2 SCRIBE

SCRIBE [2] (Spoken Corpus Recordings In British English) approximates a British version of TIMIT, including data from four U.K. dialect regions: South East, Glasgow, Leeds and Birmingham. It is divided in two subsets, the 'Few Talker' set containing speech from 30 speakers and the 'Many Talker' set containing speech from 120 speakers.

[2] https://www.phon.ucl.ac.uk/resource/scribe/

The corpus was recorded in an anechoic chamber with the use of a close talking microphone and a half-inch B&K condenser microphone, with suitable auditory feedback to the speaker. It includes both read and spontaneous speech. The read speech comprises 200 'phonetically rich' sentences, 460 'phonetically compact' sentences (which were appropriated for the British accent after TIMIT's compact sentences) and a two-minute long accent diagnostic passage. The spontaneous speech consists of constrained 'free speech' elicited through a paradigm where the speaker had to give an oral description of a picture.

For the purposes of this study, only data from 13 speakers with Birmingham accent were utilised, as it was judged to best match the Worcester accent of the PSR speakers. A set of 1654 utterances of both 'phonetically rich' and 'phonetically compact' sentences were processed. With the use of the 45 phone BEEP set, phone level transcriptions were produced automatically with forced alignment of the orthographic sentence-level annotations which were available with the corpus. Due to the limited amount of the data, ASR was conducted through a 13 fold cross validation series of experiments, each time using the data from one speaker as the test set and the rest of the data as the training set.

3.3 Anticipated outcome from the use of the different corpora

In the description of the WT corpus provided by Disney Research, it is mentioned that the individuals who carried out the corpus annotations had no formal training in phonetics. On the contrary, the annotations for TIMIT were carried out by experienced phoneticians. Equivalent information about the rest of the corpora does not apply, as they only provided word level transcriptions. This way, a direct comparison between corpora based on the proficiency of their annotators is not possible. However, it should be noted that as far as phone level transcriptions are concerned, not using trained phoneticians for coding children's speech in particular could be problematic, as any linguistically significant mispronunciations (probably caused by phonological processes) could be omitted or overseen due to the annotator's lack of expertise.

Given that some of the corpora used in this work feature speech from children as young as five years old, a lot of mispronunciation errors are expected as a result of phonological processes or in the case of older children as a manifestation of PEALA. The fact that all the corpora involved contain phonetically balanced material, renders them sufficient to elicit all the phonological phenomena in question. All substitution pairs from Table 2.3 are expected to be exhibited in these corpora, possibly at different frequencies depending on the type of substitution and the speaker's age. All types of PEALA related errors are expected to be more frequent in the age groups where phonological processes are still present (i.e. five and six-year-olds), while past those ages PEALA related error types are expected to fade out in the same order that phonological processes are reported to fade (i.e. voicing errors become less frequent faster while gliding or fricative simplification persist longer). Due to the fact that the available manual phone transcriptions of our data are not carried out by speech experts and the automatically extracted ones result from dictionary annotations of the word level transcriptions, it cannot be confirmed whether all the anticipated phenomena will be accurately captured and documented.

Chapter 4

ASR Systems

Below follows a description of the ASR systems used on this project. For each of the corpora described in Chapter 3, a customised ASR system was developed specifically for the purposes of addressing the research questions. The chapter is structured primarily into five categories, depending on which speech corpus was used to train each system, and then further in several subcategories depending on various factors such as whether each system is GMM-HMM based or DNN-HMM based. Training sets in children's systems were age independent due to the limited amount of available data. All systems performed phone level recognition.

4.1 WT ASR Systems

A tied-state triphone GMM-HMM-based ASR system was developed, based on the CMU phone set, using the HTK toolkit (Young et al., 1997). The speech was downsampled from 44.1 to 12 kHz and transformed into sequences of 39 dimensional feature vectors, comprising 12 mel frequency cepstral coefficients (MFCCs) plus $C0$, augmented with the corresponding Δ and Δ^2 parameters. A fourteen-fold cross-validation experiment was conducted, in which 13 surveys were used for training and the remaining one for testing (survey 3 was not used in the study due to annotation discrepancies). The system had approximately 700 physical states, each associated with a 32 component Gaussian mixture model (GMM). A 'flat' phone-loop grammar, in which each phone bigram is equally probable, was used in recognition. The number of GMM components, language model scale factor and word insertion penalty were

optimised on survey 15. This system scored an average phone accuracy of 37% across the 14 surveys. A summary of the phone recognition results across age groups is presented in Table 4.1.

TABLE 4.1: *Phone Accuracy Results from WT corpus.*

Age Groups	**% Accuracy**	**% Correct**	**# Deletions**	**# Substitutions**	**# Insertions**
5-year-olds	35.6	40.8	822	1594	212
6-year-olds	31.2	37.7	1052	1919	313
7-year-olds	35	40.4	1147	2013	287
8-year-olds	40.8	46.3	852	1915	285
9-year-olds	45.3	50.1	748	1597	226

In agreement with the relevant literature, the results showed progressive improvement from the youngest to the eldest age group. Minor exception to this trend was the performance on 5-year-olds' speech, which exceeded that on the 6- and 7-year-olds' data by 4.4 % and 0.6 % respectively. However, this can be attributed to the fact that, as mentioned in Chapter 3, 5-year-olds' data collection was carried out in a quiet environment, unlike the rest of the dataset which was collected in natural classroom environment. This way the 5-year-old group ended up with less noisy recordings and this contributed to the accuracy of their recognition. Besides this minor dissonance in the observed age effect, it is noteworthy that this system produced immensely low accuracy scores. A major boost in the performance accuracy was obtained after the application of a triphone language model to the recogniser, built directly from triphone frequency counts in the available phone level transcriptions. However, given the limited vocabulary that it was built upon and the use of triphones, it employed very tight context-related constraints to the recognition, which may have prevented the phone substitution errors of interest from occurring. As described in Chapter 1, a language model determines how probable a sequence of phones is, given the statistical information provided in the training data. Given that our training data is very limited, the resulting statistics will allocate very high probabilities to only a very limited amount of phone sequences, this way bypassing the acoustic model and masking any potential PEALA related effect. In order to address the research questions, such

a method that could potentially mask the impact of PEALA on the recogniser does not seem appropriate and should be avoided. Therefore, the results produced with the language model, which are shown in Table 4.2, were not further analysed and this type of language model was not further used in the next systems.

TABLE 4.2: *Phone Accuracy Results from WT corpus with the use of a triphone Language Model.*

Age Groups	% Accuracy	% Correct	# Deletions	# Substitutions	# Insertions
5&6-year-olds	60	62.9	1953	1334	253
7-year-olds	68	69.9	950	649	101
8&9-year-olds	79.1	80.9	967	916	177

4.2 Copycat ASR System

A tied-state triphone GMM-HMM-based ASR system was developed in a procedure similar to the one applied for WT, using the HTK toolkit (Young et al., 1997). All data were again downsampled from 44.1 to 12 kHz and transformed into sequences of 39 dimensional feature vectors, comprising 12 mel frequency cepstral coefficients (MFCCs) plus $C0$, augmented with the corresponding Δ and Δ^2 parameters. A 3-fold cross validation method was applied in the building of the recogniser. A 128 component Gaussian mixture model (GMM) was associated with the system , based on phone level accuracy optimisation. Following the conclusions drawn on language model use from the WT system, a 'flat' phone-loop grammar was applied. Table 4.2 features the relevant phone accuracy results.

TABLE 4.3: *Phone Accuracy Results from CopyCat corpus.*

Age Groups	% Accuracy	% Correct	# Deletions	# Substitutions	# Insertions
5-year-olds	31.5	39.6	944	1518	331
6-year-olds	39.9	48.5	737	1179	320
7-year-olds	42.3	50.1	1315	1902	501
8-year-olds	43.8	50.6	949	1332	311
9-year-olds	42.1	48.6	906	1195	266

A clear age effect can be observed as accuracy is increasing with age. Age-specific accuracy scores show slight improvement compared to the corresponding WT scores, however the overall percentage of accuracy does not rise above 40 %.

4.3 PSR ASR Systems

As described in Chapter 3, PSR consists of two subsets, PSR1 and PSR2, based on the competency of their featured speakers. PSR1 was used to train models which were employed in recognition of both PSR1 (with cross validation) and PSR2 test sets. These will be referred to as PSR1 and PSR2 systems, depending on which subset of the PSR corpus was used as a test set for the PSR1-trained recogniser. The choice of PSR1 alone as a train set was made on the grounds that it contained speech of better quality, as verified by the speakers' teachers, which would presumably produce more precise acoustic models.

A tied-state triphone GMM-HMM-based recogniser was built with the use of the HTK toolkit (Young et al., 1997). All recordings were downsampled from 20.01 to 12 kHz and transformed into sequences of 39 dimensional feature vectors, comprising 12 mel frequency cepstral coefficients (MFCCs) plus $C0$, augmented with the corresponding Δ and Δ^2 parameters. A 64 component Gaussian mixture model was associated with the system based on phone level accuracy optimisation. Instead of training a language model, a 'flat' phone-loop grammar was implemented. Testing on PSR1 data was carried out by applying a five-fold cross validation method where each time data from four speakers were used as the train set and data from the remaining speaker were used as the test set. Data from all five speakers of PSR1 were used to train the recogniser that all PSR2 speakers were tested on. Table 4.4 presents the average phone accuracy results for both PSR subsets and Table 4.5 presents the average phone accuracy per speaker for PSR2.

The highest phone accuracy was obtained for PSR1, whose speakers were judged to have good pronunciation, while PSR2, whose speakers' pronunciations varied, obtained a score lower by roughly 10%. The fact that the acoustic model for these systems

TABLE 4.4: *Phone Accuracy Results from PSR corpus.*

PSR Subset	% Accuracy	% Correct	# Deletions	# Substitutions	# Insertions
PSR1	50.1	57.4	2312	6621	1531
PSR2	39.8	46.9	521	1358	250

was trained on data with verified good level of fluency was reflected in the resulting accuracies, as both PSR1 (by far) and PSR2 outperformed WT and Copycat with respect to the corresponding age groups. Moreover, the recordings for PSR were made in a more benign environment than either WT or Copycat, and consequently produced higher ASR accuracies.

The results for each individual speaker from PSR2 are overall consistent with the teachers' judgements on the children's fluency. PSR2_A and PSR2_B, whose pronunciations were characterised as 'very good', had the second and third best accuracy scores, while PSR2_E and PSR2_F, whose pronunciations were characterised as 'poor' and 'very poor' respectively, were among the bottom three accuracy scores.

TABLE 4.5: *Individual Phone Accuracy Results from PSR2 speakers.*

Speaker ID	% Accuracy	% Correct	# Deletions	# Substitutions	# Insertions
PSR2_A	42.5	51.4	139	304	81
PSR2_B	41.9	45.2	77	122	12
PSR2_C	52.7	57.1	54	99	16
PSR2_D	33.9	39.5	51	90	13
PSR2_E	35.7	42.7	39	99	17
PSR2_F	36	43.8	161	644	111

4.4 CSLU ASR Systems

The CSLU corpus was used to build four ASR systems, two being GMM-HMM based and two being DNN-HMM based, all with the use of the Kaldi Speech Recognition Toolkit (Povey et al., 2011). A 'flat' phone-loop grammar was used in each case. All systems were trained and tested on the train and test subsets described in Chapter 3.

GMM1: This is the first, baseline recogniser. The speech was transformed into sequences of 39 dimensional feature vectors comprising 12 mel frequency cepstral coefficients (MFCCs) plus C0, augmented with the corresponding Δ and Δ^2 parameters. The system used 1951 physical states each associated with a Gaussian mixture model (GMM) whose components were chosen from a set of 15,050 shared Gaussian PDFs.

GMM2: This recogniser was developed from 'GMM1' after applying maximum likelihood linear transform (MLLT), linear discriminant analysis (LDA) and speaker adaptive training (SAT) to obtain 40 dimensional feature vectors. The resulting system had 1997 physical states each associated with a Gaussian mixture model (GMM) whose components were chosen from a set of 15,022 shared Gaussian PDFs. Finally, forced alignment was applied to obtain an alignment between the data and the 1997 senones (GMM-HMM states). This alignment was passed to the next stage.

DNN1: The initial DNN-HMM system was built using the previously trained 40 dimensional fMLLR features, which were created in the SAT stage of 'GMM2' development, and the alignment from the GMM2 system. The inputs to the DNN were feature vectors in context, with a context of ± 5 frames. The number of hidden layers used was 2 and the hidden layer dimensions were 1024. Thus the DNN can be characterized as $440 \times 1024 \times 1024 \times 1997$. DNN parameter estimation used 6 iterations of state-level minimum Bayes risk (sMBR) training.

DNN2: Alignments from 'DNN1' were used to train a new DNN which included a 9 dimensional bottleneck layer in addition to the two existing 1024 dimension hidden layers. The extracted bottleneck features (BNFs) were used instead of fMLLR features to train another DNN recogniser, following the same procedure as for 'DNN1'. The choice of 9 dimensions for the BNFs was based on the results presented in (Bai et al., 2015), which show that ASR performance using 9 BNFs is comparable with the phone recognition accuracy obtained with standard 39 dimensional MFCC-based feature vectors.

The average phone accuracy of each recogniser over all age groups is displayed in Table 4.6. As one would expect, the results are poorer than published phone recognition accuracies for adult speech. In addition, the difficulty of recognising children's

TABLE 4.6: *Overall Phone Accuracy Results per CSLU system.*

System	% Accuracy	% Correct	# Deletions	# Substitutions	# Insertions
GMM1	54.1	61.0	22566	56488	16366
GMM2	59.8	65.9	20252	48982	14516
DNN1	64.4	68.3	22331	42029	9342
DNN2	61.6	64.9	24533	46703	8561

speech will have been compounded by the use of a 'flat' phone-level grammar. As expected, the DNN systems outperformed the GMM ones, with DNN2 scoring slightly lower than DNN1. This slight difference in performance is most likely due to the fact that the various parametres for DNN2 were not optimised, and could have been potentially balanced out, had it been deemed to serve a purpose for this study. However, as the research questions of this work are focusing on interpreting the different error types produced by the ASR systems and not perfecting their performance, a discrepancy of 2.8% between the two DNN systems was not considered significant. Inclusion of all four systems in the results analysis is purposefully chosen, so that there can be a comparison between the number and types of confusions produced by systems of different configurations. This is important for the first research question as a systematic pattern of phone substitutions could potentially be revealed.

The results for each individual age group presented in Figure 4.1, confirmed the classification of the four systems and showed the anticipated gradual improvement in phone recognition accuracy with increasing age, starting from 56.6% for the first group in DDN1 and reaching 67% for the last group in the same system. According to the initial hypothesis, this difference in performance will be reflected in the phone confusion matrices of each system, which will be analysed in the next chapter. PEALA related phone substitutions are expected to be found in larger proportions in the younger speakers' matrices across systems and gradually reduce until they reach minimum proportion in the oldest age groups. The advantage of analysing confusion matrices from a great age range and four differently configured ASR systems is that it enables the emergence and confirmation of a systematicity in the substitutions of

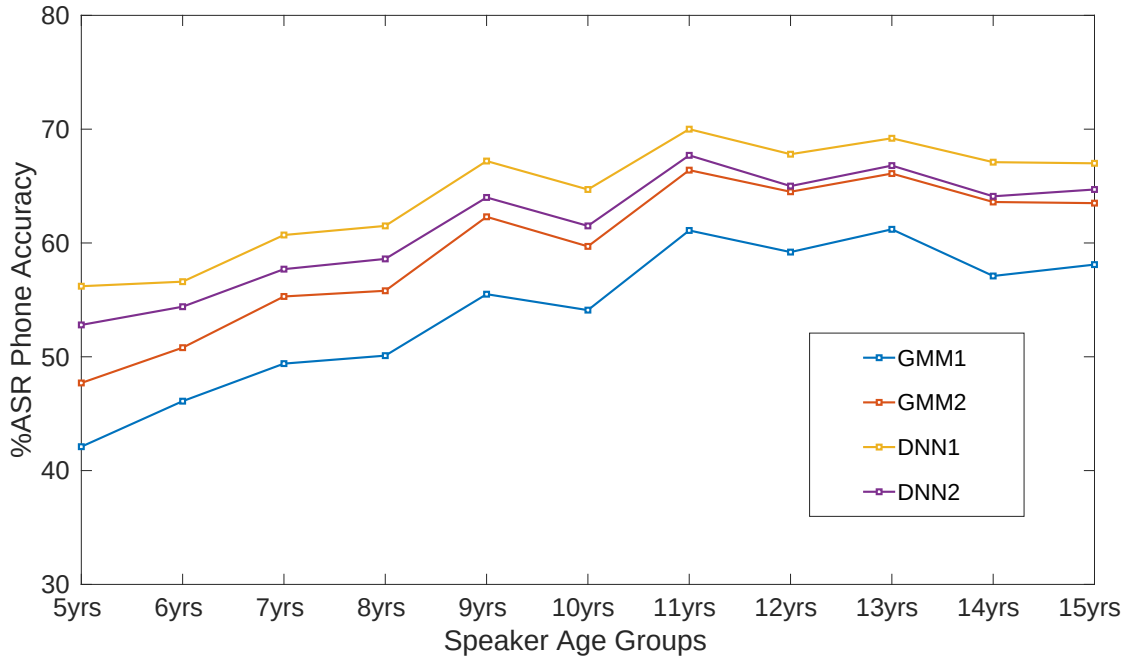


FIGURE 4.1: *CSLU percentage phone accuracy as a function of age*

interest, or lack there of.

4.5 TIMIT ASR Systems

TIMIT was used to train and test five types of ASR systems. The first was meant to be comparable to the one built on WT, with the use of the HTK toolkit (Young et al., 1997), while the other four were constructed in the same manner as the CSLU recognisers, with the use of Kaldi Speech Recognition Toolkit (Povey et al., 2011). A 'flat' phone-loop grammar was applied in all five systems. The speech was downsampled from 16 to 12 kHz and the TIMIT labels were mapped onto the CMU phone set in all cases. The resulting systems are categorised under the names: TIMIT, TIMIT_GMM1, TIMIT_GMM2, TIMIT_DNN1 and TIMIT_DNN2.

TIMIT: The system had 1445 physical states, each associated with an 8 component GMM. The train and test sets used were identical to the standard ones provided by the corpus, with the exception of the dialect specific sentences ('SA'), which were removed.

Without a grammar this system scored a phone accuracy of 57% on the full TIMIT test set. This was an anticipated result with reference to the corresponding WT phone accuracy (37%), confirming the claim that all things being equal, adults' ASR systems reach higher accuracies than children's systems. Nonetheless, by comparison with other adults' systems implementing state of the art techniques, this baseline recogniser was of mediocre performance. The following systems were an attempt to gradually raise the TIMIT accuracy score to mainstream levels.

TIMIT_GMM1: This is the equivalent of DNN1 built on the CSLU corpus, in this case with 1834 physical states and a pool of 15033 shared Gaussian PDFs.

TIMIT_GMM2: Again this is equivalent to the way DNN2 was developed for the CSLU corpus, with 1922 physical states and a set of 15019 shared Gaussian PDFs.

TIMIT_DNN1: As DNN1 was developed for CSLU, this system had 2 hidden layers with 1024 hidden units and can be characterised as $440 \times 1024 \times 1024 \times 1922$.

TIMIT_DNN2: Finally, this system follows the paradigm of DNN2 for CSLU, with 9 dimensional bottleneck features as input and a characterisation of $99 \times 1024 \times 1024 \times 1922$.

Table 4.7 demonstrates the phone accuracy scores for each system. The increase in recognition accuracy from the previous TIMIT system is evident in all four occasions, with TIMIT_DNN1 being at the top with 78.6%. All recognisers follow the same performance pattern as those for CSLU, ranking from GMM1 with the lowest accuracy, to GMM2 with a slightly higher accuracy, to DNN2 with the second best score, until DNN1 which is the best scoring system for both corpora. The difference in performance between each pair of competing recognisers is approximately steady at 15% (16.8 between GMM1 and TIMIT_GMM1, 15.7 between GMM2 and TIMIT_GMM2, 14.2 between DNN1 and TIMIT_DNN1 and 13.9 between DNN2 and TIMIT_DNN2), again confirming the hypothesis that adults' speech corpora deliver better results than children's in systems of the same configuration.

TABLE 4.7: *Overall Phone Accuracy Results per TIMIT system.*

System	% Accuracy	% Correct	# Deletions	# Substitutions	# Insertions
TIMIT_GMM1	70.9	74.8	1413	4195	879
TIMIT_GMM2	75.5	78.5	1345	3437	670
TIMIT_DNN1	78.6	82.7	874	2977	909
TIMIT_DNN2	75.5	79.3	1136	3484	836

4.6 SCRIBE ASR System

The SCRIBE corpus was used to train and test one GMM-HMM based recogniser, following a method comparable to the one applied for the PSR corpus with the use of the HTK toolkit (Young et al., 1997). Speech was downsampled from 44.1 to 12 kHz and transformed into sequences of 39 dimensional feature vectors, comprising 12 mel frequency cepstral coefficients (MFCCs) plus $C0$, augmented with the corresponding Δ and Δ^2 parameters. A 128 component Gaussian mixture model was associated with the system based on phone level accuracy optimisation. Instead of training a language model, a 'flat' phone-loop grammar was applied. Testing was carried out by applying a thirteen-fold cross validation method where each time data from twelve speakers were used as the train set and data from the remaining speaker were used as the test set.

This system produced a recognition accuracy of 44% failing to exceed one of the two children's systems it was meant to be compared against (PSR1 - 50%) and surpassing the other one (PSR2 - 40%) only by 4%. This performance was unexpected from a recogniser built on adults' speech, however the limited amount of data utilised in its development might have primarily contributed to this outcome.

With the exception of the performance of the system built on SCRIBE corpus, all recognisers produced expected accuracy results; every system built on TIMIT showed better performance than the corresponding children's corpus it was compared against. The amount of available data played a role in the performance (for example WT and Copycat were outperformed by GMM1 built on CSLU), as well as the type of HMM system (DNN systems outperformed GMM ones). An anticipated age effect was observed in all children's systems, with performance gradually improving as speakers' ages increased. Additionally, speaker competency appeared to be a predicting factor for ASR accuracy, as PSR speakers whose speech was judged as 'very good' by their teachers achieved higher ASR scores than PSR speakers whose pronunciation was judged as 'poor'.

After an extensive research on the different results previously published from all the corpora involved in this project, no comparable accuracy norms were found. It appears that the present work was the first one to use WT and Copycat in phone level ASR, and looking at the papers that followed, no one is performing phone level ASR. CSLU offers two different word error rate scores in the paper that comes with the dataset, and PSR is used to recognise proficiency of speech. SCRIBE is an incomplete set, so there are no results available. TIMIT is the only standardised corpus, which is widely used, and has many available accuracy results. However, these are not included in this section as it would not offer any perspective to only compare against one corpus's accuracy norms.

The next chapter explores the prominence of PEALA related phone substitutions in the confusion matrices produced by each of the systems described here. It is expected that PEALA related errors will adhere to the trends observed it the current chapter, proportionately affected by factors such as speaker age and competency, as well as system configuration.

Chapter 5

Confusion Analysis

The purpose of this chapter is to investigate whether the phonological effects associated with language acquisition (PEALA) defined in Chapter 3, are reflected in the performance of ASR through systematic error patterns and if so, to what extent this occurs. In order to establish a measure for substitution patterns within the speech data, phone confusion matrices were extracted from the recognisers described in Chapter 5 and utilised in a comparison between children's and adults' phone substitution errors. A statistical significance test is proposed to identify substitution errors in the children's data that cannot be explained by the expected variation in the adult data. The resulting errors are then analysed to determine if they can be attributed to speech developmental factors.

5.1 The effect of different types of phone-level annotation

An issue that has to be addressed before the analysis of ASR phone confusion takes place, is the type of phone-level annotation that is available for the different data sets. Ideally one would have accurate time-aligned phone-level annotations. In this case, differences between the true annotation and an annotation obtained from a word pronunciation dictionary would indicate pronunciation errors (PEALAs), while differences between the true annotations and the ASR outputs would indicate true phone recognition errors.

Preceding the analysis of ASR phone confusions with respect to PEALA, a comparison between the available handmade phone level annotations of the children's data

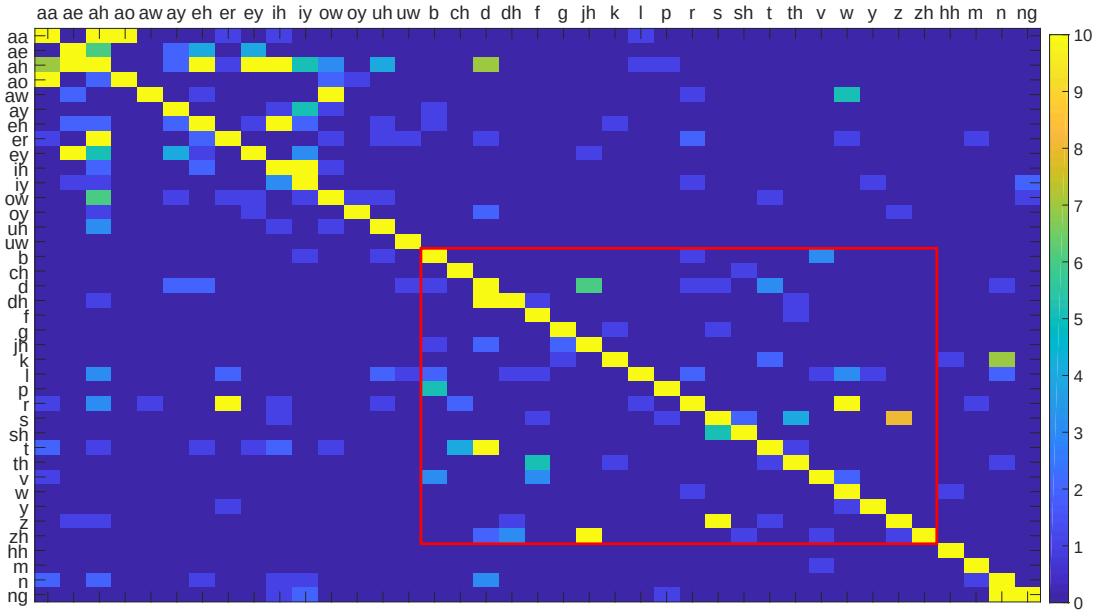


FIGURE 5.1: Confusion matrix: Annotators vs Dictionary. The y-axis of the figure represents the phone occurrences in the dictionary transciptions of the test material for the 5-year-old speakers of WT. The x-axis represents how each of the phones in the y-axis was transcribed by the annotators. The phone labels are ordered in groups so that vowels appear first, consonants which are involved in the substitutions in Table 2.3 follow, and consonants which are not involved in the substitutions in Table 2.3 are last. The red rectangle marks the area in which PEALA related errors are expected to occur.

and their pronunciation dictionary transcriptions was performed in order to ascertain that the former captured reliably the phonological variations in children's speech. Thus, several matrices were computed displaying the confusion between manual annotations from the WT corpus and their corresponding dictionary transcriptions. In general very little evidence of confusion was observed for all age groups, including the cases where substitutions would be expected due to PEALA, and despite the empirical judgement of the experimenter who collected the dataset and had reported that a lot of developmental mispronunciations had taken place. In fact it seems that the data annotations were in close alignment with the pronunciations suggested by the dictionary. Figure 5.1 illustrates the lack of confusion between the human and dictionary annotations in the case of the youngest (5-year-old) age group of WT.

The observed low percentage of confusions in the manual transcriptions suggests

that the annotators in this particular data collection were strongly influenced by what they expected to hear. Surely more data are required before any generalization is inferred from this observation, however it is hypothesized that annotators without any formal phonological training might fail to distinguish between an accurate pronunciation of a phoneme and a PEALA related substitution of it, especially in the cases where the two phonemes in question are acoustically similar. Unfortunately, accurate phone-level transcription of children's speech requires skilled phoneticians and is prohibitively expensive for large amounts of data.

In the experiments described in this thesis, the annotations of the children's recordings are based on a pronunciation dictionary (apart from those of WT which were hand transcribed but turned out to be close to dictionary based). Thus, an observed ASR phone substitution could be a genuine ASR error, or the result of a child pronunciation error, or a combination of both. We rely on the statistical significance test described in the following section, to factor out genuine phone substitution errors that are not due to PEALA.

5.2 A test for statistical significance

The premise of this Chapter is that some ASR phone confusions for children's speech will be attributable to phonological factors associated with language development. Conversely, the null hypothesis is that all such errors can be explained as random variations of errors that occur in ASR for adults. To test this hypothesis a model of phone confusion in adult ASR is needed.

Given a set of K examples of the i^{th} phone ϕ_i, it is assumed that the classification of the set is governed by a multinomial distribution whose parameters are the $N = 39$ probabilities $p_{i,1}, p_{i,2}, ..., p_{i,N}$ in the i^{th} row of the adult phone confusion matrix. If $|\phi_i \rightarrow \phi_j|$ denotes the number of occurrences of the phone substitution $\phi_i \rightarrow \phi_j$, the probability $p(|\phi_i \rightarrow \phi_j| = k)$ that k of the ϕ_is are recognised as ϕ_j follows the

corresponding marginal distribution, which is binomial with parameters $p_{i,j}$ and K:

$$p(|\phi_i \rightarrow \phi_j| = k) = \frac{K!}{k!(K-k)!} p_{i,j}^k (1 - p_{i,j})^{K-k} \tag{5.1}$$

With these assumptions it is possible to decide whether a particular set of errors in child ASR can be attributed to a random variation of the pattern of errors observed for adults, or is significantly different. Specifically, k misclassifications of ϕ_i as ϕ_j in phone recognition of children's speech is judged to be significantly large (i.e. very unlikely to occur as often in adult recognition) if the (cumulative) probability $P(|\phi_i \rightarrow \phi_j| \geq k)$ of k or more misclassifications of ϕ_i as ϕ_j, based on the adult reference, is less than 0.05. In the latter case the errors are characteristic of children and may be due to developmental factors. Similarly, k misclassifications of ϕ_i as ϕ_j is significantly small if $P(|\phi_i \rightarrow \phi_j| \leq k) \leq 0.05$.

This way the test for statistical significance takes into account the fact that phone substitutions are common in ASR experiments, and only considers the occurrence of a substitution to be significant if it occurs more frequently than would be expected as a random variation of the reference data.

5.3 Phone substitutions which are predictable from PEALA and the proportion of them which occur significantly more often than in adult speech

Based on the assumption that phone confusions in ASR can reflect linguistic confusions in the speakers' data, phone confusion matrices were extracted from each ASR system described in Chapter 5 and used to investigate whether any of these confusions could have been predicted from PEALA. Figure 5.2 demonstrates the high confusion output of the WT system on 5-year-olds' test data. The initial hypothesis is that as children grow older, their speech quality improves, thus their productions are expected to induce fewer ASR confusions that are related to PEALA. Similarly, children of the same age whose pronunciation has been judged as good by their teachers, are expected

to cause fewer PEALA related ASR confusions than their peers whose pronunciations have been deemed as poor.

The set of 27 phone substitution pairs presented in Table 2.3 was used as a reference for determining which substitutions were predictable from PEALA, these will be referred to as 'predictable' substitution errors. The statistical significance test described in section 5.2 was applied to all the confusion matrices, in order to isolate the substitutions that occurred significantly more often for children than for adults. These will be referred to as 'significant' substitution errors. The main point of interest in the analysis of the results, is the proportion of substitution errors that are predictable and significant at the same time.

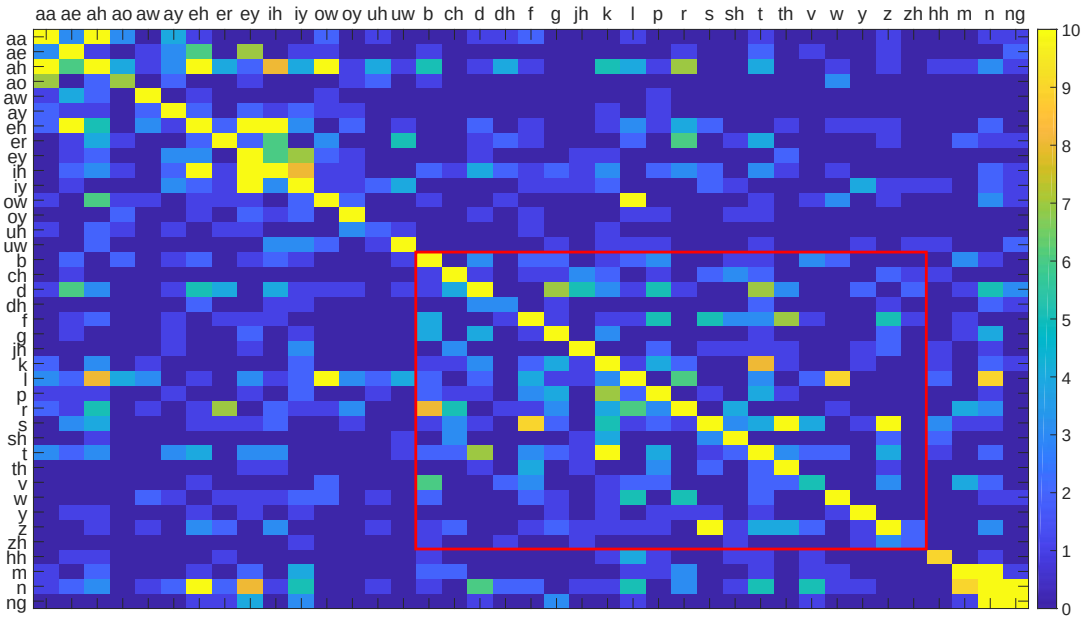


FIGURE 5.2: Confusion matrix: Annotations vs ASR. The y-axis of the figure represents the phone occurrences in the manual annotations of the test material for the 5-year-old speakers of WT. The x-axis represents how each of the phones in the y-axis were recognised by the ASR system. The phone labels are ordered in groups so that vowels appear first, consonants which are involved in the substitutions in Table 2.3 follow, and consonants which are not involved in the substitutions in Table 2.3 are last. The red rectangle marks the area in which PEALA related errors are expected to occur.

An important issue in this work is the fact that the phone-level transcriptions are

obtained from forced alignment using baseform transcriptions from a word pronunciation dictionary. If a child makes systematic pronunciation errors, then these will occur in both the training and the test set. To address this issue, the assumption was made that if there are children in the training set who exhibit a particular phonological effect, then the models for the corresponding phones will be corrupted. For example if a child uses $/t/$ for $/k/$, the $/k/$ phone models will tend to be more $/t/$-like and so there will be an increase of $/t/ \rightarrow /k/$ substitutions in the test. To cater for that implication, we looked at both directions of confusion for each of the 27 effects. Depending on the different analyses applied on the results, two measures were calculated. Firstly, the probability of each substitution error predictable from PEALA was calculated as the ratio of the number of confusions in both directions of each pair, over the total number of confusions for the phones involved:

$$P(p_A \rightarrow p_B) = \frac{(p_A \rightarrow p_B) + (p_B \rightarrow p_A)}{\sum_{i \in \{A \neq i\}}^{N} (p_A \rightarrow p_i) + \sum_{i \in \{B \neq i\}}^{N} (p_B \rightarrow p_i)} \tag{5.2}$$

Moreover, the total percentage of predictable confusions within a dataset was computed as the ratio of the total number of predictable substitutions in the dataset over the sum of all the substitutions occurring in the same set:

$$\%\text{Predictable Substitutions} = \frac{\sum_{(A,B) \in PEALA} ((p_A \rightarrow p_B) + (p_B \rightarrow p_A))}{\sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} (p_i \rightarrow p_j)} \tag{5.3}$$

$$\text{where } PEALA = \{(A, B) : p_A \rightarrow p_B \text{ is a PEALA related substitution}\}$$

The same method (equation 5.3) was applied in calculation of the percentage of predictable confusions which are significantly more frequent in children's data.

5.4 Results

5.4.1 PEALA related substitutions across data sets

The results demonstrating the probability of each predictable substitution within the various ASR systems developed on American English corpora, are displayed in Table 5.1. The highlighted values indicate which predictable substitutions were significantly more frequent in the children's than in the adults' data (TIMIT), according to the binomial test for statistical significance. At first glance, the majority of substitutions seems to be significant for all six systems, supporting the hypothesis that errors related to PEALA occur at a significantly higher rate in children's data. Indeed there seems to be a high consistency across the six different ASR systems built on three different speech corpora. However, this type of errors only represents a small proportion of the total number of phone substitution types which were found significant after the binomial test. In WT, the number of significant predictable substitution types (i.e. /p/→/b/ is one type of predictable substitution, /b/→/p/ is another one, /t/→/d/ is another type and so on) is 39 over a total of 660 random substitution types (i.e. /aa/→/ah/ or /t/→/ah/ or /t/→/f/ are random substitution types that do not follow a specific developmental pattern). For CopyCat this ratio is $\frac{22}{532}$, for GMM1, GMM2, DNN1 and DNN2 the corresponding ratios are $\frac{33}{1170}$, $\frac{39}{1132}$, $\frac{41}{1072}$ and $\frac{39}{1079}$ respectively. It thus becomes less straightforward whether these significant predictable substitutions express the effect of PEALA on the dataset or are just coincidentally occurring as part of random noise. A closer look at their specific occurrence probabilities might offer some insight.

As Table 5.1 illustrates, the probabilities of predictable phone confusions across the six ASR systems range between 0.4% and 47%, but average at roughly 10%. This means that for a given predictable substitution $/p_A/ \rightarrow /p_B/$, the phonemes involved have a 10% chance to be confused with each other, with a 90% remaining chance split among any of the other 38 phonemes of the CMU phonetic alphabet. It is remarkable that this average percentage is approximately the same for all recognisers with only

1% standard deviation between them. If these confusions really represent the linguistic variability in the children's data, then 10% seems to be the rate at which they impact the ASR output.

The substitution category which averages the highest probability is voicing with 23%, then follows fricative simplification with 15%, deaffrication with 11%, fronting with 8%, stopping with 7% and finally gliding with 6%. This ranking clashes with the chronological order in which those phonological processes are reported to disappear in figure 2.1. For example, as mentioned in Chapter 3, voicing appears to be the first process to fade away, but according to the results presented here, it is the most prominent process among the ASR confusions. Similarly, gliding is the longest lasting phonological process, but in the present results it appears to be the least probable confusion category. The five most probable confusions across recognisers are $/s/ \rightarrow /z/$ at 36%, $/ch/ \rightarrow /sh/$ at 27%, $/p/ \rightarrow /b/$ at 22%, $/k/ \rightarrow /g/$ at 18% and $/s/ \rightarrow /th/$ at 16%. They do not seem to be following a particular pattern, since they belong to different categories and involve different classes of phonemes. In general there doesn't seem to be any systematicity in the way individual substitution probabilities are distributed.

TABLE 5.1: *Probabilities of substitutions related to PEALA (FS = Fricative Simplification) in different ASR systems. Highlighted numbers indicate phone substitution rates that are significantly higher than would be expected for adult speech.*

	Substitution	WT	CC	CSLU				Mean
				GMM1	GMM2	DNN1	DNN2	
Voicing	/p/↔/b/	0.11	0.2	0.22	0.29	0.26	0.24	**0.22**
	/t/↔/d/	0.09	0.04	0.17	0.21	0.19	0.16	**0.14**
	/k/↔/g/	0.09	0.15	0.17	0.25	0.24	0.2	**0.18**
	/s/↔/z/	0.28	0.25	0.3	0.45	0.47	0.41	**0.36**
Stopping	/s/↔/t/	0.04	0.02	0.02	0.02	0.01	0.01	**0.02**
	/f/↔/p/	0.1	0.07	0.06	0.07	0.07	0.07	**0.07**
	/jh/↔/d/	0.05	0.07	0.13	0.14	0.14	0.12	**0.11**
	/v/↔/p/	0.01	0.04	0.03	0.03	0.03	0.03	**0.03**
	/ch/↔/t/	0.07	0.07	0.06	0.06	0.06	0.06	**0.06**
	/sh/↔/t/	0.04	0.05	0.01	0.01	0.01	0.01	**0.02**
	/th/↔/p/	0.05	0.05	0.07	0.05	0.04	0.05	**0.05**
	/v/↔/b/	0.1	0.08	0.09	0.08	0.08	0.08	**0.09**
	/dh/↔/d/	0.04	0.004	0.05	0.07	0.07	0.09	**0.05**
	/s/↔/th/	0.11	0.1	0.18	0.21	0.17	0.19	**0.16**
Fronting	/k/↔/t/	0.15	0.08	0.1	0.09	0.09	0.12	**0.11**
	/g/↔/d/	0.12	0.03	0.1	0.1	0.08	0.11	**0.09**
	/g/↔/t/	0.02	0.03	0.02	0.03	0.02	0.02	**0.02**
	/sh/↔/s/	0.11	0.14	0.06	0.07	0.05	0.08	**0.09**
Deaffric.	/ch/↔/sh/	0.19	0.24	0.3	0.27	0.27	0.33	**0.27**
	/jh/↔/zh/	0.13	0.16	0.07	0.05	0.06	0.19	**0.11**
	/ch/↔/k/	0.07	0.03	0.03	0.03	0.02	0.02	**0.03**
	/zh/↔/z/	0.03	0.004	0.02	0.02	0.03	0.02	**0.02**
FS	/th/↔/f/	0.14	0.11	0.2	0.14	0.14	0.14	**0.15**
Gliding	/r/↔/w/	0.06	0.03	0.03	0.04	0.04	0.04	**0.04**
	/r/↔/l/	0.06	0.07	0.03	0.02	0.02	0.03	**0.04**
	/l/↔/w/	0.11	0.1	0.15	0.12	0.12	0.16	**0.13**
	/l/↔/y/	0.01	0.002	0.01	0.1	0.01	0.01	**0.02**

5.4.2 PEALA related substitutions across speaker fluency

Table 5.2 shows the results obtained for subsets of the PSR corpus, namely PSR1 and PSR2 as well as the individual speakers from PSR2. The statistical significance test was applied twice in each subset, once with reference to PSR1, whose speakers were judged to have good pronunciations by their teachers, and once with reference to SCRIBE, which contains adult speech. As observed in Chapter 5, human judgement of pronunciation and ASR accuracy are in agreement for the most part. In the case of PEALA related predictability or significance, the results are less straightforward.

TABLE 5.2: *Phone accuracy (row 2), percentage of errors predicable from PEALAs (row 3) and those which occur significantly more often than for children with good pronunciation (PSR1, row 4) and adults (SCRIBE, row 5), for subsets of PSR.*

	PSR1	PSR2	PSR2_A	PSR2_B	PSR2_C	PSR2_D	PSR2_E	PSR2_F
% Acc.	50.1%	39.8%	42.5%	41.9%	52.7%	33.9%	35.7%	36%
% Predictable	20.2%	19.6%	16.5%	18%	21.2%	17%	21%	21.3%
Sig. (PSR1)	0.0%	7.8%	7.6%	6.6%	2%	0%	0%	12.7%
Sig. (SCRIBE)	14.9%	11.9%	4.0%	10%	2%	0%	13%	15.0%

The percentages of predictable phone substitutions for the different subsets are surprisingly similar to each other, reaching 20.2% for PSR1 and 19.6% for PSR2. One would expect PSR1, a set which consists of speakers of good-only pronunciation and which indeed scored a lower error rate than PSR2 by 10%, to exhibit fewer PEALA related errors than PSR2. In turn PSR2, consisting of speakers of various pronunciation proficiency, would be expected to exhibit a larger proportion of phone substitutions predictable from PEALA than PSR1. The fact that the resulting predictable substitution scores for these two subsets are so close to one another, and actually that PSR1 is on the lead by 0.6%, does not confirm the hypothesis which suggests that as speech pronunciation quality improves, phone substitutions related to PEALA will decrease. Within PSR2, however, individual speakers' results for predictable substitutions follow a more or less expected pattern. Speakers PSR2_A and PSR2_B, who had the best pronunciation of the group, along with PSR2_D, whose pronunciation was average,

reached the bottom three scores in predictable substitution errors, leaving the two poor speakers along with the other average one, in the highest places.

Looking at the significance of predictable substitutions, there can hardly be detected any trend. The comparison using PSR1 as a reference shows that the proportion of predictable errors which are significantly more frequent in PSR2 than in children with good pronunciation is very low and in two cases it is zero. Things get particularly perplexed in the cases of the two poor and one average speakers who had reached high percentages of predictable substitutions, but in this significance test managed to diverge with 12.7%, 0% and 2% respectively. When the same data are compared against the adult reference, more incompatibility emerges. For some speakers these percentages are higher than those extracted from the PSR1 reference (PSR2_B, PSR2_E and PSR2_F), for some they are equally low (PSR2_C and PSR2_D) and for speaker PSR2_A is almost half. Taking into consideration the fact that the PSR2 set contains very limited data from each speaker, with PSR2_A and PSR1_F at the top two durations adding up to 3.69 and 5.96 minutes respectively, and accounting for 70% of the phone errors on PSR2, the analysis could proceed with focusing on these two and ignoring the rest of the PSR2 speakers. Taking this approach, we end up with a simplified system of a good speaker with high phone accuracy and a poor speaker with low phone accuracy, whose percentages of predictable phone substitutions relate to each other as expected (16.5% for PSR2_A and 21.3% for PSR2_F) and whose proportion of significantly predictable errors exhibits large discrepancy against the adult reference (4% for PSR2_A and 15% for PSR2_F). Nevertheless, it still remains unclear why the significance measure for PSR2_A differs more from PSR1 than from adults' speech (7.6% as opposed to 4%).

5.4.3 PEALA related substitutions across age groups

As presented in Chapter 4, all children's ASR accuracy results increase with age. Intuitively one would expect the percentage of errors that are predictable from PEALA, and in particular those that occur significantly more often in children's than in adults' speech, to decrease with age. On the contrary, the proportion of predictable and significant predictable confusions for each of the systems developed in this project, exhibit the opposite effect, namely rising with age. Table 5.3 shows how predictable and significant predictable errors increase across age groups in parallel with the phone recognition accuracy. Evidently, this observation cannot reflect the development of these errors in the speakers' productions, since it is well established that such phone substitution errors typically decline with time and usually disappear by the age of six (3). There must be then, some other reason explaining this unexpected phenomenon, which is not related to the development of the speakers' speech properties, but rather represents the development of the ASR systems across age groups. Thus, it appears that both the predictability and the significance factor are correlated with phone accuracy.

Figures 5.3 and 5.4 show scatter plots of the percentage of substitutions predictable from PEALA and substitutions predictable from PEALA that occur significantly more frequently than for TIMIT, as a function of phone accuracy. The Pearson correlation coefficients between phone accuracy and percentages of predictable substitutions range between 0.82501 and 0.98486, while those between phone accuracy and percentages of significant predictable substitutions range between 0.67283 and 0.9503. Such values indicate a strong correlation between phone accuracy and these two categories of substitution errors. This finding complicates the initial purpose of the data analysis, which was to determine whether the output of ASR on children's speech contains information characteristic of the children's linguistic development. Instead of a correlation between age and PEALA related substitution errors, the obtained results suggest a counter intuitive pattern which can only partially be explained by the correlation between PEALA related errors and phone accuracy.

TABLE 5.3: Phone recognition accuracy, percentage of Errors Predictable from PEALA and those which occur significantly more often than fr adult speech, as a function of age for WT, CoyCat (CC) and the four CSLU systems (GMM1, GMM2, DNN1, DNN2)

		5yrs	6yrs	7yrs	8yrs	9yrs	10yrs	11yrs	12yrs	13yrs	14yrs	15yrs
WT	Acc.	35.6	31.2	35	40.8	45.3						
	Pred.	13.1	11.3	13.1	14.2	16.4						
	Sig.	6.8	4.3	5.1	7.3	11						
CC	Acc.	31.5	39.9	42.3	43.8	42.1						
	Pred.	10.1	11	12.5	14.3	12.3						
	Sig.	3.1	2.5	5.4	6.2	4.4						
GMM1	Acc.	42.1	46.1	49.4	50.1	55.5	54.1	61.1	59.2	61.2	57.1	58.1
	Pred.	10.09	13.16	13.21	12.84	14.39	13.93	16.95	15.76	13.80	17.03	16.96
	Sig.	2.7	7.49	7.34	6.42	6.98	8.01	8.82	9.18	9.35	10.21	12.50
GMM2	Acc.	47.7	50.8	55.3	55.8	62.3	59.7	66.4	64.5	66.1	63.6	63.5
	Pred.	11.65	13.86	14.40	14.76	15.95	15.26	17.93	17.73	18.00	18.27	18.62
	Sig.	5.15	7.67	9.35	8.44	11.32	11.08	14.50	12.69	14.00	12.18	15.65
DNN1	Acc.	56.2	56.6	60.7	61.5	67.2	64.7	70	67.8	69.2	67.1	67
	Pred.	12.15	13.74	14.16	15.07	15.62	14.96	16.81	17.25	17.87	17.03	18.00
	Sig.	4.14	6.80	8.17	10.20	10.07	8.64	11.94	13.92	12.60	13.86	14.66
DNN2	Acc.	52.8	54.4	57.7	58.6	64	61.5	67.7	65	66.8	64.1	64.7
	Pred.	12.76	13.98	14.53	14.30	16.33	15.18	17.53	17.28	15.87	17.41	16.49
	Sig.	5.26	9.81	10.06	9.94	11.77	9.67	11.52	13.23	10.41	12.30	12.37

A possible explanation is that residual effects of PEALA are present in the older children, and that the increase in recognition accuracy with age enables these effects to be seen more clearly. For example, a young child who uses /w/ for /r/ may continue to produce a '/w/-like' /r/ as he or she gets older. This 'mispronunciation' may not be sufficient for a listener to make a categorical decision that the child is exhibiting the /r/→/w/ PEALA, but it may still be sufficient to cause an ASR error. The fact that the proportion of such ASR errors increases as phone accuracy improves, suggests that some other type of random errors, not related to PEALA must be decreasing. This way, by elimination of random noise, the PEALA related substitutions become more prominent. However, this explanation is not conclusive about the role of PEALA in the interpretation of children's ASR.

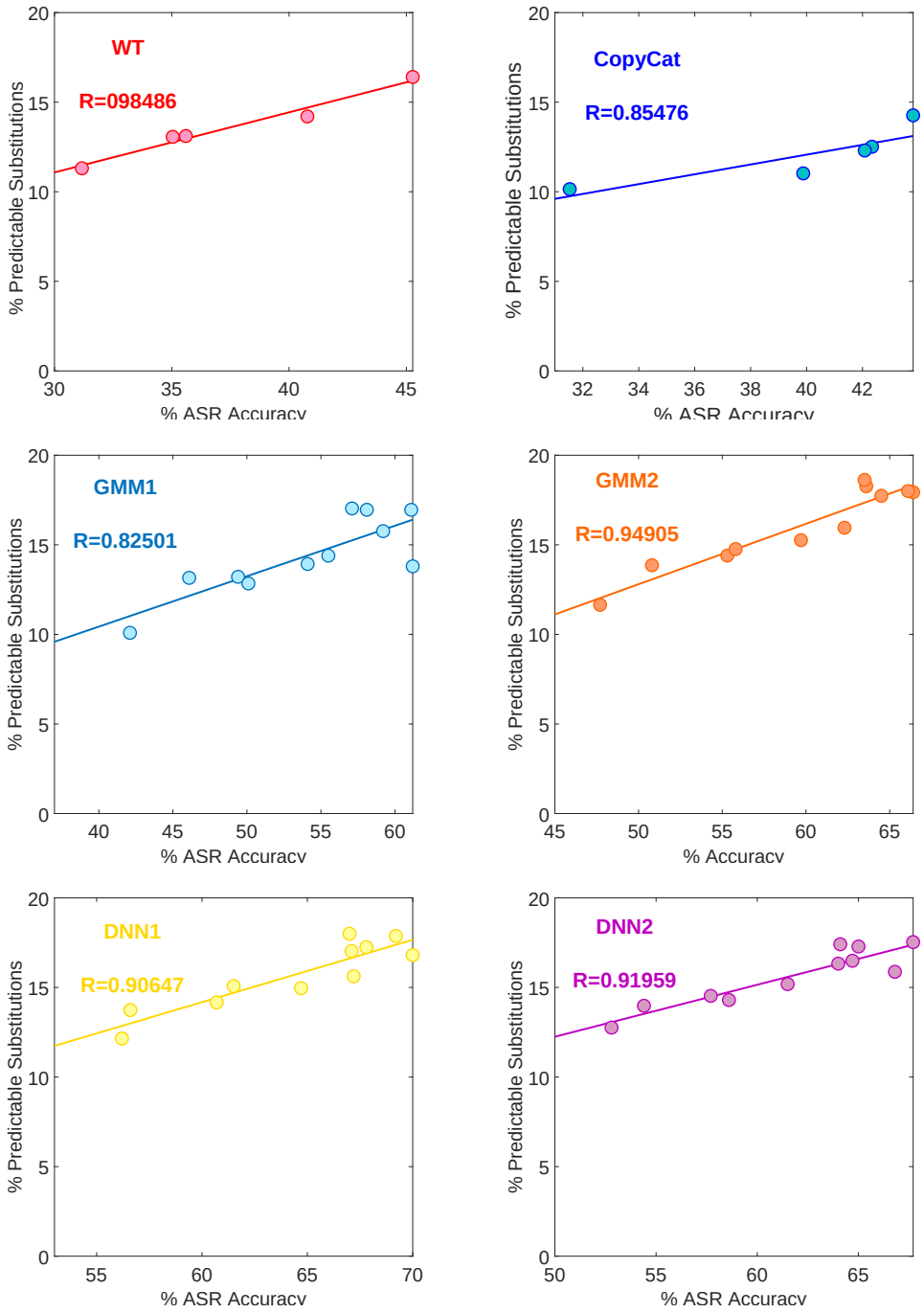


FIGURE 5.3: Scatter plots for WT, CopyCat and all CSLU systems of the percentage of predictable substitutions, as a function of phone accuracy. Each of the dots represents an age group. In all plots, the order in which age groups appear follows the progression of the corresponding ASR accuracy.

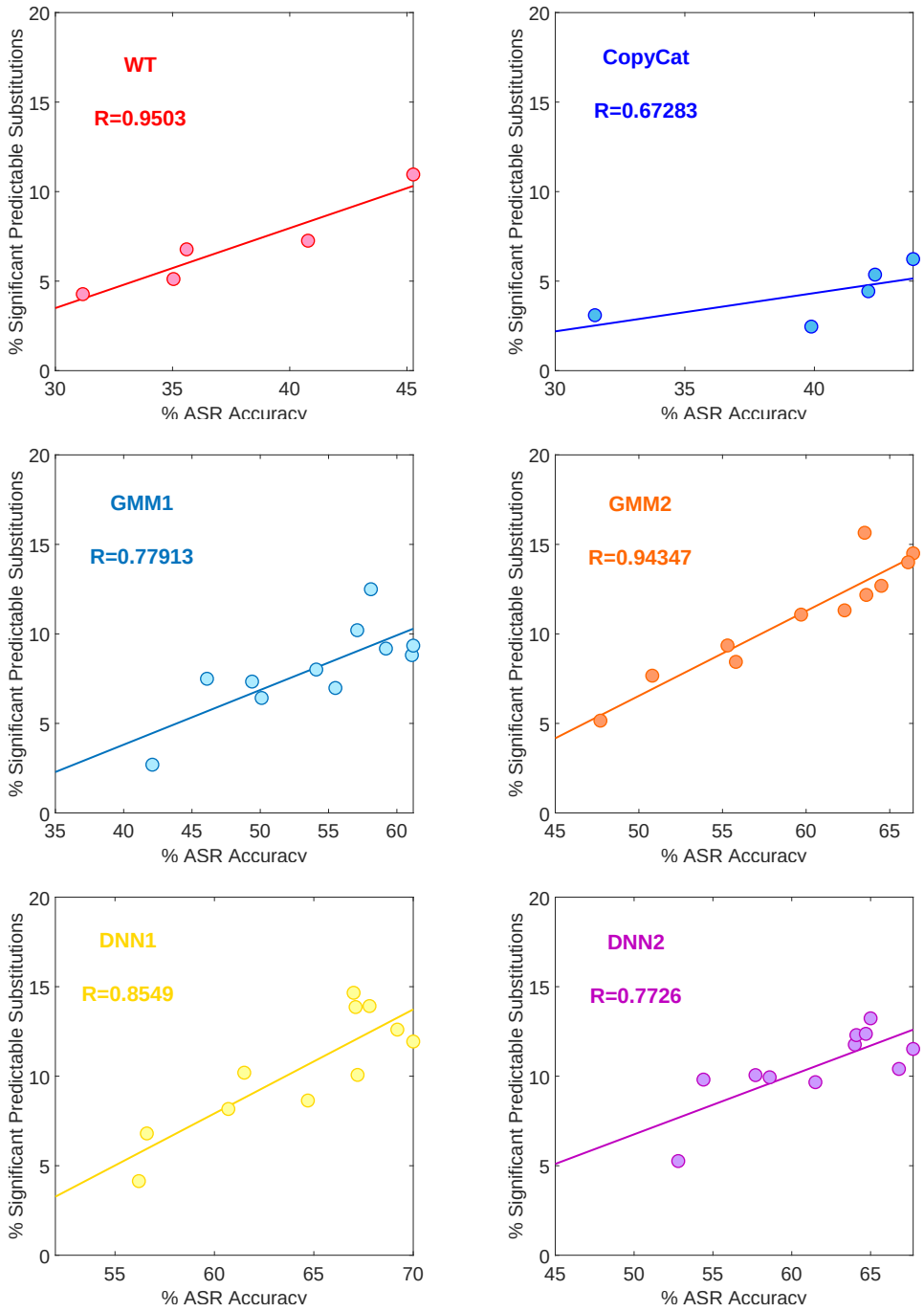


FIGURE 5.4: Scatter plots for WT, CopyCat and all CSLU systems of the percentage of predictable substitutions that occur significantly more frequently than for TIMIT, as a function of phone accuracy. Each of the dots represents an age group. In all plots, the order in which age groups appear follows the progression of the corresponding ASR accuracy.

5.5 A model for the role of PEALA in children's ASR

It has been observed so far, that the output of the ASR systems built on this project does not offer a clear image of systematic phone substitution patterns that could be attributed to PEALA. The different categories of substitution pairs that are of interest, did not reveal any pattern that could have resulted from the development of phonological processes. The two data sets of contrasting speaker competence offered almost identical proportions of predictable substitution errors, failing to exhibit a speech quality effect on the confusions under investigation. Finally, quite unexpectedly, there was a reverse age effect on the PEALA related substitutions across all systems, which appears to be correlated with ASR accuracy. In this section, an attempt will be made to interpret the combination of these findings and produce a model for describing the role of PEALA in the substitution errors of children's ASR.

Let C_0 denote the phone confusion matrix for an ASR system trained and tested on children who are judged not to exhibit PEALA. In other words,

$$C_0(i,j) = P_{ASR}(p_j|p_i) \tag{5.4}$$

For a child ch who does exhibit PEALA, the pattern of phone substitutions can be expressed in a 'pronunciation matrix' P^{ch}, where

$$P^{ch}_{i,j} = P_{ch}(p_j|p_i) \tag{5.5}$$

is the probability that the child produces the phone p_j when standard pronunciation requires p_i. In this case, the element $C^{ch}(i,j)$ of the ASR phone confusion matrix C^{ch} for child ch is given by

$$C^{ch}(i,j) = \sum_{k=1}^{K} P_{ASR}(p_j|p_k)P_{ch}(p_k|p_i) \tag{5.6}$$

where K is the number of phones. In other words,

$$C^{ch} = P^{ch}C_0 \tag{5.7}$$

For illustration purposes, imagine a system with three phones p_1, p_2, p_3. Suppose that a child ch always uses p_1 when the standard pronunciation requires p_2 and that the underlying ASR phone accuracy is 50%, with each phone misrecognised as the other two with equal probability 0.25. Then,

$$P^{ch} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad , C_0 = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix} \tag{5.8}$$

And

$$C^{ch} = P^{ch}C_0 = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix} \tag{5.9}$$

In other words, even though the child always uses p_1 for p_2, according to the phone confusion matrix for that child $P(p_1|p_2)$ is just 0.5. It seems then, in this hypothetical scenario, that the detection of each individual error pattern in a test speaker is as good as the corresponding diagonal value of the system's confusion matrix. In any real case, of course, the pronunciation dictionary would be more complicated and the product of the two matrices would also be more complicated. Moreover, in practice a child is unlikely to make such a substitution error every time. However, it appears that the higher the overall accuracy of a system is, the higher the values in the diagonal elements of the system's confusion matrix are and thus any systematic phone confusion in the child's pronunciation matrix becomes accentuated in the product of the two matrices.

This model is applicable with respect to the results presented in section 6.4.3. Any residual of PEALA in the speech of the older children in CSLU could have been augmented by the high accuracy of their systems, while any PEALA manifestation in the

younger children's speech could have been compressed by the relatively lower ASR accuracy of their recognisers. Since all the systems have been trained on the same data, the variations in their performances must be due to the different test sets. The fact that test sets containing older children's data yield better results is in line with the findings of speech development research, since speech quality is expected to improve with age. In the case of the results presented in section 6.4.2, two groups of speakers within the same age range but with confirmed differences in pronunciation quality, reached similar scores of PEALA related substitutions. According to the proposed model, PSR1 would have higher diagonal values in its C_0 matrix, and lower values in the non diagonal elements of its P^{ch} matrix, compared to PSR2. Thus, the C^{ch} matrices for both systems turned out to be similar as a result of compensation between the high and low values that were involved in the product of P^{ch} and C_0.

The introduction of such a model helps the interpretation of the present results, however it does not confirm whether ASR confusions are indeed systematically influenced by PEALA or not. In the next chapters, it is attempted to look qualitatively into the acoustic features of our data, hoping to detect more distinctive evidence regarding the presence of PEALA into ASR patterns.

Chapter 6

Acoustic Feature Visualisation: Bottleneck Features and i-Vectors

In the absence of evidence for a significant effect of factors associated with language acquisition on phone recognition output, in this chapter the focus is drawn on an investigation of whether these phenomena are evident in the acoustic data. Visualisations through bottleneck features and i-vectors are applied to the CSLU data, aiming to reveal acoustic structure relationships.

6.1 Bottleneck features

An attempt is made to examine the relationships within PEALA related pairs in the acoustic domain and how these evolve with speaker age progression. This is facilitated by the extraction of multidimensional bottleneck features (BNFs) of the training set from the CSLU corpus, which are then plotted in two-dimensional graphs after linear discriminant analysis (LDA) processing. The resulting images reflect the similarities in the acoustic features of several phones through their proximity in the BNF plane.

6.1.1 Bottleneck feature extraction

It was shown in (Weber et al., 2016a) that low dimensional projections of BNFs can represent speech sounds in a topology that broadly reflects their phonetic properties,

and that variations due to different initializations of the DNN may be compensated by suitable linear transformations. In (Bai et al., 2018) it was shown that local structure in low dimensional BNF representations does correspond to phonetic relationships and in fact it depicts phonetic classes organised with respect to phone production mechanisms. Figure 6.1 illustrates how the different phones are distributed in the 2 dimensional BNF space forming classes based on place of articulation.

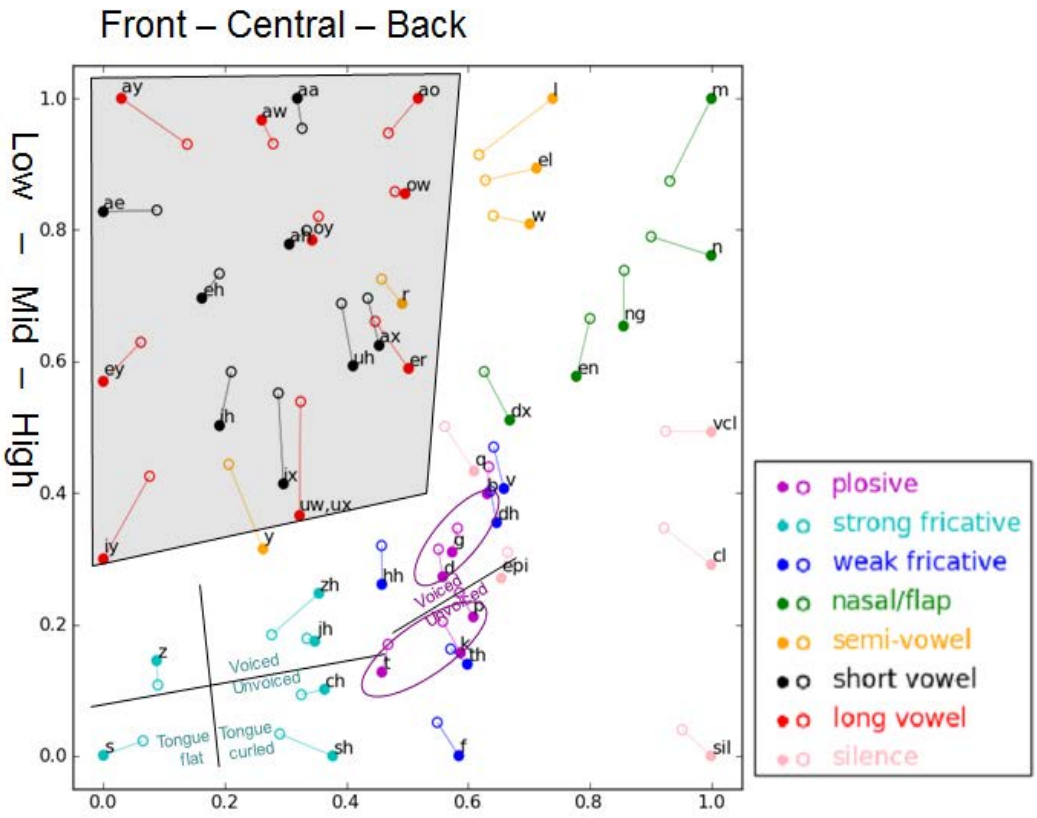


FIGURE 6.1: Taken from Bai et al., 2018.

Motivated by this, it was hypothesized that a BNF induced image might offer an insight on how speech sounds evolve as a function of age. According to this hypothesis, the acoustic realisations of phones corresponding to phonemes involved in predictable substitution pairs will be more similar to each other for younger than for older speakers. The age correlated decrease in similarity could potentially be depicted in 2 dimensional graphs as an increase in the linear distance between the BNFs representing each of the phonemes. The 9 dimensional BNFs extracted during the training of

the DNN2 system (presented in Chapter 4) were used in a visualisation process aiming to investigate this hypothesis.

In order to be able to visualise the BNFs, dimension reduction was attained by LDA aiming to separate the different phone classes and resulting in sets of 2 dimensional vectors. For each phone, an age specific Gaussian ellipse was computed and plotted in different combinations of graphs. Due to the high level of overlap among the data, the phone cluster contours were chosen to correspond to 0.1 standard deviations allowing the phone clusters to be more easily distinguished.

6.1.2 Results

Figures 6.2 and 6.3 depict Gaussian ellipses corresponding to realisations of the complete CMU phone set for speakers aged 5 and 15 years old respectively. They both share considerable similarities with figure 6.1 taken from (Bai et al., 2018). The vowels all concentrated on the top of the image, form a quadrilateral reminiscent of a reversed IPA vowel diagramme. The consonants, are grouped into the same phonetic classes based on place of articulation (nasal, plosive, strong fricative, weak fricative) and are organised into similar local structures with respect to voicing. These similarities are consistent across all age groups of CSLU and are of great importance as they replicate the findings of (Bai et al., 2018) for a completely different corpus (Bai et al. used TIMIT) and with a completely differently initialised DNN. This result confirms the validity of this approach towards illustrating the local relationships between phone representations and renders the analysis of the following results meaningful.

Having established that the BNF induced phone representations attained from CSLU do reflect phonetic relationships, it is reasonable to draw a comparison between different age groups' data, in order to examine how these relationships evolve from childhood to adolescence. Figures 6.2 and 6.3 represent the youngest and oldest age groups of the CSLU corpus. The overall progression of the phones' spatial distribution is according to expectation, as all Gaussians from the 5-year-olds' data appear clustered closer together than the corresponding ones from the 15-year-olds' data, which are more widely spaced. This trend is consistent within each of the phonetic classes,

where distances between different phones of the same class are broadened in the older speakers' data. This observation suggests that phones which are easily confused by young children become more distinguishable as they get older.

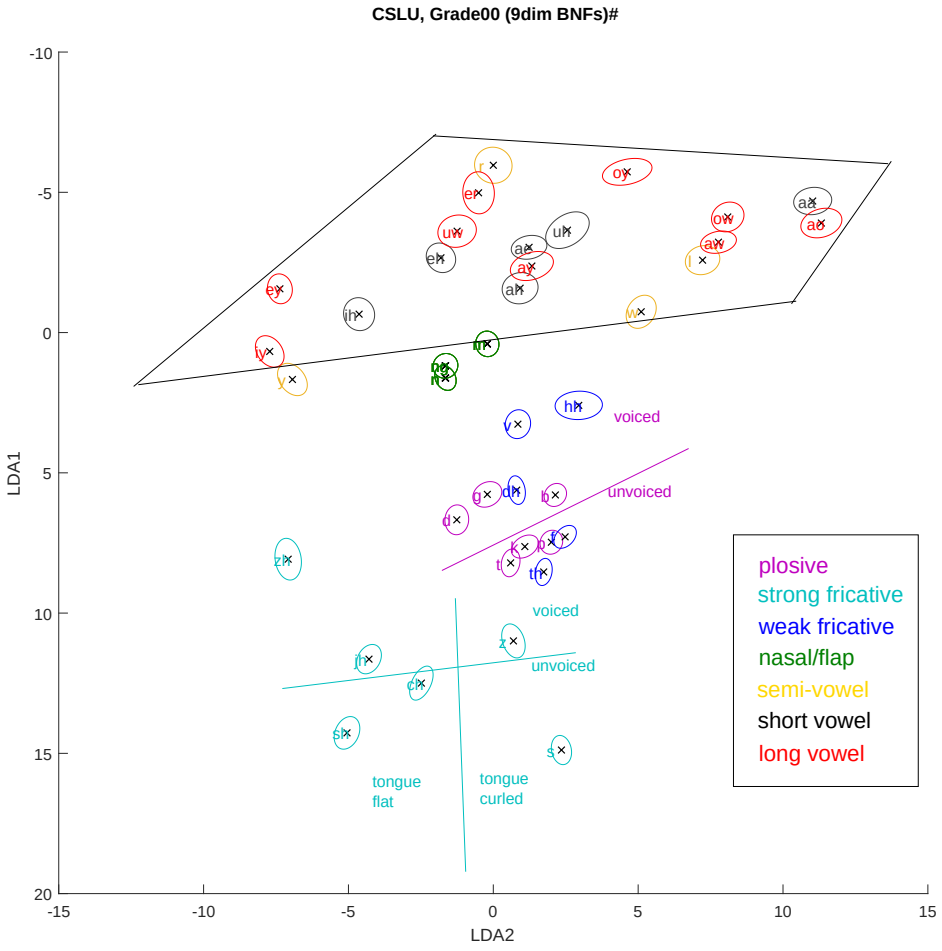


FIGURE 6.2: Plot of 2-dimensional projection of 9-dimensional BNFs. Individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours for each of the phones in the CMU set as realised by 5-year-old speakers.

A closer look at the PEALA related pairs is achieved by creating figures which isolate the evolution of each pair in question across the different age groups. The Gaussian ellipses representing the 27 predictable phoneme substitution pairs related to PEALA (as described in Table 2.3) were colour-coded to represent the different age groups, with shades of green and blue for the younger speakers and red and yellow for

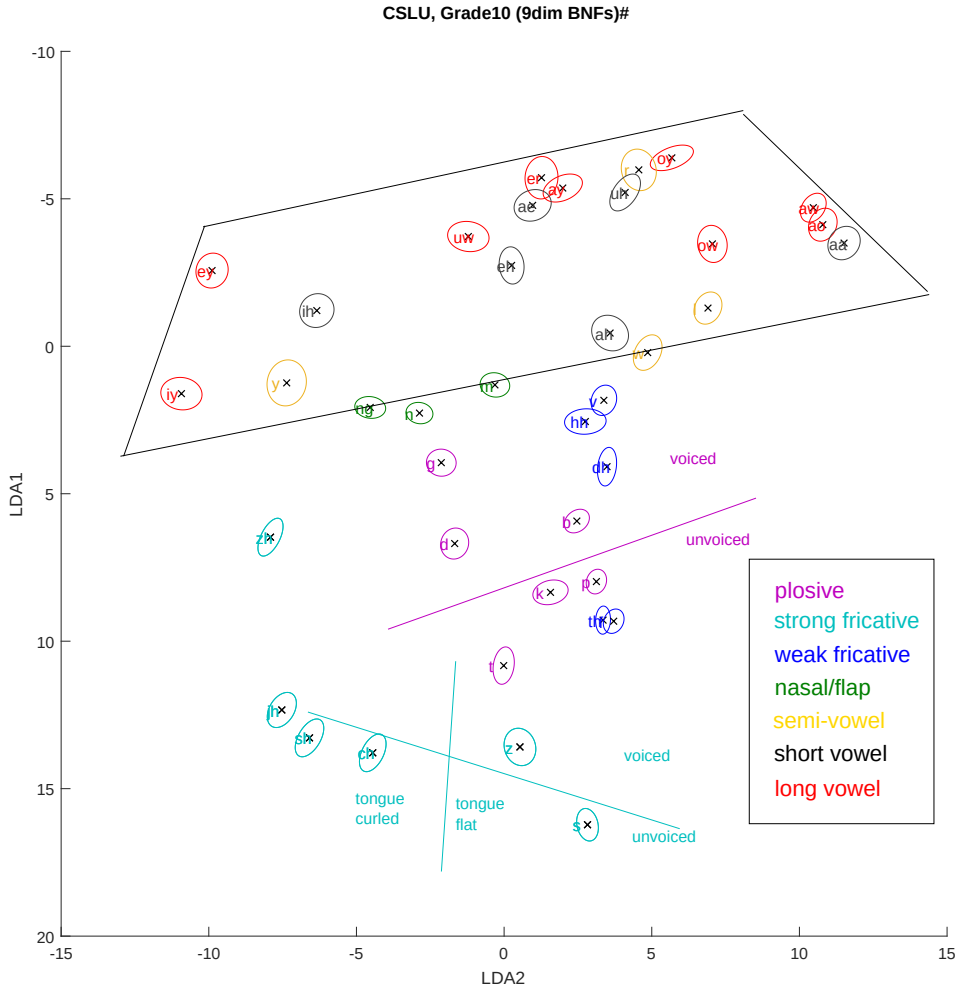


FIGURE 6.3: Plot of 2-dimensional projection of 9-dimensional BNFs. Individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours for each of the phones in the CMU set as realised by 15-year-old speakers.

the older ones, as illustrated in figure 6.4. In addition, the Euclidean distances between the mean values of the 2 dimensional data corresponding to substitution pairs were calculated and plotted into bar charts as a function of age. Each age group in the bar charts was again colour coded to match the correspondences set in figure 6.4.

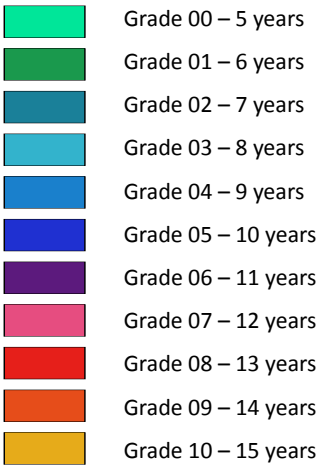


FIGURE 6.4: Colour coding of the different age groups of the CSLU corpus used in figures .

Figures in Appendix B show all the Gaussian ellipse plots for the 27 pairs of confusable phones related to PEALA, as well as their corresponding bar chart graphs of BNF distance development as a function of age. For each age group, a line connects the mean values of BNFs for each of the two phonemes. The bar charts illustrate whether the phonemes in question come closer together or move away from each other as the speaker age increases. Following visual inspection, these charts can be divided into four categories: (a) 'increasing distance as a function of age', (b) 'constant distance', (c) 'decreasing distance as a function of age', and (d) 'randomly alternating distance'. The graphs featured in figure B.1 belong in category (a), those featured in figures B.2 roughly represent category (b), those in figures B.3 roughly represent category (c) and finally figure B.4 features graphs from category (d).

In general, plosives are featured in category (a) for the most part, with a few appearances in category (b) and (d), while liquids and glides are mostly featured in category (c). Fricatives are spread throughout the four categories. This partition could potentially reveal some motif of linguistic importance, it is unclear though, whether it is linked to speech development. If phone separation increases with age, one would expect the vast majority of pairs to be in category (a), which is the case. However, the surprising outcome in pairs such as $/p/ - /th/$ or $/l/ - /r/$ complicates the analysis.

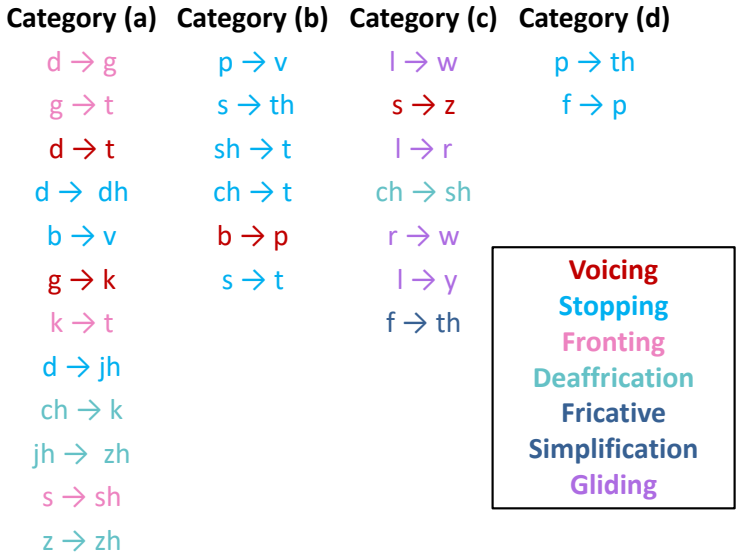


FIGURE 6.5: The distribution of different PEALA related pairs into four categories of BNF distance development as a function of age. Each pair is colour coded based on the phonological process it is linked with, according to the legend.

The distribution of the different PEALA related pairs from Table 2.3 is presented in Figure 6.5. After close examination, several patterns emerge. All fronting errors belong to category (a) and all gliding errors belong to category (c), along with fricative simplification. Deaffrication appears mostly in category (a) with one exception in category (c). Stopping is spread in categories (a), (b) and (d), while voicing is spread in categories (a), (b) and (c). This points out that category (c) is comprised of gliding and fricative simplification, with two exceptions from different processes. These two exceptions happen to involve acoustically confusable pairs of phones ($/s/ \rightarrow /z/$, $/ch/ \rightarrow /sh/$), therefore their appearance in this category where the BNF distance decreases with age, could be explained as a consequence of the high confusability between the two sets of sounds. Gliding and fricative simplification though, appear to be exhibiting a phonological phenomenon, which is not predictable by PEALA. Apart from two stopping occasions ($/p/ \rightarrow /th/$, $/f/ \rightarrow /p/$), which as explained below, are considered to be outliers, the rest of the PEALA related pairs are occupying categories (a) and (b), with the majority of them being in category (a). This confirms the hypothesis that BNF distance increases as a function of age, and PEALA related processes

such as voicing, stopping, fronting and deaffrication are all asserting this. The arising question at this point is what is it that separates gliding and fricative simplification from this trend. One possible explanation is that this particular projection of the 9 dimensional BNFs completely ignores the dimension in which the expected divergence is expressed for these processes. In spite of this possibility though, the current projection has pointed out enough predictable patterns to make a case that they do not occur randomly. A closer look at a few characteristic examples might be helpful.

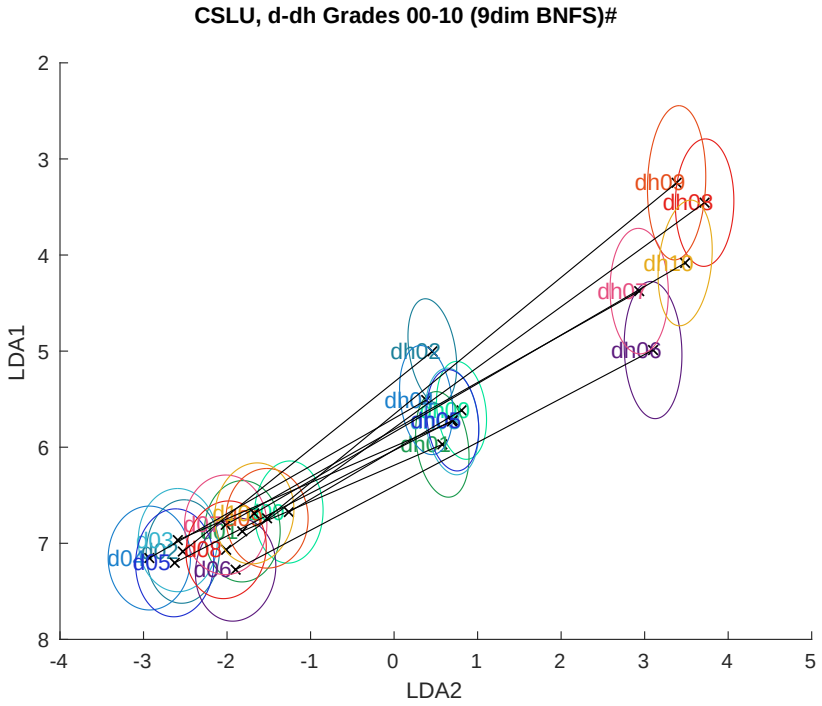


FIGURE 6.6: 2-dimensional LDA projections of 9-dimensional BNFs. Individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours for the PEALA related pair of phones: **d-dh**. Each ellipse is colour coded to reflect the different age groups.

Figure 6.6 (B.1f in Appendix B) shows the plot for the pair $/d/-/dh/$ from category (a). It is apparent that there is a trend for the distance between the mean values of the two phonemes to increase with age. The observation that the $/dh/$ ellipses are closer to the $/d/$ ellipses for the youngest children is consistent with the hypothesis that they move progressively away from $/d/$ with increasing age. The figure shows a distinct cluster for $/d/$ in the bottom left-hand corner, while the values for $/dh/$ form two age-dependent clusters, which separate the data in a non-gradual manner

in two groups; one below and one above age twelve.This distinct progression of age-dependent clusters could be a consequence of boys' voices breaking around that age, combined with the fact that $/dh/$ has been reported to be one of the last consonants to be acquired (Dodd et al., 2003) and thus is more linguistically challenging. Indeed, this pattern of distinct age clusters was again observed in the cases of $/v/$, $/r/$ and $/l/$ which are also among the lastly acquired phonemes. On the contrary, phonemes that appear first in the phonemic acquisition repertoire such as $/p/$, $/b/$ and $/d/$ (Dodd et al., 2003), progress more gradually and in more compact clusters.

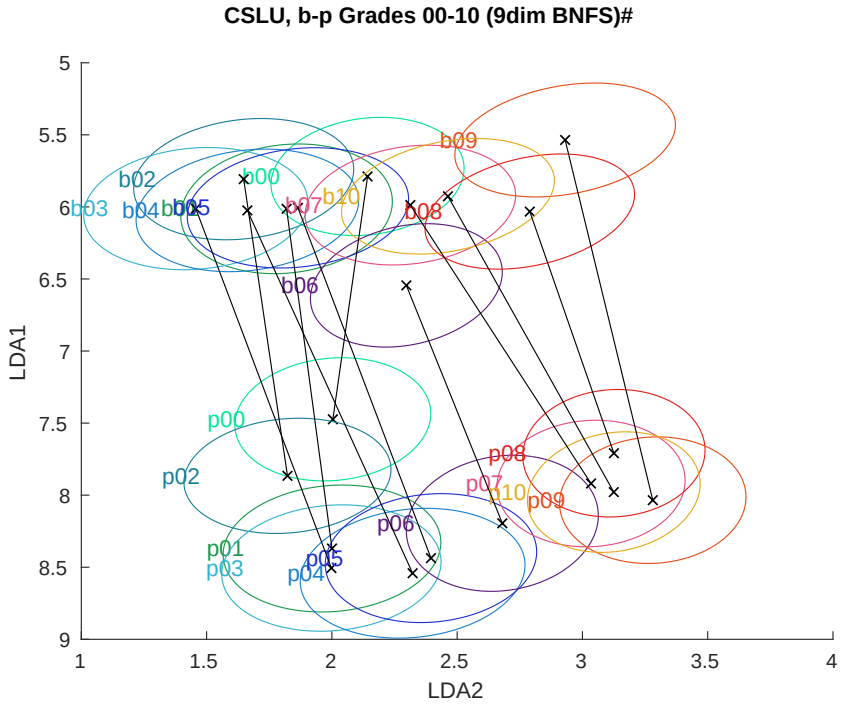


FIGURE 6.7: 2-dimensional LDA projections of 9-dimensional BNFs. Individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours for the PEALA related pair of phones: **b-p**. Each ellipse is colour coded to reflect the different age groups.

Figure 6.7 (B.2i in Appendix B), corresponding to the pair $/b/ - /p/$, is an example from category (b), where the separation of phones is considered approximately constant since the youngest age group does not differ much from the older groups, there are however a few irregular increases and drops of distances. It suggests that there is considerable variability between the realisations of $/b/$ and $/p/$. The evolution of these realisations as a function of age is quite smooth for both phonemes, with $/p/$

describing a semi-circle converging for age 12 and above, offering another indication of male voice breaking.

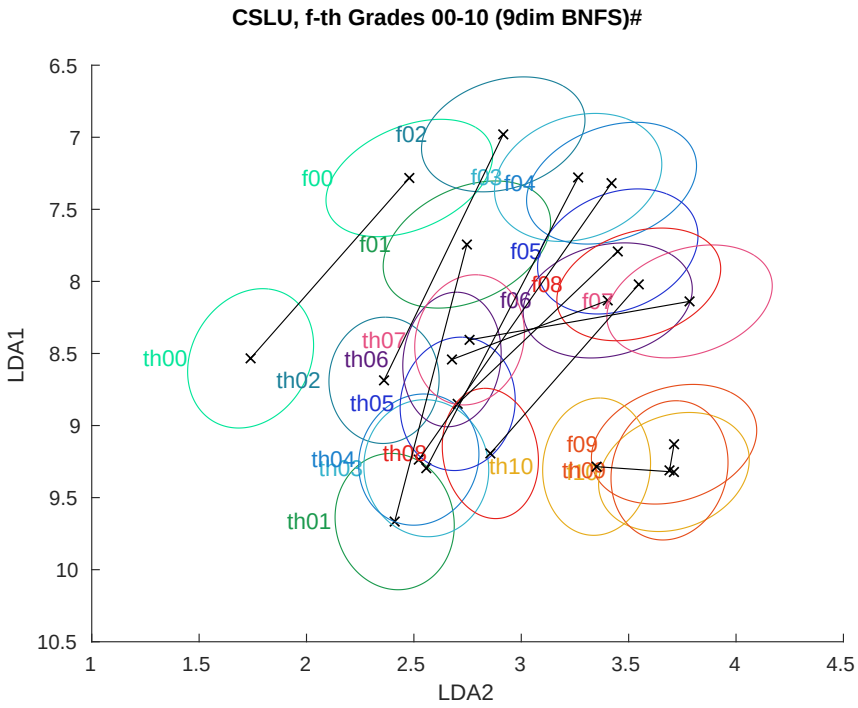


FIGURE 6.8: 2-dimensional LDA projections of 9-dimensional BNFs. Individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours for the PEALA related pair of phones: **f-th**. Each ellipse is colour coded to reflect the different age groups.

Figure 6.8 (B.3m in Appendix B), corresponding to the pair $/f/ - /th/$, is an example from category (c), where contrary to expectation, separation decreases with increasing age. The clusters for $/f/$ and $/th/$ are diverse and overlapping compared with the previous two figures. In fact the trend for $/f/$ and $/th/$ to become more confusable with increasing age is supported by the results of Chapter 5. If this coincidence indicates a true phenomenon, that young teenagers' production of $/f/$ and $/th/$ becomes more alike with age, then it seems very unlikely that this behaviour can be attributed to language acquisition, and a different explanation needs to be found, possibly of sociolinguistic nature.

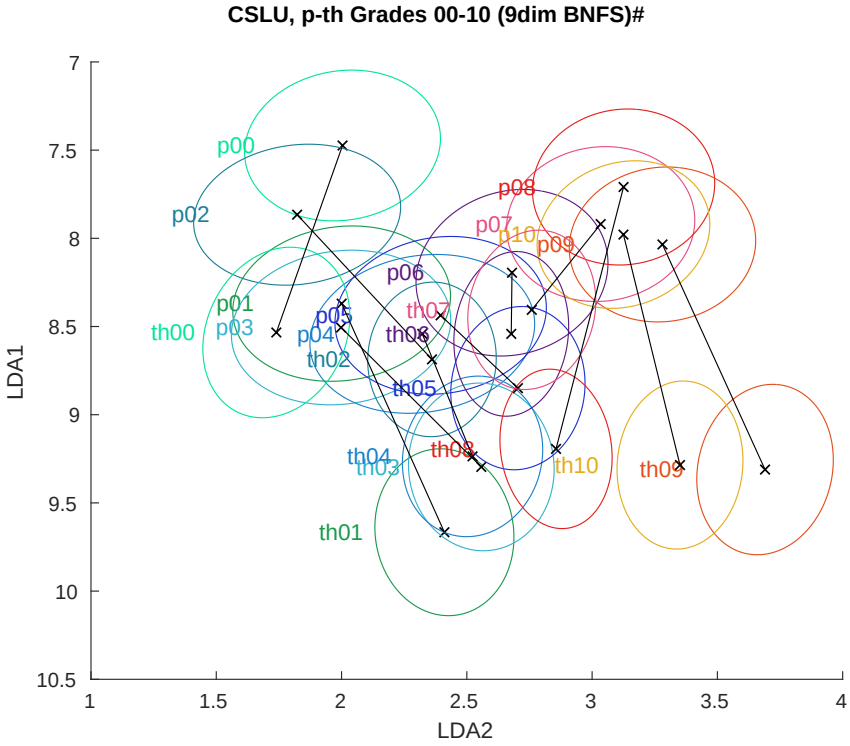


FIGURE 6.9: 2-dimensional LDA projections of 9-dimensional BNFs. Individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours for the PEALA related pair of phones: **p-th**. Each ellipse is colour coded to reflect the different age groups.

Figure 6.9 (B.4b in Appendix B), corresponding to the pair $/p/ - /th/$, is one of the two occasions of category (d) where distances for the different age groups alternate in random manner, in this case resembling a co-sinusoidal curve. The other such occasion (B.4a in Appendix B) is the pair $/f/ - /p/$ whose distances for the different age groups progress in a sinusoidal manner. These two pairs comprising category (d) could be considered as some sort of outliers, as they do not match any other category and are difficult to interpret. A possible explanation could be that differentiating between one of these sounds and a third sound takes priority at a particular stage of development, causing these two sounds to become temporarily more confusable. However, this is just speculation.

6.2 Speaker Specific Feature Visualisation: i-Vectors

The focus of the previous sections has been on systematic changes into particular speech sounds as a result of language development. In this section, the focus is shifted towards systematic changes affecting the entire speech system of developing children. I-vectors are known to yield state of the art performance for speaker identification and have been used successfully in children's speech (Safavi, Russell, and Jancovic, 2018). The aim of the current section is to employ i-vectors in order to visualise and interpret the way individual speakers of the same age group relate to each other and how these relationships differ among different age groups.

6.2.1 I-vector extraction

I-vectors have been previously described in Chapter 1. They are relatively low dimensional vector representations of speakers which emerge from a linear mapping T of a 'total variability space' V whose dimension is lower than the supervector space, to the supervector space S which is created by the concatenation of the mean vectors of a speaker dependent GMM ($T : V \rightarrow S$) (Dehak et al., 2011). This way a supervector s is expressed as $s = Tw + \epsilon$, where w is an i-vector and ϵ stands for error. Training of i-vectors is an iterative process which uses posterior probabilities like those used in the E-M algorithm, referred to as 'sufficient statistics', in order to find T and w such that will maximise the probability of the training data.

The extraction of 200 dimensional ivectors was implemented with the use of the training set of the CSLU speech corpus and the Kaldi Speech Recognition Toolkit (Povey et al., 2011). Dimension reduction was achieved with the application of LDA, which performed class separation with respect to the different speakers in the data set. The resulting 2 dimensional ivectors, each corresponding to a speaker from the CSLU training set, were subsequently plotted into Gaussian ellipse graphs, with each ellipse representing a single speaker. Likewise the BNF graphs in the previous section, there was a high level of overlap among the data, so in order to be able to distinguish the ivector cluster contours, they were chosen to correspond to 0.1 standard deviations.

Each cluster was colour coded based on its corresponding speaker age following the same patterns as described in figure 6.4.

6.2.2 Results

Figure 6.10 contains Gaussian ellipse plots representing each speaker from the whole CSLU data set, colour coded according to the different age groups. As stated in Chapter 2, younger speakers are known to present higher levels of variability which decreases as a function of age. Therefore, the expected outcome in the i-vector plot would include larger Gaussian ellipse contours for the younger speakers, expressing within-speaker variability, and in general broadly distributed clusters for the younger age groups, which would gradually become compact clusters for the oldest ones, expressing between-speaker variability. On the contrary, the depiction of the i-vector clusters in figure 6.10 indicates no substantial difference in the size of the ellipses between the different age groups and presents the younger speakers' plots in more compact neighbourhoods of closer proximity than the older speakers' plots, which are more spread out in sparse distribution. Moreover, instead of progressing in a straight line, the placement of the ellipses forms a semicircle, arching near the 11-year-old speakers' data.

The interpretation of the unanticipated patterns in these results is not clear-cut, but rather speculative. The fact that the clusters are more compact for the youngest children and do not show any signs of within subject variability, can be supported by evidence found in (Safavi, Russell, and Jancovic, 2018) where the same speech corpus is used in speaker identification. In this study, age groups were further arranged into three categories of increasing age. According to their findings, the youngest of these categories exhibited by far the most successful results in speaker identification and in fact, in the occasional confusions, children from the youngest age groups were only likely to be confused for children within their own category, i.e. children one or two years apart from them, while children belonging in the oldest age groups were found likely to a small extent, to be mistaken for children from the other two categories as well as their own. This is in line with the closeness in the younger speakers' plots

and the high separation in those of older speakers, however it does not provide an explanation for them.

The semicircular shape of the graph might be explained by the fact that the ellipse plots which begin to deviate from a straight line alignment and take this semicircular form, are representing 11-year-old speakers, could be signifying the point where boys' voices start breaking and thus are separated from the rest of the speakers (girls and boys whose voices haven't broken yet). Figures 6.11a and 6.11b were created in order to investigate this hypothesis, as they present the same data separately for male and female speakers. The progression is clear, the data start in compact distributions, of overlapping clusters for male and female i-vector plots, and as age increases, the male data diverge towards the same direction as the semicircle observed in figure 6.10. However the separation is not perfect, and ellipses corresponding to female speakers are also slightly shifted for the older age groups, therefore gender alone cannot justify the unexpected shape of the data in figure 6.10. Alternatively, it might be an indication that the particular LDA dimensions chosen for this projection are not the most appropriate to describe age differences, and potentially had LDA been conducted with respect to age group separation instead of speaker separation, the image would be closer to expectation.

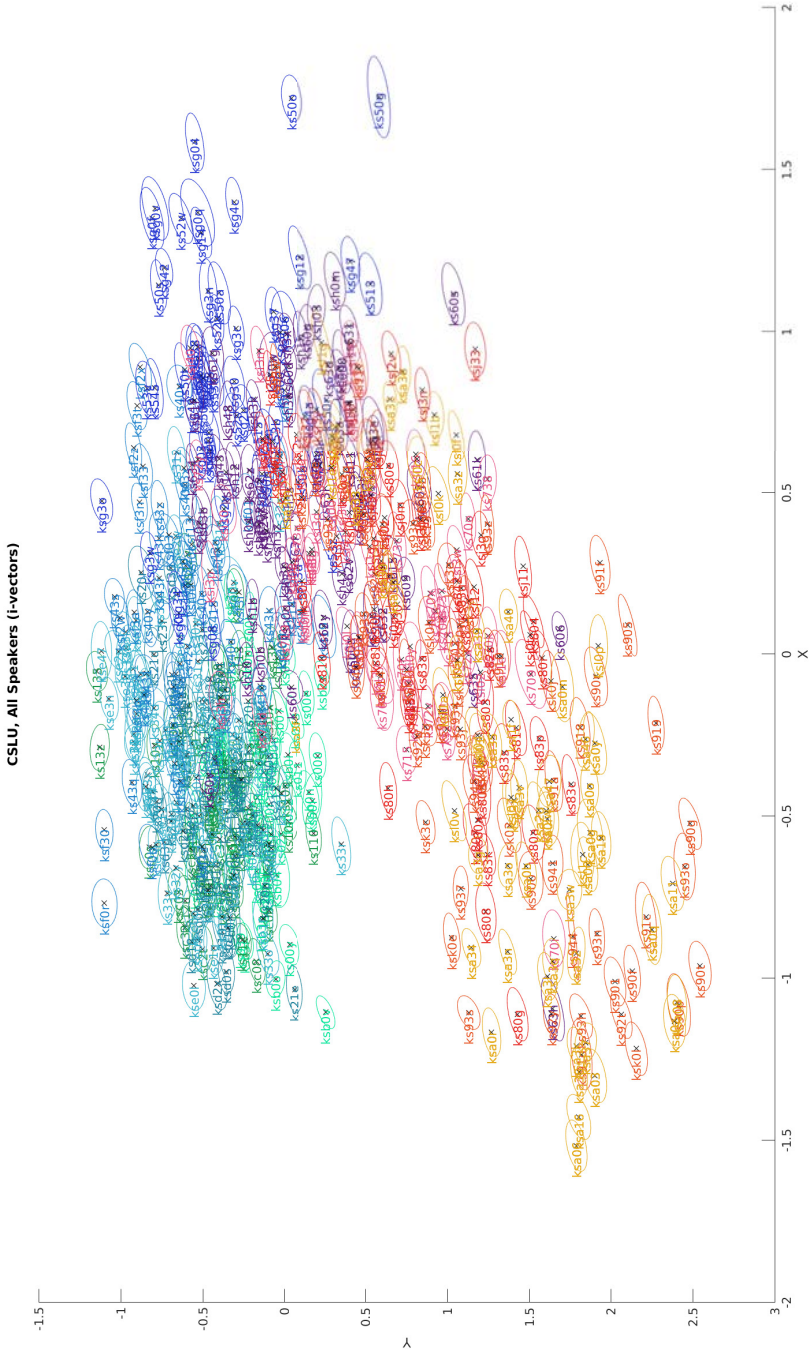
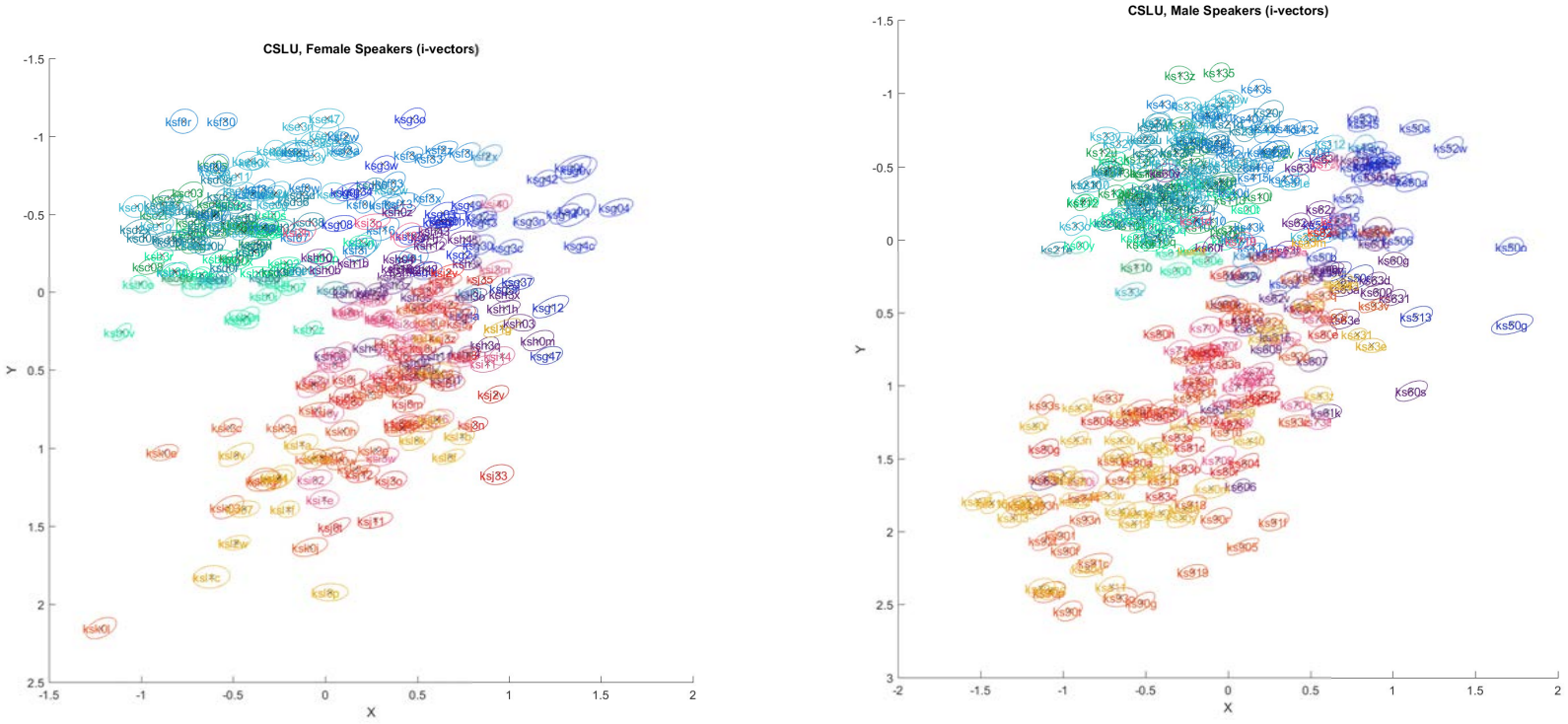


FIGURE 6.10: *LDA projections of 200-dimensional i-vectors in the 2-dimensional space. Each point corresponds to a speaker. The data have been colour coded, with blue and green shades for the younger speakers, and red and yellow shades for the older speakers.*



(A) Female Speakers

(B) Male Speakers

FIGURE 6.11: LDA projections of 200-dimensional i-vectors in the 2-dimensional space for female (left) and male (right) speakers. Each point corresponds to a speaker. The data have been colour coded, with blue and green shades for the younger speakers, and red and yellow shades for the older speakers.

Chapter 7

Conclusions

The present thesis aimed to investigate the extent to which ASR errors on children's speech can be attributed to common phonological effects associated with language acquisition (PEALA). Findings from studies on speech development were used as the framework upon which the two research questions were formulated: Can PEALA be detected in systematic patterns of ASR phone confusion errors? and, Can PEALA be evidenced in systematic patterns of acoustic feature structure? A set of predictable patterns of error were defined and guided the analysis of the experimental results reported. ASR experiments were conducted with the use of four children's and two adults' speech corpora comprising both American and British English.

To address the first question, ASR phone confusion matrices were extracted and analysed with the use of a statistical significance test, which was proposed for the purposes of the current work. The emerging results suggested a strong correlation between PEALA-related predictable errors and ASR accuracy which produced a reverse age effect, against what was expected. A simple mathematical model was introduced in order to interpret this finding within the developmental framework set previously. This way the research question was answered, as PEALA were detected in ASR phone confusions at a steady but low rate across speech corpora, however, the accuracy correlation effect (the fact that these errors become more difficult to detect from poor performing ASR) impeded the establishment of a developmentally meaningful systematicity across ages. The direct implication for children's ASR is that addressing

PEALA related phone substitutions will not lead to significant performance improvement, especially for the youngest speakers.

In addition, to answer the second question, bottleneck features and i-vectors representing the acoustic features in one of the systems developed, were extracted and visualised in 2 dimensional plots with the help of linear discriminant analysis. A qualitative analysis was conducted with reference to patterns that can be predicted through PEALA. The related findings confirmed the hypothesis that visualisation of acoustic structures through the use of BNFs can reflect phonetic properties, by replicating the findings of (Bai et al., 2018), and affirmed the research question. The fact that these differences are visible in the acoustic BNF space and the acoustic BNF space is locally interpretable in terms of phonetic features indicate that there is a potential application in areas such as speech therapy, where the phonetic interpretation of acoustic differences could be used to provide feedback to the child and understanding for the clinician.

The first important contribution of the thesis is the compilation of findings from speech therapy and language acquisition, introducing phonological effects associated with language acquisition (PEALA) as a predictor for ASR errors on children's speech, into a single model appropriate for the analysis of ASR performance. This is not a straightforward procedure, as different papers define different criteria to decide whether of not a particular sound is acquired and therefore come up with different conclusions about the age of acquisition. Providing a detailed inventory of developmental phenomena of speech, and relating them directly to the output of ASR is an original approach to children's ASR which has not been previously explored.

Furthermore, the introduction of a methodology, based on the statistical analysis of confusion matrices, for identifying ASR errors in children's speech that may be due to PEALA, and proposing a statistical significance test that aims to isolate the error patterns that occur significantly more frequently in children's than in adults' speech, offers a structured and reliable process to run the analysis of ASR results.

Next, the application of this methodology through ASR experiments on different children's speech corpora, including corpora comprising spontaneous and read

speech as well as speech of documented varying competence, offers a broad selection of results which cover many different examples of speech recordings. This adds validity to the ASR results which follow the same trends in spite of the differences in the speech corpora. These experiments suggest that PEALA do not have a significant effect on ASR accuracy and, contrary to expectation, the proportion of errors that are associated with PEALA increases with age. This way the initial hypothesis of a PEALA related effect on the output of ASR is refuted and there is evidence to support it.

Additionally, the introduction of a simple model of the relationship between ASR phone error rate (and hence age) and the proportion of ASR errors attributable to PEALAs that are detectable in the phone confusion matrices offers an explanation of the potential underlying reason for the refutation of the initial hypothesis.

Graphic visualisation of children's speech features, using bottleneck features (BNFs) towards the analysis of the development of the realization of different speech sounds as a function of age is a novel method that incorporates state of the art techniques, in the context of PEALA. Similarly, the use of i-vectors to visualise the progression of individual children's entire speech systems as a function of age, offers valuable insights in the context of speech development.

One of the assertions of the thesis is that the proportion of ASR substitution errors attributable to PEALA will increase as a consequence of improved ASR performance. As ASR technology evolves it will be possible to confirm this hypothesis. However, no matter how good ASR performance is, the observed pattern of errors will be due to a combination of child pronunciation errors and ASR errors. To understand the true extent of actual mispronunciations, a large quantity of data hand transcribed at the phonetic level would be required, preferably by trained professional phoneticians. This would provide a gold standard, enabling the true and expected transcriptions to be compared. However, this is unlikely to happen because of the costs involved in the attainment of the expertise and resources needed for this level of annotation.

This thesis has focused on substitution errors that are likely to be caused by phenomena associated with language acquisition but has ignored deletion and insertion

errors. Since phone deletions and insertions are also factors in language acquisition (with phonological processes such as cluster reduction, deletion of final consonants, deletion of unstressed syllables and reduplication) it is important to extend the current models to explain these types of errors. Therefore, another line of future research would be to investigate the extent to which ASR deletion and insertion errors are attributable to PEALA.

Over the past thirty years the trend has been to move away from the use of human knowledge towards machine learning in speech technology, for example to achieve higher accuracy in ASR. This is epitomised in the current trend towards end to end ASR. However, applications such as interactive language learning and speech therapy rely on the presence of representations of human knowledge as a means to provide feedback. Therefore, there is a tension between the need to move towards data driven approaches to achieve good performance and the need for the systems to be interpretable in order to provide feedback. The findings from the present research provide a solution to this problem by showing that the structures that emerge from machine learning can be interpreted in terms of phonetic features. This area warrants further research.

The BNF-based visualisations of the acoustic realisations of phones in children's speech are dependent on the particular initialisation of the DNN parameters prior to training. Previous work has suggested that differences in the realization of vowels in BNF space can be accommodated using linear transformations (Weber et al., 2016b). More recently it has been suggested that the differences between entire BNF spaces that arise due to different initialisations can be removed by a piece-wise linear (Bai, 2017) or a one-to-one non-linear continuous and differentiable mapping (a diffeomorphism) implemented with a simple neural network (Bosch and Boves, 2018). It would be useful to explore this further and examine BNFs corresponding to different DNN initialisations for children's speech.

The BNF-based visualisations of the realisations of phones that are difficult to acquire, such as /dh/, suggest that young children adopt an intermediate realisation of /dh/ that is located between the final realisations of /d/ and /dh/ in BNF space. It

would be interesting to pursue this further to see if there is any evidence for this type of phenomenon in the literature on speech development in children. There has been very recent work on human perceptual learning and computer recognition of 'confusable' phones which lie on the continuum between /l/ and /r/ (Scharenborg et al., 2018) and this may also prove a useful route to explore.

In (Safavi, Russell, and Jancovic, 2018) it is shown that i-vectors can support high-accuracy speaker, gender and age recognition from children's speech. An i-vector characterises the complete inventory of sounds that an individual makes (at least those that occur in the training data). Therefore it might be expected that a visualization of i-vectors for children of various ages might throw light on how children's overall speech production skills evolve with age. However, the results of the analysis of children's speech systems using i-vectors are not as expected. This may be because the criterion used in the derivation of the LDA, namely speaker classification, is not the best visualising the effects of age. Projections that use LDA transforms trained on different class structures, for example, age-groups, might lead to more intuitive results.

Appendix A

Phone Set Mappings

TABLE A.1: *Phone set mappings.*

TIMIT 61 symbol set	BEEP 45 symbol set	CMU 40 symbol set
	ia	ih
ih	ih	
ix		
iy	iy	iy
ae	ae	ae
ah	ah	ah
ax-h		
ax	ax	
eh	eh	eh
	ea	
uh	uh	uh
	ua	uw
uw	uw	
ux		
aa	aa	aa
	oh	
ao	ao	ao

er	er	er
axr		
ey	ey	ey
ay	ay	ay
oy	oy	oy
aw	aw	aw
ow	ow	ow
p	p	p
t	t	t
k	k	k
b	b	b
d	d	d
dx		
g	g	g
pcl	sil	sil
tcl		
kcl		
bcl		
dcl		
gcl		
q		
pau		
h#		
epi		
f	f	f
v	v	v

s	s	s
z	z	z
sh	sh	sh
zh	zh	zh
th	th	th
dh	dh	dh
ch	ch	ch
jh	jh	jh
m	m	m
em		
n	n	n
nx		
en		
ng	ng	ng
eng		
l	l	l
el		
r	r	r
w	w	w
y	y	y
hh	hh	hh
hv		

Appendix B

BNF plots

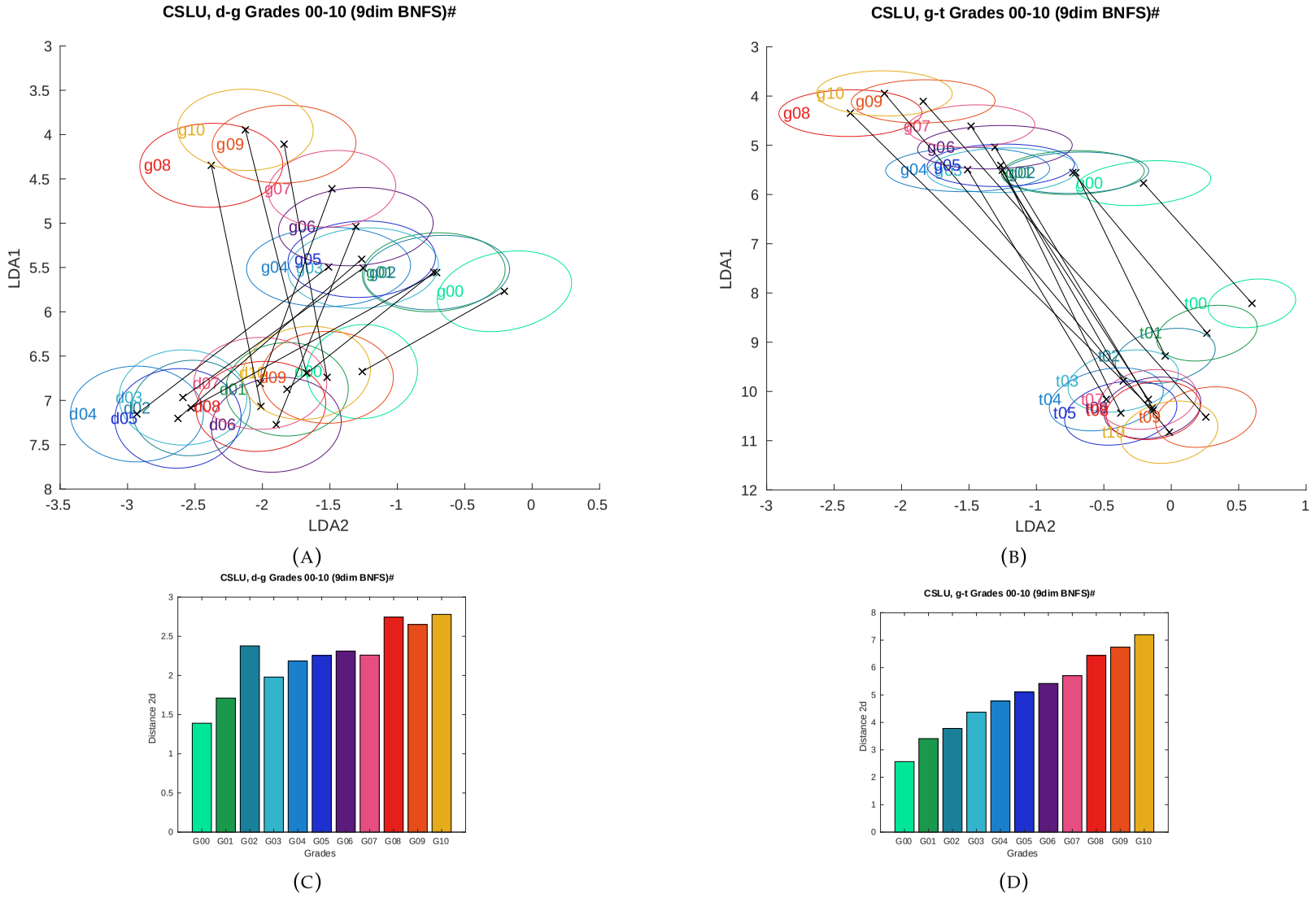


FIGURE B.1: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(a) 'increasing distance as a function of age'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.

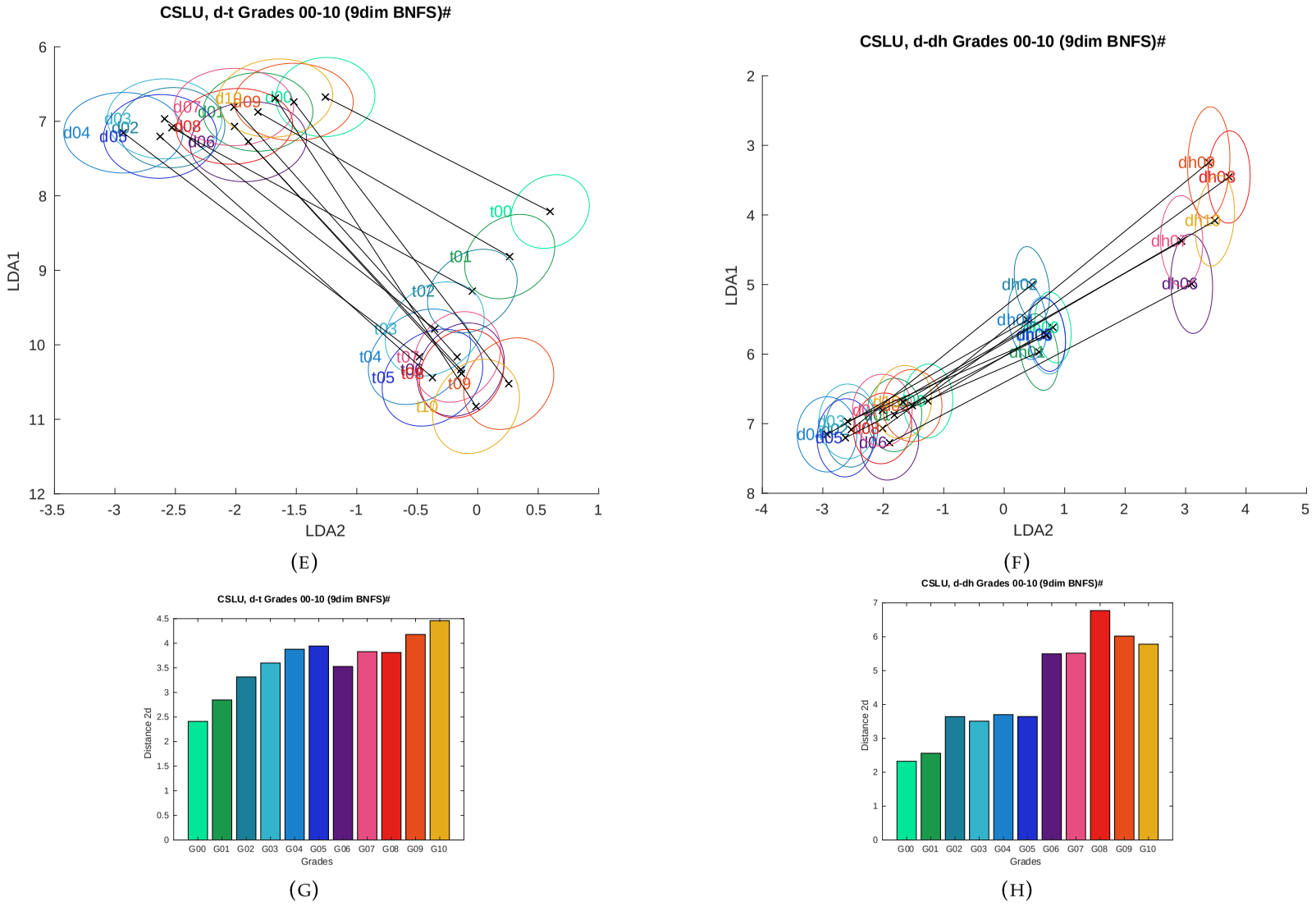


FIGURE B.1: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(a) 'increasing distance as a function of age'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.

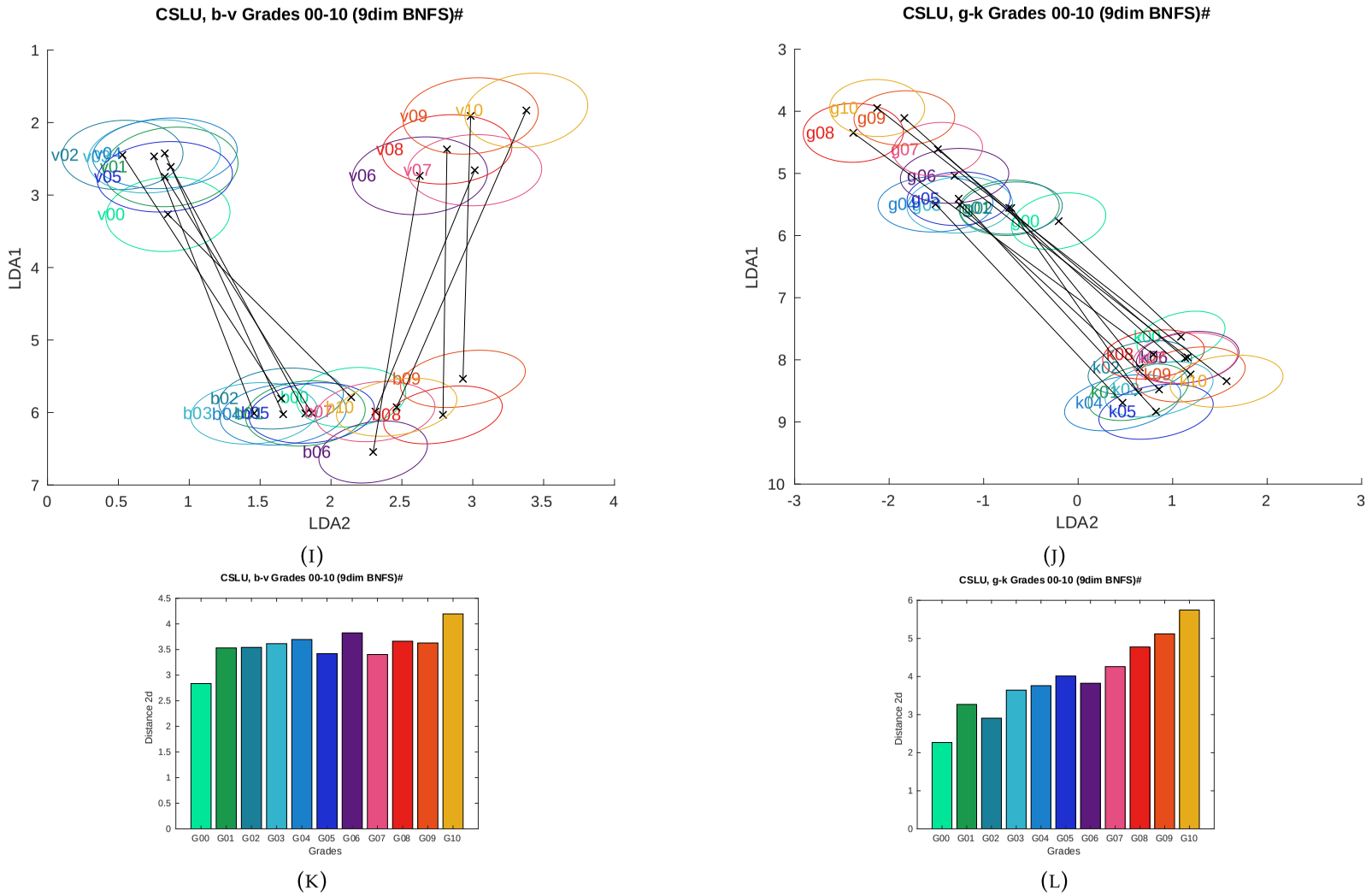
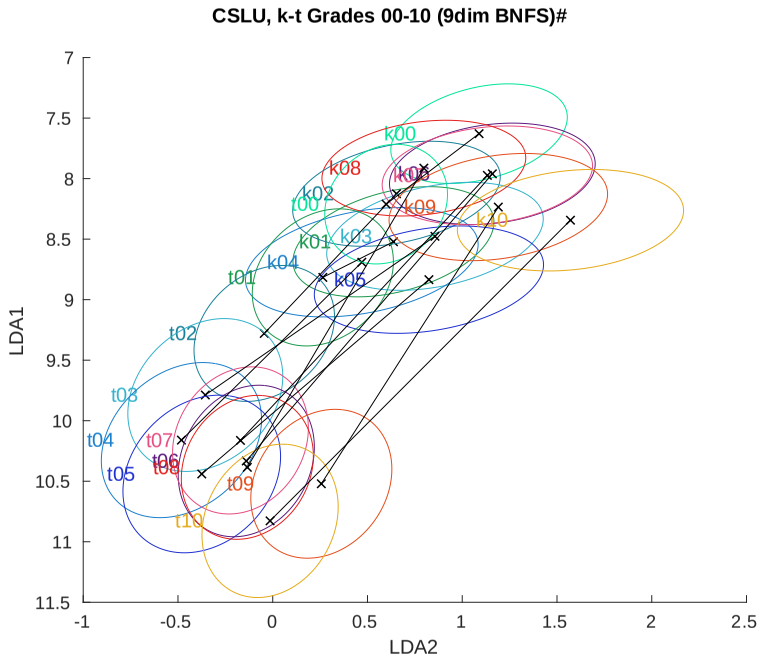
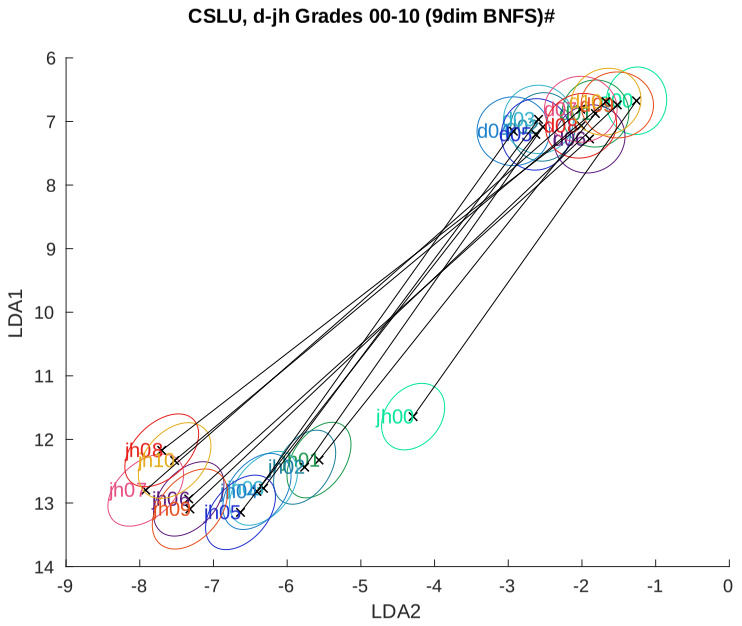


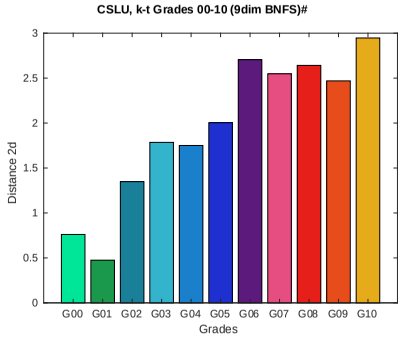
FIGURE B.1: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(a) 'increasing distance as a function of age'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.



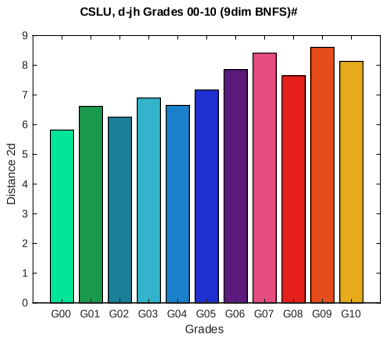
(M)



(N)



(O)



(P)

FIGURE B.1: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(a) 'increasing distance as a function of age'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.

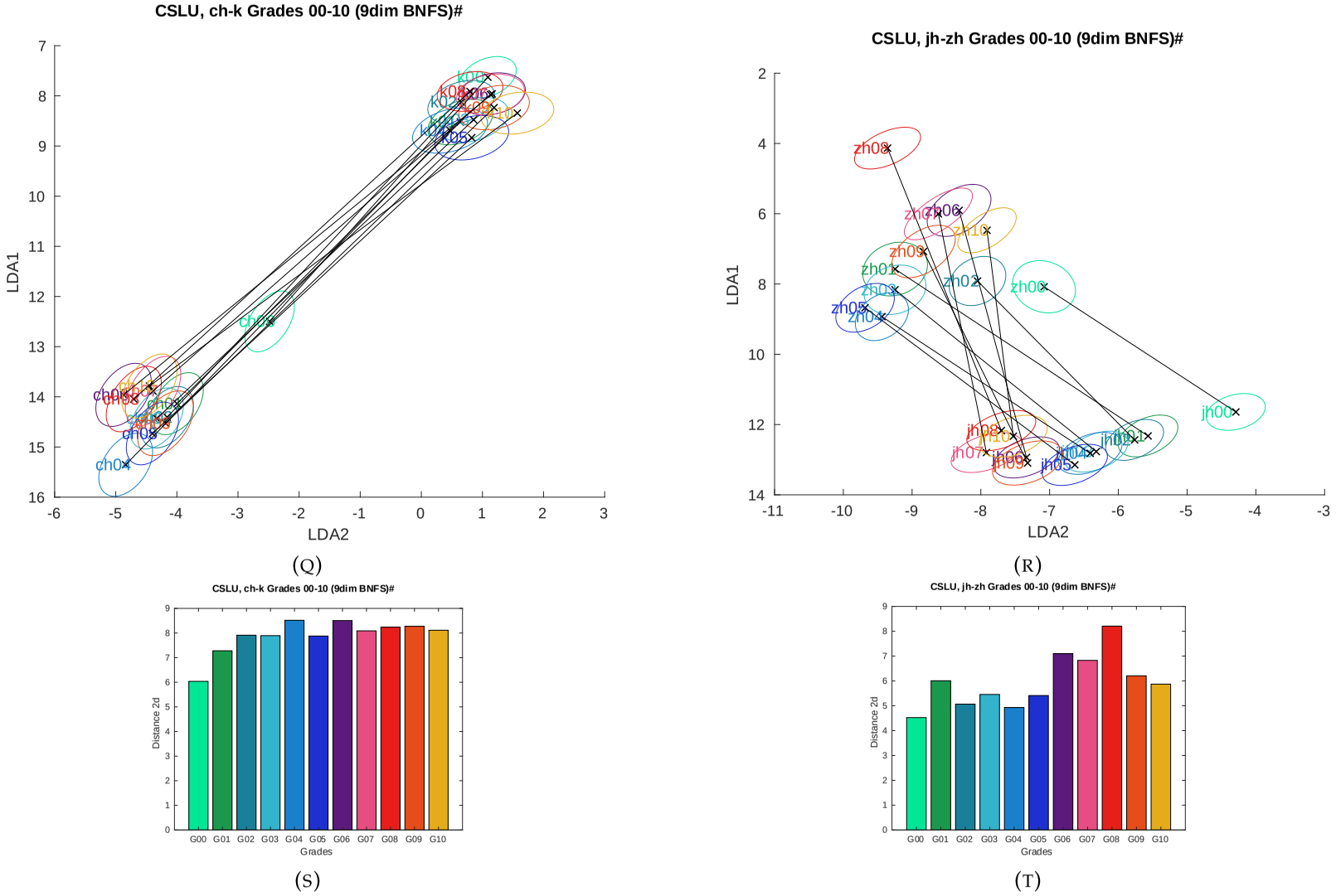
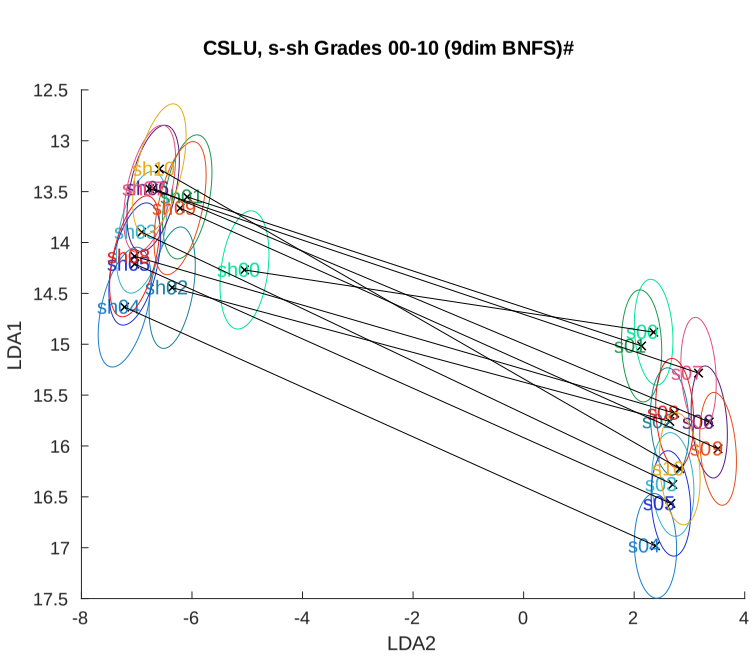
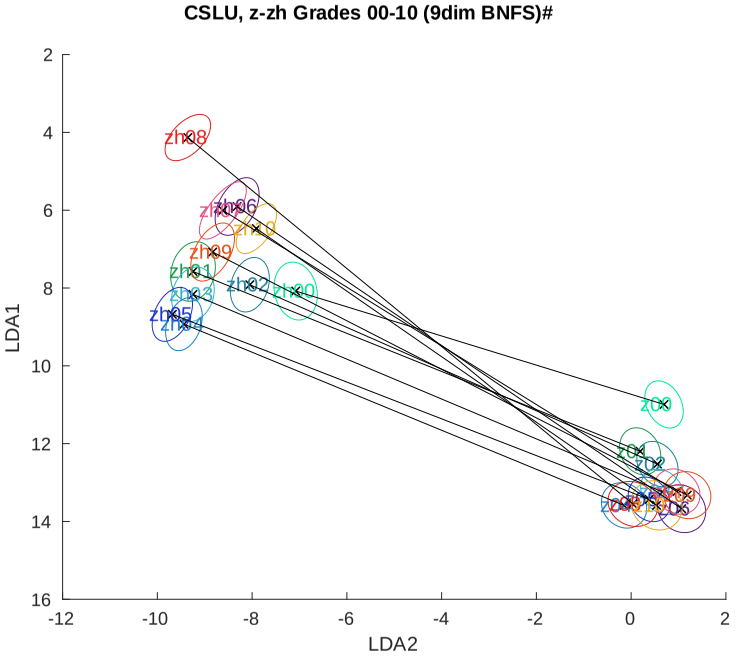


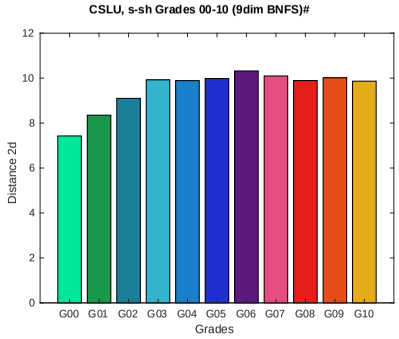
FIGURE B.1: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(a) 'increasing distance as a function of age'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.



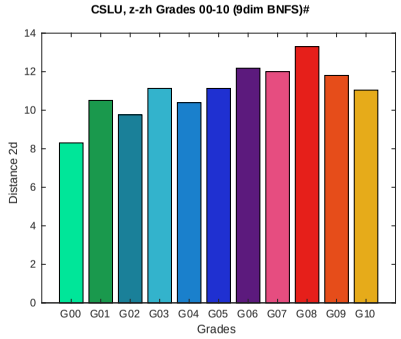
(U)



(V)



(W)



(X)

FIGURE B.1: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(a) 'increasing distance as a function of age'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.

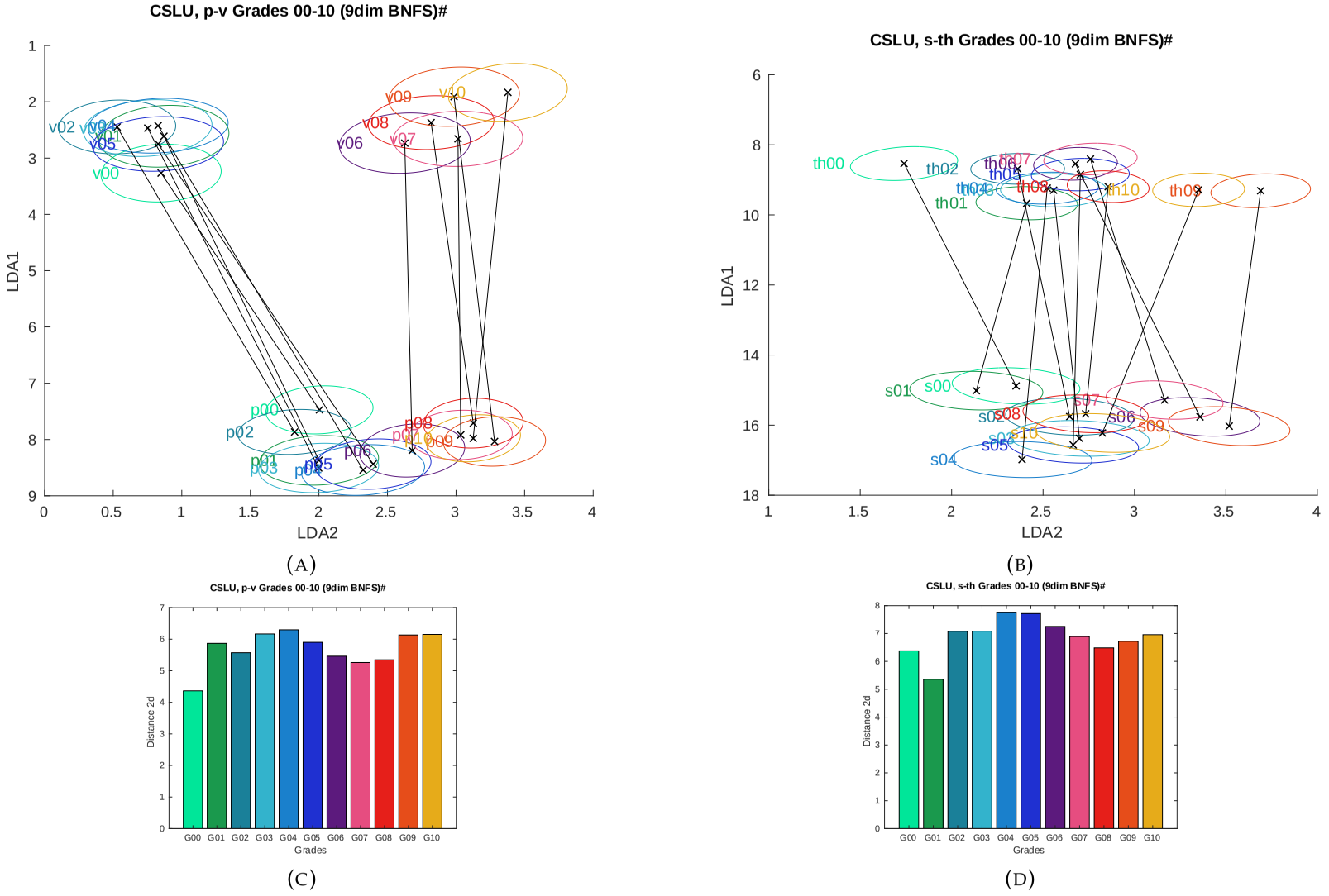


FIGURE B.2: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(b) 'constant distance'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.

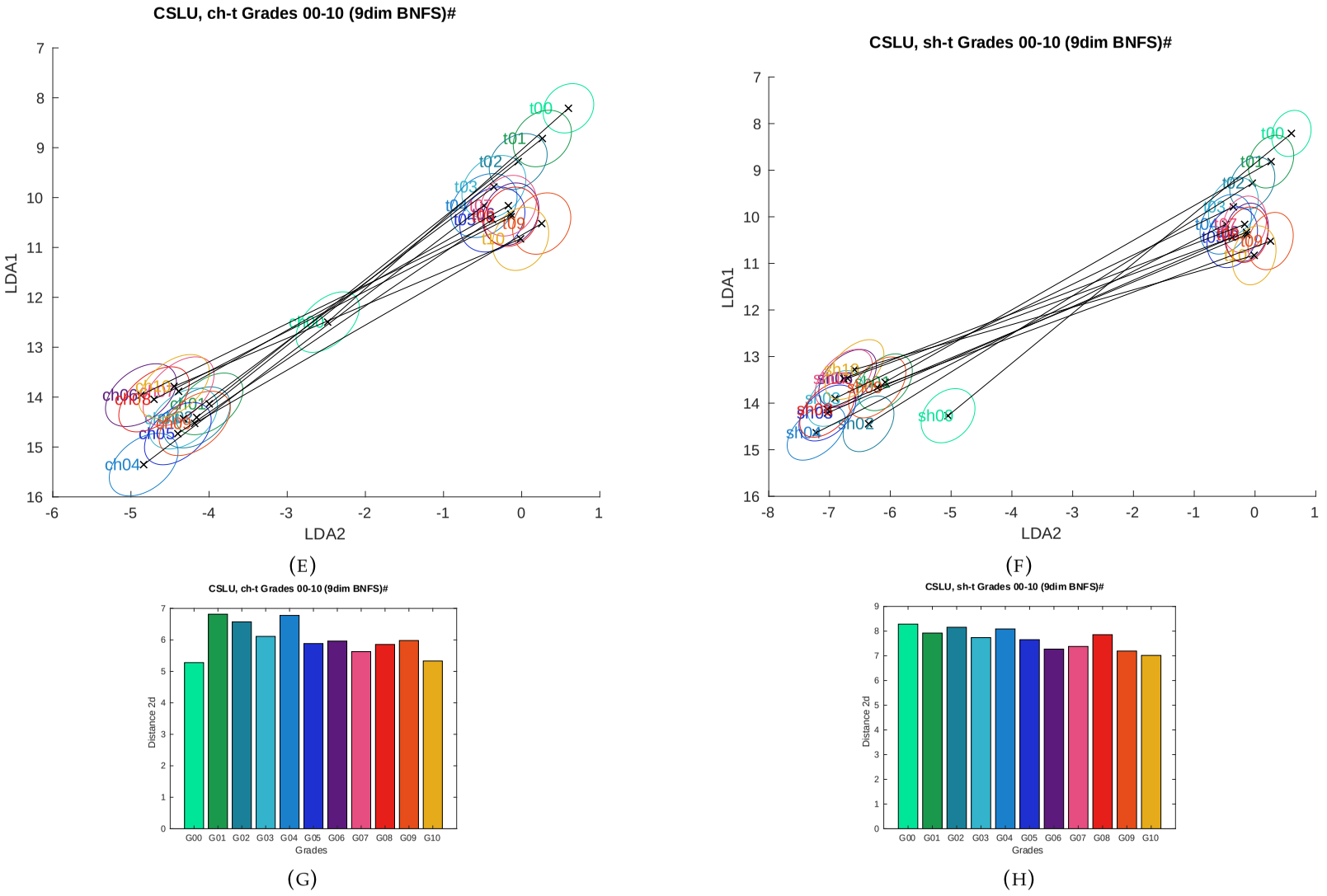
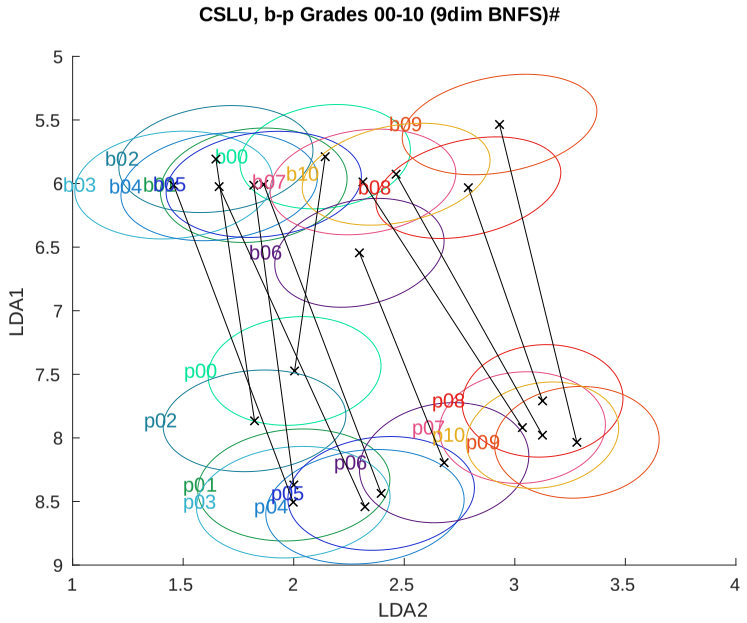
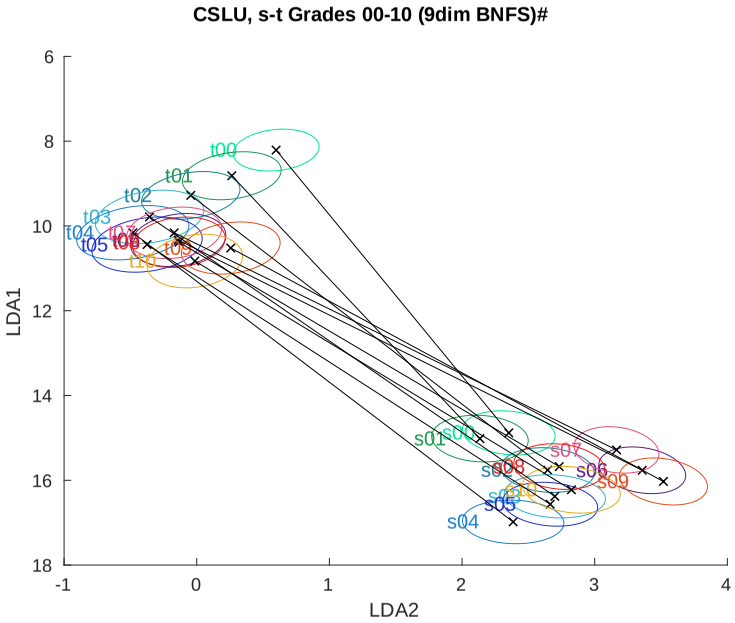


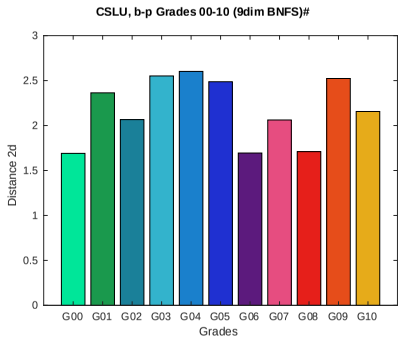
FIGURE B.2: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(b) 'constant distance'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.



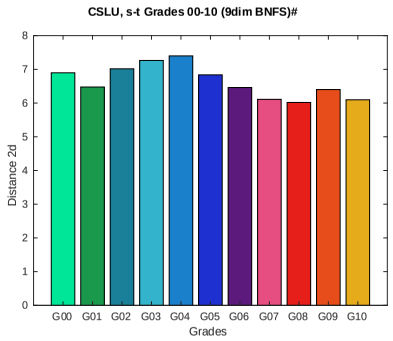
(I)



(J)



(K)



(L)

FIGURE B.2: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(b) 'constant distance'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.

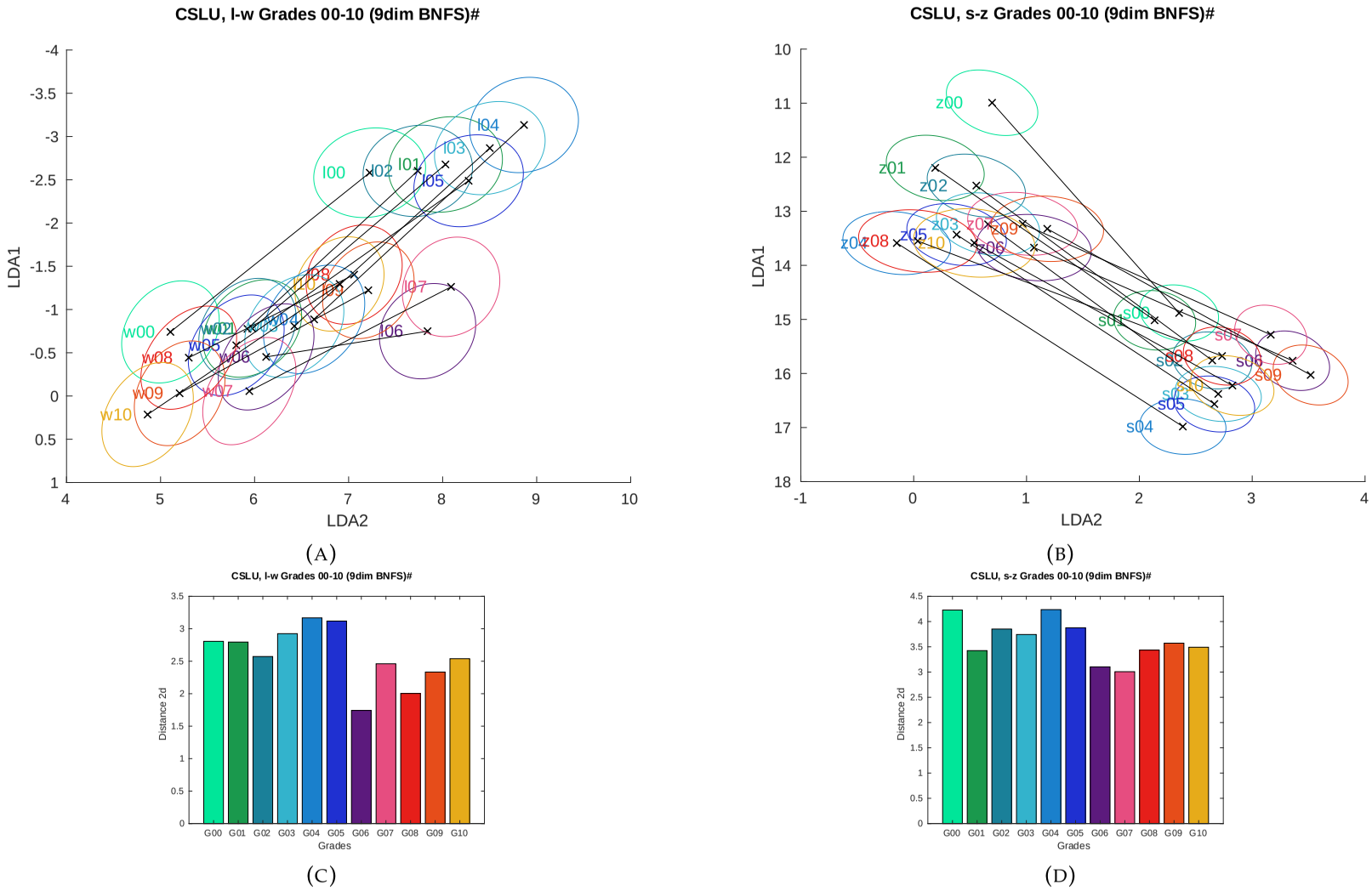
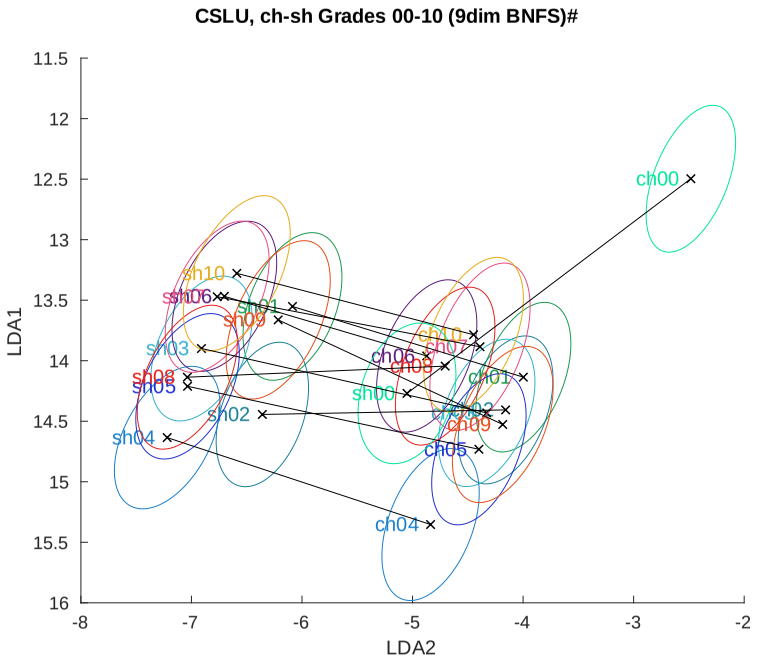
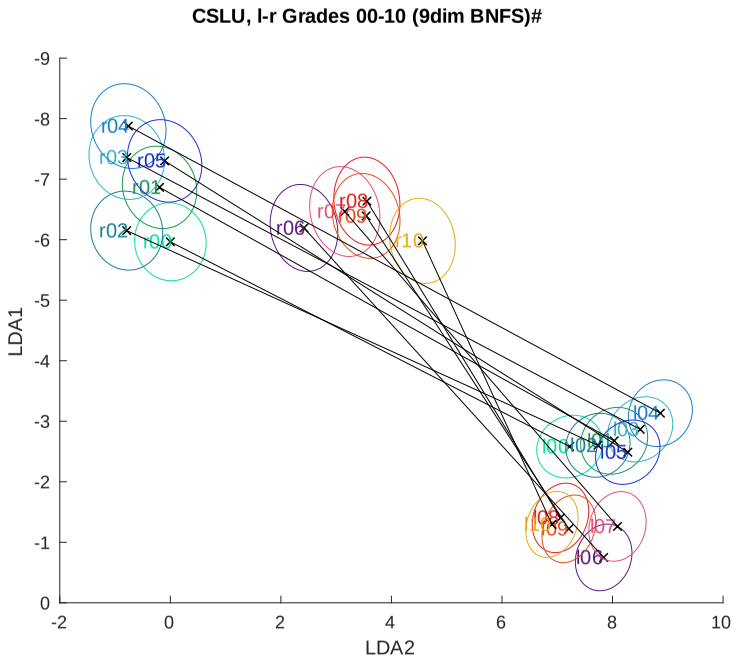


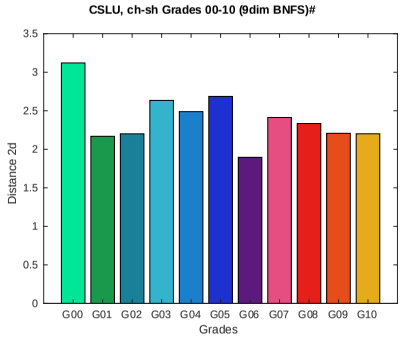
FIGURE B.3: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(c) 'decreasing distance'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.



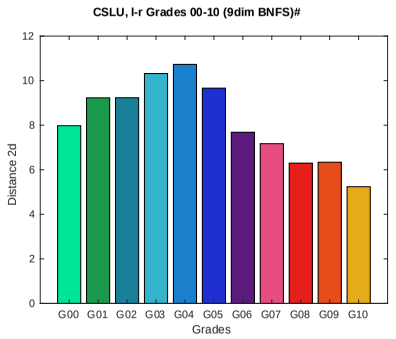
(E)



(F)



(G)



(H)

FIGURE B.3: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(c) 'decreasing distance'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.

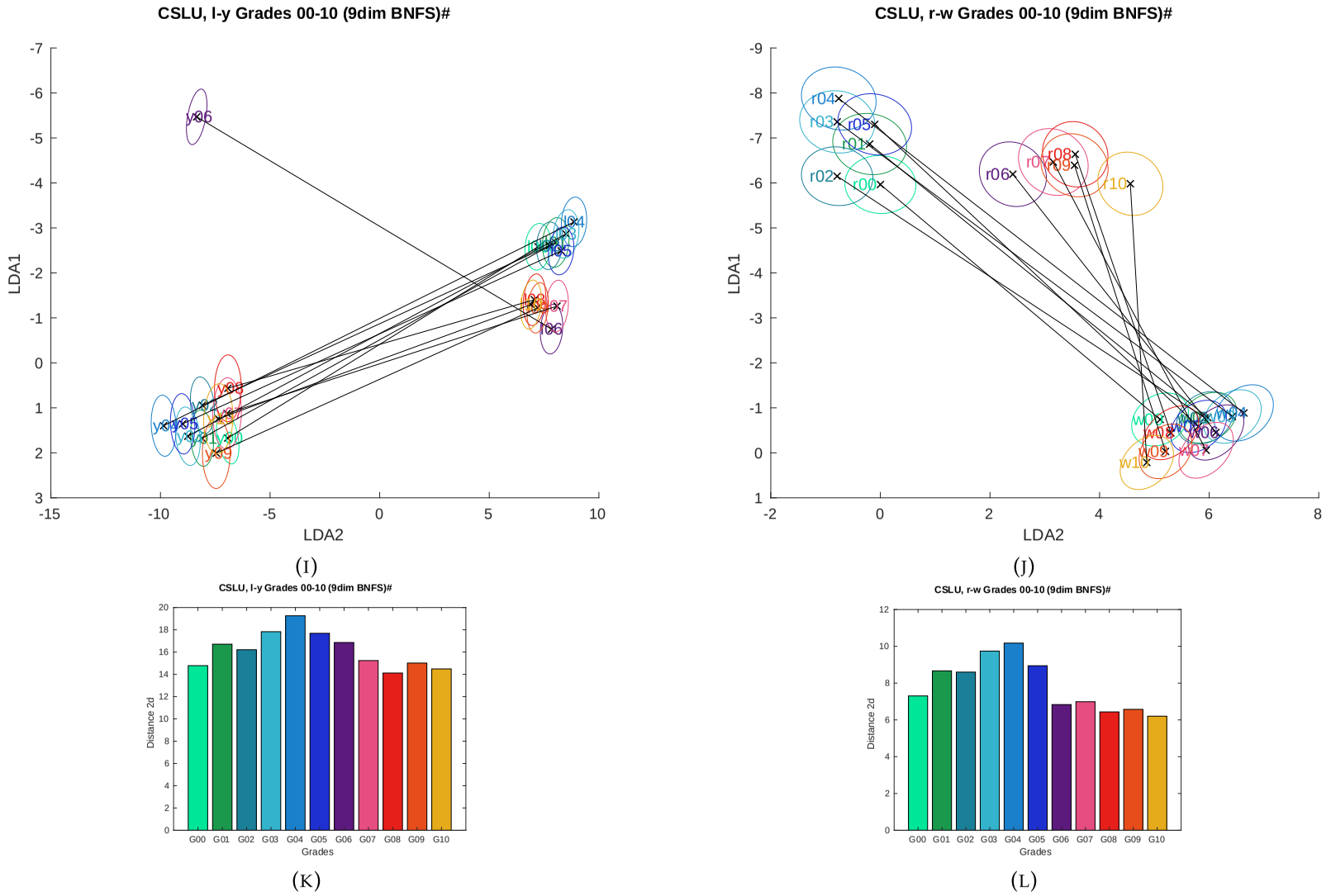
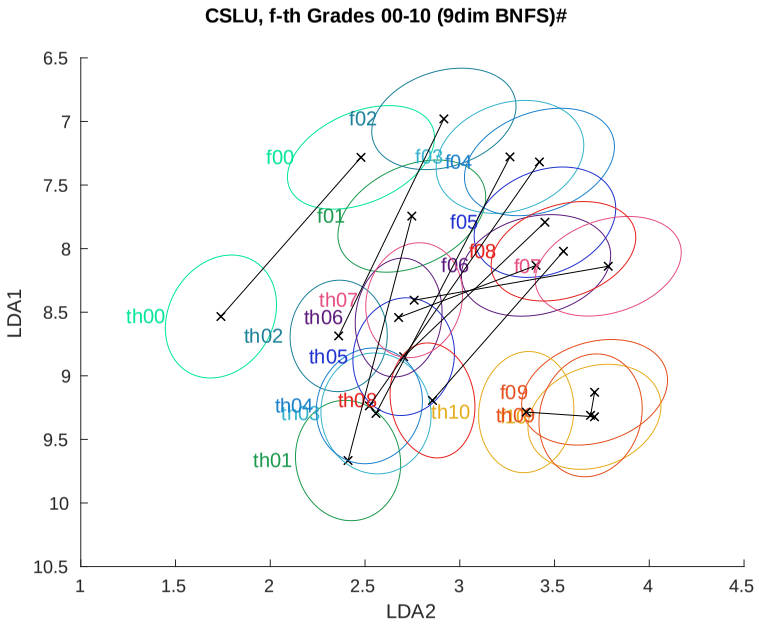
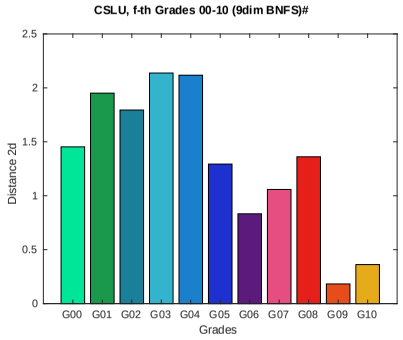


FIGURE B.3: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(c) 'decreasing distance'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.



(M)



(N)

FIGURE B.3: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(c) 'decreasing distance'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.

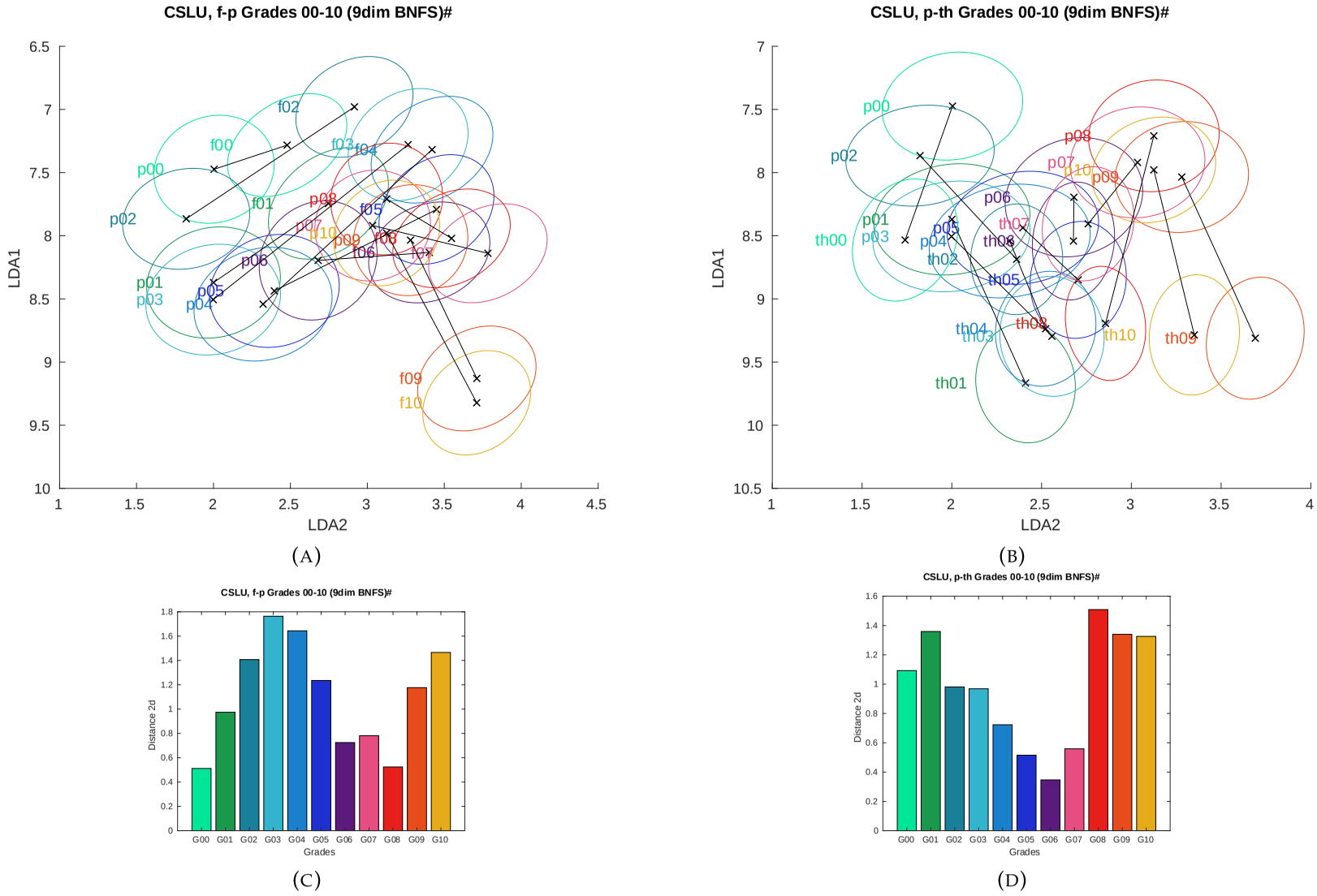


FIGURE B.4: 2-dimensional LDA projections of 9-dimensional BNFs. Gaussian ellipse plots and bar chart graphs for the pairs of phonemes which belong to category **(d) 'randomly alternating distance'**. In the ellipse plots, individual data points (crosses) and ellipses represent mean values and 0.1 standard deviation contours of the BNF features for both of the phones in each PEALA related pair from Table 2.3. The bar charts represent the Euclidean distances between the mean values of each phone in the pair as a function of age.

Bibliography

Al Moubayed, Samer and Jill Lehman (2015). "Design and Architecture of a Robot-Child Speech-Controlled Game". In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. HRI'15 Extended Abstracts. Portland, Oregon, USA: ACM, pp. 79–80. ISBN: 978-1-4503-3318-4. DOI: 10.1145/2701973.2702041. URL: http://doi.acm.org/10.1145/2701973.2702041.

Arlt, P. and T. Goodban (1976). "A comparative study of articulation acquisition as based on a study of 240 normals, aged three to six". In: 7, pp. 173–180.

Bai, L. et al. (2015). "Analysis of a Low-Dimensional Bottleneck Neural Network Representation of Speech for Modelling Speech Dynamics". In: *INTERSPEECH 2015*, pp. 583–587.

Bai, L. et al. (2018). "Exploring how phone classification neural networks learn phonetic information by visualising and interpreting bottleneck features". In: *INTERSPEECH 2018*.

Bai, Linxue (2017). "Speech Analysis Using Very Low-dimensional Bottleneck Features and Phone-class Dependent Neural Networks". PhD thesis. University of Birmingham.

Barrett, M. (1995). "Early Lexical Development". In: *The Handbook of Child Language*. Ed. by P. Fletcher and B. MacWhinney. Oxford: Blackwell Publishing. Chap. 13, pp. 362–393.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: OUP.

Bosch, Louis ten and Lou Boves (2018). "Information Encoding by Deep Neural Networks: What Can We Learn?" In: *interspeech 2018*.

Burnham, D. K. (1986). "Developmental loss of speech perception: Exposure to and experience with a first language". In: 7, pp. 207–240.

Campbell, W.M. et al. (2006). "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation". In: *icassp06*. Vol. 1, pp. I–I.

Cohen, W. and C. Anderson (2011). "Identification of phonological processes in preschool children's sinlge-word productions". In: *International Journal of Language and Communication Disorder* 46.4, pp. 481–488.

Davis, S. and P. Mermelstein (1980). "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences". In: 28.4, pp. 357–366.

Dehak, N. et al. (2011). "Front-End Factor Analysis for Speaker Verification". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4, pp. 788–798.

Dodd, B. et al. (2003). "Phonological Development:A normative study of British-English speaking children". In: *Clinical Linguistics and Phonetics* 17.8, pp. 617–643.

Eguchi, S. and I. J. Hirsh (1969). "Development of speech sounds in children". In: 257, pp. 1–51.

Eimas, P.D. et al. (1971). "Speech Perception in Infants". In: 171, 303–306.

Elenius, D. and M. Blomberg (2004). "Comparing Speech Recognition for adults and children". In: *FONETIK 2004*. Vol. 1, pp. 156–159.

Ferguson, C. A. and C. B. Farwell (1975). "Words and sounds in early language acquisition". In: 51, pp. 419–439.

Flege, J. E. and W. Eefting (1986). "Linguistic and developmental effects on the production and perception of stop consonants". In: 43, pp. 155–171.

Gangji, N., M. Pascoe, and M. Smouse (2015). "Swahili speech development: preliminary normative data from typically developing pre-school children in Tanzania". In: 50.2, pp. 151–164.

Garofolo et al., J. S. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium. Univ. Pennsylvania, Philadelphia, PA.

Gauvain, J. L. and C. Lee (1994). "Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains". In: 2, pp. 291–298. URL: citeseer.ist.psu.edu/gauvain94maximum.html.

Gerosa, M., D. Giuliani, and F. Brugnara (2007). "Acoustic Variability and Automatic Recognition of Children's Speech". In: 49, pp. 847–860.

Gerosa, M. et al. (2006). "Analysing Children's Speech: An Acoustic Study of Consonants and Consonant-Vowel Transition". In: *icassp06*. Vol. 1, pp. 393–396.

Ghai, Shweta (2011). "Addressing Pitch Missmatch for Children's Automatic Speech Recognition". PhD thesis. Indian Institute of Technology Guwahati.

Giuliani, D. and M. Gerosa (2003). "Investigating Recognition of Children's Speech". In: *icassp03*, pp. 137–140.

Golonka, Ewa M. et al. (2014). "Technologies for foreign language learning: a review of technology types and their effectiveness". In: *Computer Assisted Language Learning* 27.1, pp. 70–105. DOI: 10.1080/09588221.2012.700315. eprint: https://doi.org/10.1080/09588221.2012.700315. URL: https://doi.org/10.1080/09588221.2012.700315.

Graves, Alex and Navdeep Jaitly (2014). "Towards end-to-end speech recognition with recurrent neural networks". In: *International Conference on Machine Learning*, pp. 1764–1772.

Grunwell, P. (1981). "The development of phonology: a descriptive profile". In: 2.6, pp. 161–191.

Hämäläinen, Annika et al. (2014). "Correlating ASR Errors with Developmental Changes in Speech Production: A Study of 3-10-Year-Old European Portuguese Children's Speech". In: *Workshop on Child Computer Interaction-WOCCI 2014*, pp–1.

Hazan, V. and S. Barrett (2000). "The development of phonemic categorization in children aged 6+12". In: *Journal of Phonetics* 28, pp. 377–396.

Holm, A., S. Crossbie, and B. Dodd (2007). "Differentiating normal variability from inconsistency in children's speech: normative data". In: *International Journal of Language and Communication Disorders* 42.4, pp. 467–486.

Holmes, J.N. and W.J. Holmes (2001). *Speech synthesis and recognition*. 2nd. London and New York: Taylor and Francis.

Hori, Takaaki et al. (2017). "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM". In: *arXiv preprint arXiv:1706.02737*.

Huang, Jui-Ting, Jinyu Li, and Yifan Gong (2015). "An analysis of convolutional neural networks for speech recognition". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, pp. 4989–4993.

Ingram, D. (1974). "Phonological rules in young children". In: *Journal of child language* 1.2, pp. 49–64.

— (1979). "Phonological patterns in the speech of young children". In: *Language Acquisition*. Ed. by P. Fletcher and M. Garman. Cambridge: Cambridge University Press. Chap. 7, pp. 133–149.

Johnson, C. E. (2000). "Children's phoneme identification in reverberation and noise". In: 43, pp. 144–157.

Juang, B.-H. (1985). "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains". In: *AT&T Technical Journal* 64.6, pp. 1235–1249.

Jusczyk, P.W. and C. Derrah (1987). "Representation of Speech Sounds by Young Infants". In: 23.5, pp. 648–654.

Kenny, Patrick, Gilles Boulianne, and Pierre Dumouchel (2005). "Eigenvoice modeling with sparse training data". In: *ieeesap* 13.3, pp. 345–354.

Kent, R. (1992). "The Biology of Phonological Development". In: *Phonological Development: Models, Research, Implications*. Ed. by C. A. Ferguson, L. Menn, and C. Stoel-Gammon. York: York Press, pp. 65–90.

Kent, R. D. (1976). "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies". In: 19, 421–447.

Kitzing, Peter, Andreas Maier, and Viveka Lyberg Åhlander (2009). "Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders". In: *Logopedics Phoniatrics Vocology* 34.2, pp. 91–96.

Kuhl, P. K. et al. (2006). "Infants show a facilitation effect for native language phonetic perception between 6 and 12 months". In: 9.2, F13–F21.

Lee, S., A. Potamianos, and S. Narayanan (1999). "Acoustics of Children's Speech: Developmental Changes of Temporal and Spectral Parameters". In: 105.3, pp. 1455–1468.

— (2014). "Developmental Acoustic Study of American Englisg Diphdongs". In: *interspeech14*, pp. 1–16.

Lee, S. and R. Rose (1998). "A Frequency Warping Approach to Speaker Normalization". In: *icassp98*. Vol. 6. 1, pp. 49–60.

Leggetter, C.J. and P. C. Woodland (1995). "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models". In: 9.2, pp. 171–185.

Lehman, J and Samer Al Moubayed (2015). "Mole madness–a multi-child, fast-paced, speech-controlled game". In: *AAAI Symposium on Turn-taking and Coordination in Human-Machine Interaction. Stanford, CA*.

Li, Jinyu et al. (2015). "LSTM time and frequency recurrence for automatic speech recognition". In: *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, pp. 187–191.

Liao, Hank et al. (2015). "Large vocabulary automatic speech recognition for children". In: *Sixteenth Annual Conference of the International Speech Communication Association*.

Liporace, L.R. (1981). "Maximum likelihood estimation for multivariate observations of Markov sources". In: *IEEE Trans. Inform. Theory* IT-28, pp. 729–734.

Luria, A.R. and F.Ia. Yudovich (1959). *Speech and the development of mental processes in the child*. Staples Press.

Lust, B. (2006). *Child Language: Acquisition and Growth*. Cambridge University Press.

McLeod, S. and J. Arciuli (2009). "School-Aged Children's Production of /s/ and /r/ Consonant Clusters". In: 61, pp. 336–341.

Miao, Yajie, Mohammad Gowayyed, and Florian Metze (2015). "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding". In: *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, pp. 167–174.

Mostow, J. et al. (1994). "A prototype reading coach that listens". In: *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI94), Seattle, WA*, pp. 785–792.

Narayanan, S. and A. Potamianos (2002). "Creating Conversational Interfaces for Children". In: *icassp02*, pp. 65–78.

Olmsted, D. (1971). *Out of the Mouth of Babes: earliest stages in language learning*. The Hague: Mouton.

Palaz, Dimitri, Mathew Magimai Doss, and Ronan Collobert (2015). "Convolutional neural networks-based continuous speech recognition using raw speech signal". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, pp. 4295–4299.

Palaz, Dimitri, Mathew Magimai-Doss, and Ronan Collobert (2015). "Analysis of cnn-based speech recognition system using raw speech as input". In: *Sixteenth Annual Conference of the International Speech Communication Association*.

Pfau, T., R. Faltlhauser, and G. Ruske (2000). "A Combination of Speaker Normalization and Speech Rate Normalization for Automatic Speech Recognition". In:

Potamianos, A. and S. Narayanan (2003). "Robust Recognition of Children's Speech". In: *icassp03*. Vol. 11. 6, pp. 603–616.

— (2007). "A Review of the Acoustic and Linguistic Properties of Children's Speech". In: *icassp07*, pp. 22–25.

Povey, D. et al. (2011). "The Kaldi Speech Recognition Toolkit". In: *IEEE 2011 - Workshop on Automatic Speech Recognition and Understanding*.

Priester, G. H., W. J. Post, and S. M. Goorhuis-Brouwer (2011). "Phonetic and phonemic acquisition: Normative data in English and Dutch speech sound development". In: 75, pp. 592–596.

Rabiner, L. R. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". In: *Proceedings of the IEEE*. Vol. 77. 2, pp. 257–286.

Reynolds, D. A. (1992). "A Gaussian mixture modeling approach to text independent speaker identification". PhD thesis. Georgia Institute of Technology.

Reynolds, D.A. and R.C. Rose (1995). "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". In: *IEEE Trans. Speech and Audio Proc.* 3.1, pp. 72–83.

Romeo, R., V. Hazan, and M. Pettinato (2013). "Developmental and gender-related trends of intra-talker variability in consonant production". In: *Acoustical Society of America* 134.5, pp. 3781–3792.

Rummelhart, D.E., G.E. Hinton, and R.J. Williams (1986). "Learning representations by back-propagating errors". In: *Nature* 323, pp. 533–536.

Russell, M. et al. (2000). "The STAR system: an interactive pronunciation tutor for young chidren". In: 14.2, pp. 161–175.

Russell, M. J., S. D'Arcy, and L. Qun (2007). "The effects of bandwidth reduction on human and computer recognition of children's speech". In: *IEEE Signal Processing Letters* 14.12, pp. 1044–1046.

Russell, Martin and Shona D'Arcy (2007). "Challenges for computer recognition of children's speech". In: *Workshop on Speech and Language Technology in Education*.

Saaristo-Helin, K., S. Kunnari, and T. Savinainen-Makkonen (2011). "Phonological development in children learning Finnish: A review". In: 31.3, pp. 342–363.

Safavi, S., M. Russell, and P. Jancovic (2018). "Automatic speaker, age-group and gender identification from children's speech". In: *Computer Speech and Language* 50, pp. 141–156.

Sak, Haşim et al. (2015a). "Fast and accurate recurrent neural network acoustic models for speech recognition". In: *arXiv preprint arXiv:1507.06947*.

Sak, Haşim et al. (2015b). "Learning acoustic frame labeling for speech recognition with recurrent neural networks". In: *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, pp. 4280–4284.

Sander, E. (1972). "When are speech sounds learned?" In: 37, pp. 55–63.

Scharenborg, Odette et al. (2018). "Visualizing Phoneme Category Adaptation in Deep Neural Networks". In: *interspeech18*.

Shivakumar, P.G. et al. (2014a). "Improving Speech Recognition for Children's Speech using Acoustic Adaptation and Pronunciation Modeling". In: *Proc. Workshop on Child-Computer Interaction, WOCCI*.

Shivakumar, Prashanth Gurunath et al. (2014b). "Improving speech recognition for children using acoustic adaptation and pronunciation modeling". In: *WOCCI*, pp. 15–19.

Shobaki, K., J. P. Hosom, and R. A. Cole (2000). "The ogi kids speech corpus and recognizers". In: *ICSLP 2000 - Sixth International Conference on Spoken Language Processing*. Vol. 4, pp. 258–261.

Simon, C. and A. Fourcin (1978). "Cross-language study of speech pattern learning". In: *Journal of the Acoustical Society of America* 63, pp. 925–935.

Sinha, Rohit and Shweta Ghai (2009). "On the use of pitch normalization for improving children's speech recognition". In: *Tenth Annual Conference of the International Speech Communication Association*.

Smit A., B. et al. (1990). "The Iowa Articulation Norms Project and its Nebraska Replication". In: 55, pp. 779–798.

Smith, A. and H. Zelaznik (2004). "Development of functional synergies for speech motor coordination in childhood and adolescence". In: 45.1, pp. 22–33.

Soltau, Hagen, Hank Liao, and Hasim Sak (2016). "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition". In: *arXiv preprint arXiv:1610.09975*.

Sosa, A. V. and C. Stoel-Gammon (2006). "Patterns of intra-word variability during the second year of life". In: 33, pp. 31–50.

— (2012). "Lexical and phonological effects in early word production". In: 55, pp. 596–608.

Stampe, D. (1969). "The acquisition of phonetic representation". In: *Papers from the Fifth Regional Meeting, Chicago Linguistic Society*. Ed. by R. Binnick et al., pp. 126–133.

Stark, R.E. (1978). "Features of infant sounds: the emergence of cooing". In: 5.3, 379–390.

Stoel-Gammon, C. (1985). "Phonetic inventories, 15-24 months: a longitudinal study". In: 28, pp. 505–512.

Templin, M. (1957). *Certain language skills in children: their development and interrelationships*. Vol. 26. Institute of Child Welfare Monographs. Minneapolis, MN: University of Minnesota Press.

Trigeorgis, George et al. (2016). "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network". In: *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, pp. 5200–5204.

Vihman, M. M. (1996). *Phonological Development: The origins of language in the child*. Blackwell.

— (2014). *Phonological Development: The first two years*. Wiley-Blackwell.

— (2017). "Learning words and learning sounds: Advances in language development". In: 108, pp. 1–27.

Walsh, B. and A. Smith (2002). "Articulatory movements in adolescents". In: 45, 1119–1133.

Wang, Hongcui, Christopher J. Waple, and Tatsuya Kawahara (2009). "Computer Assisted Language Learning system based on dynamic question generation and error prediction for automatic speech recognition". In: *Speech Communication* 51.10. Spoken Language Technology for Education, pp. 995 –1005. ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2009.03.006. URL: http://www.sciencedirect.com/science/article/pii/S0167639309000430.

Weber, P. et al. (2016a). "Interpretation of low dimensional neural network bottleneck features in terms of human perception and production". In: *INTERSPEECH 2016*.

Weber, P. et al. (2016b). "Progress on Phoneme Recognition with a Continuous State HMM". In: *icassp16*.

Weninger, Felix et al. (2015). "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR". In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer, pp. 91–99.

Werker, J. F. and R. C. Tees (1983). "Developmental Changes Across Childhood in the Perception of Non-native Speech Sounds". In: 37.2, pp. 278–286.

— (2002). "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life". In: 25, 121–133.

Werker, J. F. et al. (1981). "Developmental Aspects of Cross-Language Speech Perception". In: 52.1, pp. 349–355.

Wilpon, J. and C. Jacobsen (1996). "A Study of Speech Recognition for Children and the Elderly". In: *icassp96*.

Young, S. J., J. J. Odell, and P. C. Woodland (1994). "Tree-Based State Tying for High Accuracy Acoustic Modelling". In: *Proceedings of the Workshop on Human Language Technology*, pp. 307–312.

Young, S. J. et al. (1997). *The HTK Book*. v2.1. Cambridge, UK: Entropic Camb. Res. Lab.

Yu, D. and L. Deng (2014). *Automatic speech recognition: A deep learning approach*. Springer.

Zeyer, Albert et al. (2017). "A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition". In: *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, pp. 2462–2466.

Zhang, Ying et al. (2017). "Towards end-to-end speech recognition with deep convolutional neural networks". In: *arXiv preprint arXiv:1701.02720*.