Exploring Phone Recognition in Pre-verbal and Dysarthric Speech

Syed Sameer Arshad

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

University of Washington

2019

Committee:

Gina-Anne Levow

Gašper Beguš

Program Authorized to Offer Degree:

Department of Linguistics

University of Washington

Abstract

Exploring Phone Recognition in Pre-verbal and Dysarthric Speech

Chair of the Supervisory Committee:

Dr. Gina-Anne Levow

Department of Linguistics

In this study, we perform phone recognition on speech utterances made by two groups of people: adults who have speech articulation disorders and young children learning to speak language. We explore how these utterances compare against those of adult English-speakers who don't have speech disorders, training and testing several HMM-based phone-recognizers across various datasets. Experiments were carried out via the HTK Toolkit with the use of data from three publicly available datasets: the TIMIT corpus, the TalkBank CHILDES database and the Torgo corpus. Several discoveries were made towards identifying best-practices for phone recognition on the two subject groups, involving the use of optimized Vocal Tract Length Normalization (VTLN) configurations, phone-set reconfiguration criteria, specific configurations of extracted MFCC speech data and specific arrangements of HMM states and Gaussian mixture models.

Preface

The work in this thesis is inspired by my life experiences in raising my nephew, Syed Taabish Ahmad. He was born in May 2000 and was diagnosed with non-verbal autism as well as apraxia-of-speech. His speech articulation has been severely impacted as a result, leading to his speech production to be sequences of babbles. He is receptively bilingual in English and Urdu, and communicates via Assistive and Augmentative Communication (AAC) apps on his iPad. He undergoes speech therapy, behavior therapy and special-needs education throughout the week. He has come a long way in life and he gets better at communicating every day. As a linguist and an engineer, I have observed his babbles and read up a lot on speech disorders and the way they are treated by speech-language pathologists. Through discussions with my nephew's speech-language pathologist and reading texts on the topic, I identified an under-served aspect of the speech-recognition and speech-therapy research literature that I feel I can contribute to, using the wisdom from my lived experiences as a caregiver for my nephew and as a volunteer worker for all the social organizations that he participates in. My hope is that the work in this thesis inspires other researchers to take this kind of research further, in order to create new methods of speech-therapy that can help people like my nephew, who is transitioning from childhood to adulthood with a severe speech-articulation disorder as well as a neuro-cognitive disorder that impacts his social communication.

I would like to thank Professors Gina Levow, Emily Bender, Fei Xia, Richard Wright, Ellen Kaisse, Toshiyuki Ogihara and Gašper Beguš for all their help.

Table of Contents

List of Tables

	Model 28 on all scenarios that were applicable.	
24	The number of human-transcribed phones clustered by Age At Recording and Language Environment.	100
25	The results for the second set of experiments we performed for Phase 5. We explored 9 values for VTLN Warp Factor and 17 values for Age At Recording for the English-environment subset of the entire Talkbank dataset, running Model 28 on all scenarios that were applicable.	102 - 104
26	The results for the third set of experiments we performed for Phase 5. We explored 9 values for VTLN Warp Factor and 17 values for Age At Recording for the French-environment subset of the entire Talkbank dataset, running Model 28 on all scenarios that were applicable.	104 - 106
27	The results for the third set of experiments we performed for Phase 5. We explored 9 values for VTLN Warp Factor and 17 values for Age At Recording for the German-environment subset of the entire Talkbank dataset, running Model 28 on all scenarios that were applicable.	106 - 110

List of Figures

1. Introduction

Human speech is a complex feat involving the coordination of multiple systems of articulation and acoustics to accomplish very specific vocalization tasks in the context of a linguistic environment. Many of us take this process for granted and are only confronted by the complexity of this achievement when something goes wrong somewhere.

This study explores the use of computers in recognizing the speech utterances of people who have dysarthria and also the speech utterances of young pre-verbal children. This is an understudied aspect of speech recognition, which usually caters to commercial applications involving the speech of able-bodied adults. Children are not expected to call customer-service and navigate a spoken-dialog system. Adults and children with speech-articulation disorders are not the target market for voice-controlled intelligent-assistants . Moreover, even if children without speech-articulation disorders are targeted as users for any of these systems, they are expected to be fully verbal children beyond the age of 7. Therefore, we have found very little literature on the subject of speech recognition for these speakers. Before embarking on the approach of this study, it is important to provide some motivation for the concept, which will shed light on why the approaches described in this paper were taken.

1.1 Motivation for this study

Speech articulation disorders such as dysarthria and apraxia-of-speech remind us that many things have to go right for human speech to happen. Speech language pathology is the field that focuses on studying these disorders and focuses on providing services to patients to help them improve their speech and language production skills within their social environments. This field is relatively new but has an extensive library of publications and research, as well as a professional association, known as the American Speech-Language-Hearing Association (ASHA).

1.2 The work of speech-language pathologists

When talking about how speech therapists work with their patients, we need to keep in mind how painstaking this kind of work can be. An SLP therapy session usually lasts one hour, unless they are doing an assessment, in which case, these can go as long as three hours, with interactions being complexly navigated due to ongoing communication challenges faced by the patient (Paul & Norbury, 2012).

We know from the speech language pathology literature (Paul & Norbury, 2012) that studying the language acquisition pathway of children without speech disorders can give us a lot of clues on how to treat people who have speech disorders. There has been research (Paul & Norbury, 2012) showing that the sequence of achievement milestones that speech-disorder patients take is similar to the development pathway of young children who are learning to speak, all things considered. This gives us a clue about

how important it can be for a speech language pathologist to know how both scenarios play out, in order to plan speech-therapy sessions accordingly.

When a patient presents with speech sound disorders of any kind, part of the initial assessment of the patient involves acquiring a phonetic inventory. A phonetic inventory is the list of phones that can be produced by the patient, which the SLP has verified by hearing them a certain number of times (sometimes just once) in a specific sample of speech. The speech sample is collected by either recording or observing about 10 to 15 minutes of speech events from the patient. These speech events could be units of speech elicited by the clinician. They could also be units of connected speech that occur in a conversation with the clinician. Any elicitation or guided conversation conducted for this purpose is done with the goal of getting the patient to produce sounds across the phoneme inventory of the language of instruction, which is usually English in the United States. However, ASHA currently has a recommendation towards being sensitive to patients who come from bilingual and multicultural families (Paul & Norbury, 2012). There is some documented research effort to take bilingualism into account in acquiring a phonetic inventory for patients (Paul & Norbury, 2012). Some clinicians in the United States go further and acquire a “phonemic” inventory, which accounts for whether or not the patient has acquired the expected correct allophone distributions for the phonemes in their dialect of English. Getting this representative list of the phones that the patient can produce is an efficient starting point for teaching new phones. This list can also be used towards helping them learn how to imitate the phones that they can already produce, especially if someone produces them in their presence in an elicited

repetition task. We know such tasks are central towards early-stage speech therapy activities (Paul & Norbury, 2012).

However, for some patients who have neurocognitive disorders like non-verbal autism, preparing a phonetic inventory can be challenging. Sometimes, a clinician cannot elicit any speech from the patient and neither can the clinician engage the patient in conversation. Any attempts to do so may be responded to with smiles and nods, but no elicited speech. If there is a speech event, it may not be imitating anything the clinician said and may not be intelligible either. It could be a couple of seemingly unrelated interspersed syllables, or a squeal or even just laughter. Ideally, an SLP clinician would have to look up the known phonemes of all languages in the patient's language environment, even for languages that the SLP clinician doesn't know. This would be important, especially in English-speaking countries, so as not to come up with a phonetic inventory that is English-centric. There are some phones that don't exist in English and they should be looked out for (and looked after) because the goal is to align the patient with their language environment.

An example of a typical patient would be a verbal child who is semi-intelligible, or a talkative child who has an articulation issue with producing rhotic sounds. However, a patient who has non-verbal autism and/or childhood apraxia of speech, cannot be expected to produce all the phones that they can possibly produce, upon elicitation, in the 15 minutes of contiguous time that an SLP would have allotted to acquiring their speech sample. This means that any attempt from an SLP to acquire a phonetic inventory

in the regular manner would lead to a very incomplete list. It is impractical to go about doing this in the usual way that a clinician would do so for a typical patient.

Once a phonetic inventory is acquired, other diagnostic criteria are checked for before the SLP would make a diagnosis on whether or not a speech articulation disorder exists. If a language impairment also exists, a therapy plan is made to work on acquiring proficiency with various language forms while also acquiring phones that are missing from the patient's repertoire (Paul & Norbury, 2012).

1.3 Phonetic Inventories: Easier said than done.

There is a lot of data in a patient's natural day-to-day spoken utterances that can be considered useful towards producing a phonetic inventory. Sometimes the caregivers of the patient are interviewed about which phones have been heard by them from the patient's past speech utterances, but this data is not verifiable first-hand information. It is also prone to error because caregivers are not trained to hear and make judgments regarding phonetic transcriptions. In the event that a reliable phonetic inventory cannot be acquired in an elicited clinical setting, some clinicians may opt to spend several hours with the patient, in different naturalistic situations, in different social environments such as the home, the classroom or during a hobby activity, monitoring the speech utterances of the patient. In the event that the clinician cannot spend this kind of time with the patient, sometimes the caregivers are asked to record the speech utterances of the patient over a prolonged period of time, performing different activities, and provide those recordings to the clinician. Recordings like these involve the clinician spending a lot of time analyzing and transcribing the patient's speech. A minute of recorded speech can

take several minutes to analyze and/or transcribe. Most clinicians will avoid trying to tally every instance of every phone and instead will check for a phone occurring once or twice to confirm that it can be produced by the patient. Most clinicians stop transcribing after hearing around thirty minutes of recorded speech. Even though this gives better data than than trying to work with 15 minutes of mostly silence that would have emerged from a recording in a therapy session, it is still quite a burden. Sometimes a clinician will outsource this work to an assistant. In spite of this, even thirty minutes of a patient's recorded speech arranged by a caregiver can possibly be unrepresentative of the articulated sounds that a patient is capable of. There is no way to predict which possible articulations will happen at any given time of the day. A caregiver could be expected to record hours and hours of patient speech, but that would be adding a further burden to the clinician (or their assistants) to transcribe this speech. The cost of this burden gets added to the charges for the clinician's services, which is something that either has to be approved by an insurance provider or paid for out-of-pocket by the patient's family. So a compromise needs to be achieved between the quality of the data and the expense incurred to acquire that data. We can see here that even the basic initial task of collecting a phonetic inventory for children with non-verbal autism is fraught with obstacles that cannot be practically overcome. This is a contributing factor towards why SLP clinicians are unanimous in not recommending speech production therapy for patients with non-verbal autism, opting instead to focus on language therapy that does not involve using the vocal tract, instead involving the use of sign language, the Picture Exchange

Communication System (PECS) (Bondy & Frost, 1994) and speech-producing AAC devices to take the role of the vocal tract.

1.4 Using speech-recognition technology in speech-language pathology

With the creation of regimented and usable speech recognition packages, such as HTK and Kaldi, more opportunities to apply speech recognition to novel problems have emerged. However, most speech recognition systems work with a language model built from known word occurrence patterns for the target language, and an acoustic model built from an extensive phonetically-transcribed audio dataset that contains representative data for how certain words sound. These models are often combined to form a Hidden Markov Model (HMM) lattice to perform a phone-by-phone and/or word-by-word recognition for utterances that are assumed to be intelligible to a human listener in the language environment being considered. Sometimes, other machine learning approaches such as neural networks are applied to improve accuracy and exploit domain-specific patterns in the data.

A literature review on the intersection between speech recognition and speech-language pathology (SLP) yields only a few instances where speech recognition was used in the realm of SLP. Many of these ideas involve apps that computationally compare the speech samples of word utterances provided by patients to recorded samples of exemplar speech from speakers who don't have disorders. A feedback interaction is provided to the patient to let them know how close their utterance was to the defined ideal for

intelligibility training. This is done as an at-home activity that a clinician would ask a patient to carry out independently between sessions. For patients with non-verbal autism and other neurocognitive disorders that can impact speech production, a comparison-based approach like this is not something that can be used to build a phonetic inventory, due to the aforementioned struggle to elicit structured responses from the patient. One would require long durations of naturalistic speech recorded (with caregiver consent) via a wearable audio recording device, such as the numerous models that are affordably available via e-commerce platforms. This would be paired with a tool that provides a phonetic transcription of a prolonged duration of recorded speech from the patient, if the transcription has a high enough accuracy to be reliable.

Spending three or more hours on manually transcribing thirty minutes of babble speech is a difficult thing to get an insurance provider to approve payments for, when three other patients could receive one-hour therapy sessions for the same amount of billable time. But an affordably-priced automated transcription system that can do the same task in minutes without human supervision can definitely free up the clinician (and their assistants) for other tasks. A tool like this can be provided to speech language pathologists as a cloud-based service, allowing many hours of patient recordings to be uploaded by caregivers and quickly processed by modern cloud-based computer infrastructure to extract phonetic transcriptions, which can be analyzed to produce a phonetic inventory for the patient in a format that can be consumed for analysis by the speech language pathologist. This allows a better-informed speech-production therapy plan to be prepared. The entire process can be repeated after some amount of

therapy has been administered, to see if target phones are being produced spontaneously by the patient in unsupervised settings.

This will have to be a purpose-built piece of technology that is different from regular consumer-level speech recognition systems, which are designed to transcribe and detect entire words and sentences. The patients being discussed here are not uttering sentences or even intelligible words. But they are uttering phones that can be detected.

2. Related Work

Due to the multi-disciplinary approach of this paper, we will be referencing work from multiple-fields, such as language-acquisition, speech language pathology, speech-recognition, and phonetics.

2.1 Related studies on child speech production

The acquisition of speech production has classically had two theoretical approaches. One based on "competency" and one based on "performance" (Vihman, 1996). Studies focusing on competency (see review in Goldsmith, 1990) involve principles from generative phonology, distinctive features, Universal Grammar and innate phonological knowledge. Studies focusing on performance (Kent, 1984; Lindblom et al.., 1993) involve acoustic phonetics, the biological aspects of motor articulation for speech as well as the functional and perceptual aspects of speech production. In the context of an automatic speech recognition project like ours, a competence-based phonological approach is difficult to accomplish because it involves attempting to measure what is going on at the phonological level. Focusing on the physically measurable data from phonetic information ends up being a more tractable way to explore this space computationally, particularly when the speech involves young children who are still learning language. There has been a study that contrasts the two approaches, (Davis et al., 2002), which concluded that phonological approaches "do not adequately establish the proposed mental entities in infants of this age, and are nonexplanatory in the sense of not considering the causes of the structures and constraints that they posit". In a context of speech-language pathology, we would want to follow

approaches that yield insights about what is going right and what is going wrong. A performance-based phonetic approach ends up providing these insights for us.

2.1.1 Related studies on babbling

Studying the babbling of children has been proven to be useful when it comes to predicting future speech and language disorders (Oller et al., 1999). Extensive investigations on infant speech production (Oller 1980) and the organization of babbling (Davis & MacNeilage, 1994) have yielded insights on the various acoustic and articulatory shapes and forms that canonical child babbles can take in early life. Moreover, from observing the surprise-responses of infants to new sound sequences, we now know that even 8-month old infants can infer the segmentation of words from fluent English speech, from just 2 minutes of exposure (Saffran et al., 1996). There have also been longitudinal studies on the vocalizations of young children in English-environments, such as Selby et al., 2000, that document the timelines of phones emerging into usage. In a study (Davis & MacNeilage 1990) that involved quantitative approaches, Davis & MacNeilage reported that high front vowels in the environment of alveolar consonants are favored, hinting at consonant-vowel interdependence between the 14 to 20 month age range. Figure 1 below, (taken from Selby et al., 2000) shows information on which English vowels emerge in the 15 to 36 month range, across 4 young children participating in the study.

Group vowel inventory at each age interval based on 75% occurrence across the four children

Age (months)	Inventory	Size
15	ɑ, ɪ, ʊ, ʌ	4
18	ɑ, i, u, ʊ, ʌ, ɔ, æ	7
21	ɑ, i, ɪ, ɛ, u, o, ʌ, ɔ	8
24	ɑ, i, ɪ, ɛ, e, u, o, ɔ, æ	9
36	ɑ, i, ɪ, ɛ, e, u, ʊ, o, ʌ, ɔ, æ, ɝ	12

Figure 1: From Selby et al., 2012, a table showing vowels that have a 75% chance of occurrence among 4 English-environment children, across five age milestones during the toddler years.

One of the most definitive studies on this topic is Davis et al., 2002 (Davis et al., 2002), which collected extensive data on child babbling that we use in this thesis as well. This study analyzed recorded data from children between the ages of 9 months and 25 months. Their findings, which are oriented towards children in an English-language environment, are quite revealing. 58.7% of the utterances were monosyllables while 34.1% were disyllables. The most dominant consonant places-of-articulation were labials (41.3%), coronals (36.1%) and dorsals (9.94%) with the remaining sounds mostly being glottal stops and other sounds. The most dominant manners-of-articulation were oral stops (55.3%), nasals (21.6%) and glides (8.5%), with the rest (14.4%) being liquids, fricatives, affricates and glottal stops as well as the [h] sound. Of these, glottal stops and [h] accounted for 6.1% and fricatives, affricates and liquids accounted for 8.3%. When it came to vowel height, the vowels were similarly distributed across high, mid and low dimensions with little variability across the children. When it came

to vowel frontness, central vowels were the most common (37.6%), followed by front vowels (36.3%) and then back vowels (23.5%). Similar studies have emerged to shed light on the speech development of children living in other language environments, such as Persian (Fotuhi et al., 2016) and Brazilian Portuguese (Teixeira & Davis, 2002). Follow-up work has focused on children in bilingual environments, such as a recent study involving fraternal twins growing up in an English-Serbian language environment (Zlatić et al., 1997).

2.1.2 Speech-recognition for children

Throughout the 1990s and early 2000s, the body of work on speech recognition for children has been small. We first see the research community publishing about it in 1996 (Russel et al., 1996) and noticing that it is something to focus on specifically due to higher error rates (Wilpon & Jacobsen, 1996). The first dedicated papers on the topic emerged in 1997 (Potomianos et al., 1997) with further investigations into robust speech recognition for children emerging in the early 2000s (Potomianos & Narayanan, 2003). These papers note the distinct higher levels of speech-recognition errors that occur on child speech for children under 18, from models trained on adult speech. They note how these errors reduce the closer a child gets to young adulthood. They also discuss the vowel space shape differences between adults and children while also noting differences in vocal-tract length and variability in acoustic features. Almost none of these papers focus on babbles and pre-verbal speech from children who are learning to talk, mostly because the commercial applications for speech-recognition for this age group are not completely convincing to researchers. There has been recent research on “child-directed speech” and

creating computational models of how young children could learn how to speak. In 2016, Ma et al. looked into creating a phone-embedding model (Ma et al., 2016) to learn word segmentation and phonological structures from child-directed speech from a dataset on CHILDES that contained child-directed speech from mothers, that was phonetically transcribed. Such models give us interesting clues on how children could actually be learning how to speak the language of their caretakers, but these models may not be useful to speech-language pathologists who are working at a different level than people making computational models of child-language acquisition.

2.2 Related studies on dysarthria

Dysarthria is defined by ASHA as a speech disorder that occurs due to muscle weakness in the face, lips, tongue, throat and in the muscles for breathing (Paul & Norbury, 2012). There are several types of dysarthria, classified by the impacted area of the vocal tract and the specific neurological disturbance that causes the impact. Sometimes, multiple components in the respiratory, laryngeal and supralaryngeal articulatory subsystems can be impacted (Kent et al., 1999), which can happen in conditions like Parkinson's disease, amyotrophic lateral sclerosis and stroke (Kent et al., 1999). The symptoms for this condition can vary across speakers, but they include "strained phonation, imprecise placement of the articulators, incomplete consonant closure resulting in sonorant implementation of many stops and fricatives, and reduced voice onset time distinctions between voiced and unvoiced stops" (Kent et al.., 1999). There have been important works of research focusing on the acoustic qualities of dysarthric speech since the 1980s. In a 1989 study by Kent et al., (Kent et al., 1989), 19 acoustic-phonetic contrasts

were identified that were likely to be affected by dysarthric impairment that would influence speech intelligibility. The paper goes on to systematically provide a structure for intelligibility testing for dysarthric speech. Follow-up research emerges in the 1990s with important papers outlining seminal studies in acoustic analysis of speech in the context of detecting speech disorders (Kent, 1990; Kent & Read, 1992) and further methods of acoustic studies for dysarthria in particular (Kent et al., 1999). Research continues into the 2000s focusing on specific types of dysarthria, such as spastic dysarthria (Roy et al., 2001). At this point, this psychobiology and phonetic research is able to inform the speech-pathology field about how to treat the various forms of dysarthria. For example, in Roy et al., 2001 (Roy et al., 2001), the researchers talk about identifying the "loci of intelligibility deficit", the "features of deviant speech whose improvement would lead to the greatest gains in treatment" and the expected changes that lead to an improvement over a 30-month recovery period. By mapping out the speech-pathology landscape, several studies in the speech-pathology field emerged on this topic (Finch et al., 2018). A recent review of the speech-pathology literature for non-progressive forms of dysarthria can be found in Finch et al., 2018 (Finch et al., 2018)

2.2.1 Speech-recognition for dysarthric speech

Early work during the 1980s on this topic recognizes the distinctly lower recognition accuracy for dysarthric speech (Rodman et al., 1985) and proposes several ideas to remedy this issue that do not take the articulatory behavior of dysarthria in account (Rudzicz 2012). Some of the most compelling pieces of research emerge in the 2000s and 2010s, where novel speaker adaptation methods and signal-processing

adjustments are employed to improve the accuracy of speech-recognition for dysarthric speech. Hosom et al., 2003, attempts to use Gaussian-mixture mapping techniques to transform the audio mathematically into a more intelligible form. Kain et al., (Kain et al., 2007) propose concatenating original unvoiced segments with synthesized voiced segments to improve intelligibility. Tolba & Torgoman (Tolba & Torgoman 2009), improved intelligibility by running transformations on the first and second formants in dysarthric speech. Rudzicz (Rudzicz 2013) proposed a sequence of reverse-transformations on the speech data that themselves come from mathematic models of the dysarthria-specific articulations themselves, while Sharma & Hasegawa-Johnson (Sharma & Hasegawa-Johnson 2013), use a novel speaker-adaptation algorithm in conjunction with the maximum a posteriori (MAP) algorithm to bring an improvement in intelligibility for dysarthric speech. Around this time, the TORGO experiments at the University of Toronto, (Rudzicz et al., 2012) focused on creating a publicly available dataset of dysarthric speech (which we make use of in this thesis) with phonetic transcriptions and phone-aligned timing information, as well as articulatory information through various sensors planted on the mouth and face. These novel approaches have led to very exciting breakthroughs in reducing error-rates while also contributing to insights in the speech-language pathology domain (Finch et al., 2018).

2.3 Studies on phone recognition

Phone-recognition, as a speech recognition problem, is quite different from word recognition. Both involve using large amounts of transcribed utterances to produce models of uttered speech that can be used to detect forms. However, words have the added complexity of having alternate pronunciations.

They also benefit from extra predictive power via n-gram techniques that exploit information on how common a word is and how common sequences of words are as well, when it comes to the language in question. Word-recognition as a problem can also be constrained to focus on specific domains based on the nature of the problem being solved. For example, one can create a special-purpose digit-recognition system that just specializes in recognizing the ten English digit words, knowing that a speaker has been asked to limit their utterances to just speaking digit words. This kind of domain-specific constraint can rarely be arranged when it comes to phone recognition scenarios.

Furthermore, phone recognition for people with disordered speech becomes more difficult problem because most of the datasets available to prepare language-models and acoustic-models come from people who don't have speech disorders. For patients with speech disorders and for young pre-verbal children, it is important to correctly identify what was produced in the context of what was intended to be produced. However, even intention is sometimes hard to figure out, especially when we are trying to do phone recognition for very young children who are naturally producing canonical babbles. In all these situations, language-models don't help us much and all we can do is prepare an acoustic model that can attempt to accurately transcribe what was said. Preparing such an acoustic model involves an extensive phonetically-transcribed speech dataset that is also phone-aligned.

2.3.1 Phone recognition with the TIMIT corpus

Many researchers in the academic and corporate worlds have come up with their own such datasets that are not publicly available. But among the publicly available datasets, the most widely-cited and most

widely known such dataset is the TIMIT dataset (Garofolo et al.. 1993). The TIMIT dataset provides speech data from 630 speakers, across 8 dialects of American English, each of whom reads 10 sentences, providing 6300 utterances. The prompts for these sentences consist of 2 dialect sentences (SA), 450 phonetically compact sentences (SX) and 1890 phonetically-diverse sentences (SI). Everything is fully transcribed at the word-level and the phone level, with time-alignments provided for words and also for phones. This dataset took years of effort to enrich it to this level of detail and has become a de-facto standard for benchmarking speech-recognition experiments and tasks. Very few other publicly available datasets exist with this level of detail and richness. The phones of the TIMIT dataset are shown in Tables 1a to 1h on the next page, alongside their International Phonetic Alphabet equivalents.

Category	Timit phone	IPA equivalent	Notes
Stop Closures	bcl	[bˈ]	Just the closure portion
	dcl	[dˈ]	
	gcl	[gˈ]	
	pcl	[pˈ]	
	tcl	[tˈ]	
	kcl	[kˈ]	
	dx	[ɾ]	Tap
	q	[ʔ]	Glottal stop

Table 1a: Stop closures in the TIMIT phone set.

Category	Timit phone	IPA equivalent	Notes
Stop Releases	b	[b]	Just the release portion
	d	[d]	
	g	[g]	
	p	[p]	
	t	[t]	
	k	[k]	

Table 1b: Stop releases in the TIMIT phone set.

Category	Timit phone	IPA equivalent	Notes
Fricatives	s	[s]	
	sh	[ʃ]	
	z	[z]	
	zh	[ʒ]	
	f	[f]	
	th	[θ]	
	v	[v]	
	dh	[ð]	

Table 1c: Fricatives in the TIMIT phone set.

Category	Timit phone	IPA equivalent	Notes
Semivowels and Glides	l	[l]	
	r	[ɹ]	
	w	[w]	
	y	[j]	
	hh	[h]	
	hv	[ɦ]	
	el	[l̩]	

Table 1d: Semivowels and glides in the TIMIT phone set.

Category	Timit phone	IPA equivalent	Notes
Nasals	m	[m]	
	n	[n]	
	ng	[ŋ]	
	em	[m̩]	Syllabic
	en	[n̩]	Syllabic
	eng	[ŋ̩]	Syllabic
	nx	[ɾ̃]	Tap

Table 1e: Nasals in the TIMIT phone set.

Category	Timit phone	IPA equivalent	Notes
Vowels	iy	[i]	
	ih	[ɪ]	
	eh	[ɛ]	
	ey	[eɪ]	Diphthong
	ae	[æ]	
	aa	[ɑ]	
	aw	[aʊ]	Diphthong
	ay	[aɪ]	Diphthong
	ah	[ʌ]	
	ao	[ɔ]	
	oy	[ɔɪ]	Diphthong
	ow	[oʊ]	Diphthong
	uh	[ʊ]	
	uw	[u]	
	ux	[ʉ]	
	er	[ɝ]	Rhotic and Syllabic
	ax	[ə]	
	ix	[ɨ]	
	axr	[ɚ] or [ɹ̩]	Rhotic and Syllabic
	ax-h	[ə̥]	Devoiced

Table 1f: Vowels in the TIMIT phone set.

Category	Timit phone	IPA equivalent	Notes
Affricates	jh	[dʒ]	
	ch	[tʃ]	

Table 1g: Affricates in the TIMIT phone set.

Category	Timit phone	IPA equivalent	Notes
Others	pau	N/A	Pause
	epi	N/A	Epenthetic silence
	h#	N/A	Begin/End marker

Table 1h: Other symbols in the TIMIT phone set.

Much of the speech recognition experiments that have been done with the TIMIT corpus involve the use of Hidden Markov Models (HMMs). Sometimes they are used alone (Lee & Hon, 1989), sometimes they are used in a hybrid configuration with artificial neural networks (ANNs) (Rose & Momayyez, 2007; Scanlon et al.., 2007; Siniscalchi et al.., 2007), and sometimes they are used in a hybrid configuration with Condition Random Fields (CRFs) (Morris & Fosler-Lussier, 2008).

If we're going to explore phone recognition on the most widely known and more widely used phonetically-transcribed and phonetically-time-aligned speech recognition dataset, it is important to discuss the history of phone recognition with this dataset in particular, something that was the subject of a review paper by Lopes & Perdigão (Lopes & Perdigão, 2011). In 1989, Lee and Hon (Lee & Hon, 1989) conducted phone recognition experiments on the newly released TIMIT dataset using discrete-HMMs, linear prediction cepstral coefficients as features and a bigram language model, with phones from the TIMIT dataset being modelled as 1450 right-context diphones. The 61 TIMIT phones (shown in Tables 1a

to 1h) were folded into each other to form 39 phone classes that would be the recognition targets. Table 2 below shows the mapping of how this folding was carried out. It is relevant to share this table here because the technique used in our study took inspiration from this material.

Before	**After**
aa, ao	aa
ah, ax, ax-h	ah
er, axr	er
hh, hv	hh
ih, ix	ih
l, el	l
m, em	m
n, en, nx	n
ng, eng	ng
sh, zh	sh
uw, ux	uw
pcl, tcl, jcl, bcl, dcl, gcl, h#, pau, epi	sil
q	-

Table 2: Replacement criteria for phone-folding used in Lee & Hon, 1989.

With this adjustment, 23 TIMIT phones are removed from the transcriptions while the “sil” symbol representing silence is introduced to the transcriptions. After training their TIMIT phone recognizer system against the resulting 39 identification targets, they achieved a correctness[1] rate of 73.80% and an accuracy[2] rate of 66.08% using 160 utterances from one test set. In 1992, Young revisited this work, this time performing phone recognition experiments using a state-tying approach using triphone models

1 The percentage of phones that were correctly identified.

2 The number of all phones minus the number of insertion errors minus the number of deletion errors minus the number of substitution errors, times 100%

(Young, 1992) while switching to a different method to create features for the recognition system involving the use of standard Mel-frequency cepstral coefficients (MFCCs), their log energy and their first order regression coefficients (deltas). The best results from this experiment were 73.7% correctness and 59.9% accuracy on 160 sentences randomly picked from the TIMIT test set. In the review of research literature that was done for our study, we discovered that for a purely HMM-based phone recognition system, these were the best results so far. Anything better than these results was obtained either without an HMM or through hybrid methods featuring an HMM and another machine learning mechanism, such as neural networks or conditional random fields. These results from HMM-only experiments have set a benchmark in the industry and we will use them to assess our own work in this study. There have been several other attempts to improve upon phone recognition on the TIMIT dataset. In their review paper (Lopes & Perdigão, 2011), Lopes & Perdigão provide extensive information on various attempts to use non-HMM and HMM-hybrid approaches after Lee & Hon 1989, that report better results. We would refer the reader to this paper by Lopes & Perdigão to take a look at the descriptions of these experiments and their results, which give a good history of the research on phone recognition with the TIMIT dataset from 1989 to 2011. In more modern times, the recent experiments done by Lohrenz et al.. (Lohrenz et al.., 2018) involved the use of convolutional neural networks (CNNs) and HMMs to achieve a correctness rate of 83.09%, without the use of a language-model to provide context, for the same 39 phones as identification targets. There have also been experiments that don't feature HMMs at all, such as the work

from Abdel-Rahman et al., (Abdel-Rahman et al., 2011) where deep belief networks (DBNs) alone are able to obtain a 77% accuracy in phone recognition on the TIMIT test set.

2.3.2 Other phone recognition approaches

Due to the extensive popularity of the TIMIT corpus, most phone recognition studies in English focus on it for the purpose of benchmarking and comparisons. To find studies that don't involve the TIMIT corpus, one needs to look at phone recognition work done on other languages. Recent work on Italian phone recognition has emerged (Cosi 2016) where deep neural networks were employed via the Kaldi toolkit (Povey et al., 2011) on a 29-phone set for Italian speech data from the ArtiPhon corpus (Badino 2016). A lot of the work has focused on performing cross-lingual experiments and multilingual phone models. For example, Lamel & Gauvain (Lamel & Gauvain, 1993) trained phone recognition HMMs on speech data from the WSJ-CSR corpus (Paul & Baker, 1992), which is in English, and the BREF corpus (Lamel et al., 1991), which is in French. Their phone recognition results suggested that phone recognition trained and tested on French leads to better results (23.6% error with the BREF corpus) than a similar approach done on English (30.1% error with the WSJ-CSR corpus), though this might just be due to the BREF dataset having transcriptions and recordings of a better quality. A more extensive study involving multi-lingual phone models appears from Köhler in 2001 (Köhler 2001), where corpora from six languages (American English, French, German, Italian, Portuguese, Spanish) brought in from the SpeechDat database (Hoge et al., 1997) and the Macrophone database (Taussig & Bernstein, 1994) are used to identify 232 language-dependent and context-independent phone models, which are then clustered and organized in three

different approaches to see which approach yields the best results using an HMM-based phone-recognizer.

2.4 How our study is different

Our study attempts to explore how similar the phone recognition results of child speech are with the phone recognition results of disordered speech, using inspiration from speech-language pathology literature (Paul & Norbury, 2012) that suggests that the therapy-pathway for people who have disordered speech has parallels to the speech-acquisition pathway of very young children. We want to come up with results that show similar phone recognition performance between both sets of subjects, when the same phone recognition model is applied to both scenarios. We will limit our scope to dysarthria as the single speech-disorder of choice that we will focus our work on. We will purposely focus on adults with dysarthria and not children with dysarthria. We will attempt to make comparisons with other language-environments, if we are able to easily acquire datasets accordingly within our budget. We will also explore best practices towards adapting the child speech data in a manner that draws a fair comparison to adult dysarthric speech.

3. Research Questions

We can see from the history of phone recognition experiments above that the current state-of-the-art (Lohrenz et al.., 2018) still has a long way to go in creating a phone recognition system for the English language that can be considered reliable enough for clinical work (the best-case accuracy is 83.09%). Moreover, this body of research mostly focuses on English language environments only and the voices in the TIMIT phone set are all from adults who don't have speech disorders.

Because this is a Master's thesis, we definitely did not set out to create a system that can beat the phone recognition results of state-of-the-art systems, in multiple languages, even for child-speech and even for disordered speech. Such as task is out of scope in the present work. Instead, we focus on using the TIMIT corpus and some kind of purpose-built speech recognition system to inform us on how "recognizable" the speech of young children and the speech of people with speech disorders really can be, using the TIMIT corpus as a starting point for our training data.

In our search for datasets containing the transcribed speech of young children as well as the transcribed speech of people with speech disorders, we were unable to find a publicly available dataset that was as voice-diverse as the 630-speakers that the TIMIT corpus has.

What we did find was the CHILDES database (MacWhinney 2000) providing access to child speech data from various other child language acquisition studies that have been published in various journals over the years. Some of these provide phonetic transcription at the phone-level but none of them provided time-aligned phonetic transcription. We sifted through many of these datasets and identified four of them that we could use in our study as representation of canonical babbling in childhood, from various children at different points of early life. We go into detail on these datasets later in this paper.

We also found the TORGO corpus (Rudzicz et al., 2012), prepared by the University of Toronto for the purposes of a study in the speech articulation of people who have dysarthria. A part of the phonetically transcribed recordings of elicited utterances from these patients with dysarthria were made available by the TORGO research team at the University of Toronto to the public via their team website. This material proved useful as a representation of disordered speech for the purposes of our study.
We were able to use our own University of Washington membership to the Linguistic Data Consortium (LDC) to get access to TIMIT corpus materials for out study.

In order to limit the scope of this study to the size of a Master's thesis, we opted not to introduce the usage of non-HMM-based speech techniques in this work. We decided to focus our research questions and experiments on a purely HMM-based speech recognition system, in order to get a ballpark figure on how the different sets of data can be used to predict for each other and themselves. We do not intend to

replicate the techniques of state-of-the-art approaches. Therefore, we avoided the use of neural networks and conditional random fields in this study.

With the TIMIT corpus, the child-language-acquisition recordings from CHILDES and the material from the TORGO corpus and our knowledge of HMM-based speech recognition techniques, we are able to formulate the following research questions:

1. How well can a TIMIT-trained phone recognizer recognize phones produced by very young children, who are in the canonical babbling stage of development?
2. How well can a TIMIT-trained phone recognizer recognize phones produced by people with dysarthria?
3. Are some phones less recognizable than others? If so, what is the implication of this for questions 1 and 2?
4. How well would a phone-recognizer trained on child-utterances recognize child utterances it has not seen before?

4. Choice of Tools

We made use of several software tools to reach our goals.

4.1 HTK

We focused on using HTK (Young et al., 2006) for creating our HMM-based phone-recognizer systems, utilizing Baum-Welch expectation maximization with the Viterbi algorithm. The exact settings used for each experiment will be provided in later sections of this paper. We also used HTK's HCopy tool to convert .wav files to their MFCC counterparts and also the HResults tool to figure out how the resulting phonetic transcriptions from our phone-recognizer systems performed by calculating values for correctness, accuracy and providing confusion matrices for every phone we were taking under consideration.

4.2 Python and Bash

Numerous scripts were written in Python (https://www.python.org/) and Bash (https://www.gnu.org/software/bash/) to prepare the data for processing in HTK as well as to move intermediate forms of the data to the right places between various runs of HTK tools. This allowed us to create an end-to-end pipeline to streamline our procedures and automate a lot of the work.

4.3 SoX

We utilized the audio-processing libraries of SoX (http://sox.sourceforge.net/), available for use in Bash as well as in Python, to slice relevant spans of audio and organize the creation of utterance-sample .wav files from larger recordings.

5. Methodology and Approach

With the three datasets available to us, we attempted several different experiments to explore the recognizability of phones uttered by speakers who have dysarthria and speakers who are very young. We did this by designing phone-recognizers in HTK, with various configurations and parameters, with various training data, and testing it out on various arrangements of our testing data.

There were two pipelines of work that we employed to create our phone-recognizers:

The first pipeline was set up to handle training data that did not have time-aligned transcription at the phone-level, such as our Talkbank dataset. We will call this Pipeline A. The second pipeline was set up to handle data that had time-aligned transcription at the phone-level, such as our TIMIT dataset. Pipeline B is expected to yield better results due to the existence of time-aligned phonetic transcription, which allows for more precise extractions of phone information for training the model. These pipelines are closely tied to the workflow of the HTK software. After consulting The HTK Book (Young et al., 2006) extensively, we discovered sequences of best-practices for each pipeline, which we have summarized in Tables 3a and 3b.

There are three key differences between Pipeline A and Pipeline B:

- HInit is employed in Pipeline B and not in Pipeline A.
- HRest is employed in Pipeline B and not in Pipeline A.
- HErest is run ten times in Pipeline B and is run only 3 times in Pipeline A.

Pipeline A
Step 1. Decide on an MFCC configuration.
Step 2. Convert training data and testing data to this MFCC configuration.
Step 3. Decide on an HMM configuration.
Step 4. Construct a prototype HMM accordingly, fill it with dummy values.
Step 5. Run HTK's HCompV tool with this prototype HMM, against the training data, creating a prototype that contains the global means and variances of the training data.
Step 6. Form an HMM list file, consisting of copies of the resulting prototype from Step 5, as many copies as there are phones you wish to recognize.
Step 7. Run HTK's HERest tool on the result of Step 6, against the training data, to re-estimate the values for each phone. Do this for 3 iterations in sequence. The resulting HMM list is now your phone recognition model.
Step 8. Use this model to run HTK's HVite tool to run the Viterbi algorithm to make phone recognition predictions against the test data. The result is an output transcription, attempting a phonetic transcription for all the utterances in the test data.
Step 9. After performing any necessary post-processing on this output transcription, feed it to HTK's HResults tool, which compares the output transcription with the human-transcriptions in specified HTK Label files, making calculations for correctness, accuracy, hits, insertions, deletions and substitutions, generating a phone confusion matrix as well.

Table 3a: The sequence of actions for Pipeline A.

Pipeline B
Steps 1 to 5. Same as Pipeline A
Step 6. Run HTK's HInit tool against the training data, for every phone we wish to recognize, which uses phone-level time-alignment information in the training set to set suitable initial values for all the numerical values for each phone's HMM. The resulting set of phone HMMs is now properly initialized.
Step 7. Run HTK's HRest tool on each of the initialized phone HMMs, against the training data, which will use the phone-level time-alignment information in the training set to re-estimate the values further. Set the iteration limit to 100, which forces the HRest tool to keep iterating for re-estimations for each phone's HMM until numerical values converge, or until 100 iterations occur, whichever happens first.
Step 8. Form an HMM list file by aggregating all the phone HMMs prepared at the end of Step 6.
Step 9. Run HTK's HERest tool on the result of Step 8, against the training data, to re-estimate the values for each phone. Do this for 10 iterations in sequence. The resulting HMM list is now your phone recognition model.
Step 10. Same as Step 8 in Pipeline A
Step 11. Same as Step 9 in Pipeline A

Table 3b: The sequence of actions for Pipeline B.

To address our research questions, it is important to recognize that we can make two groups of phone-recognizers: one trained on TIMIT data and one trained on Talkbank data. The Torgo dataset has too few speakers and too few utterances to create enough speaker and utterance diversity to form a reliable model for phone recognition. We can see that Pipeline A will be required to prepare any models trained on the Talkbank training dataset, while Pipeline B will be required to prepare any models trained on the TIMIT training data. We know, already, that anything prepared with Pipeline B will probably perform at a superior level. But it is a valuable exercise to appreciate, in a numerical way, the advantage that we get from a time-aligned phonetically-transcribed dataset. Therefore, our experiment plan will proceed in the following phases:

- Phase 1: Use Pipeline A to train the best phone recognition model we can with our Talkbank training dataset, testing against our Talkbank testing dataset.
- Phase 2: Use Pipeline B to train the best phone recognition model we can with the TIMIT training dataset, testing against the TIMIT test dataset.
- Phase 3: Use the best phone recognition model we got out of Phase 2 to test against the Talkbank testing dataset and any variants of this dataset as needed.
- Phase 4: Using the best phone recognition model we got out of Phase 2 to test against the TORGO dataset.
- Phase 5: Using the best phone recognition model we got out of Phase 2, explore optimal parameters for age-sensitive phone recognition for the young children in the Talkbank dataset.

6. Gathering and Preparing Data

The three sets of data that we focused on for this study will be referred to as the “TIMIT dataset”, the “TORGO dataset” and the “TalkBank dataset”.

6.1 The TIMIT Dataset

As mentioned earlier, the University of Washington membership with the Linguistic Data Consortium allows us to access the TIMIT corpus of 6300 utterances involving 630 voices uttering 10 sentences each. All of the utterances are provided in the form of a .wav file, with 16-bit signed integer encoding for each sample, at a sample rate of 16000 samples per second. The default training-testing split in the data provided by the corpus is what we also used in our work. The training set contains 4620 utterances while the testing set contains 1680 utterances. We proceeded to use HTK (Young et al.., 2006) to convert these .wav files into their MFCC equivalents. The exact settings used to generate these MFCC data points are provided in Appendix A. There were various configurations used that we will describe in detail in later sections. We decided to relabel all the TIMIT phones as their IPA equivalents. The reason for this is that the Talbank Dataset (which we will describe later) consisted of child speech that was transcribed in IPA. We wanted to make the IPA phone set the default phone set for our study, because IPA symbols are standard in the linguistics community as well as in the speech language pathology community. Along with this change in the TIMIT transcription labelling, a few changes were made to the original set of 61 phones, which were folded into each other in a technique similar to Lee & Hon 1989, demonstrated in

Table 2 above, with a few adjustments, which you can see in Table 4. The only places we diverged from the Lee & Hon method was not converting “ao” to “aa” and not converting “zh” to “sh”. We also had to make a few specifications to account for some TIMIT phones being labelled with capital letters and some curiously presented TIMIT phones we found in the transcriptions, such as “riy” and “rtcl”.

Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
aa	ɑ	d	d	f	f	KCL	sil	Q	sil	V	v
AA	ɑ	D	d	F	f	l	l	r	ɹ	w	w
ae	æ	DCL	sil	G	g	L	l	R	ɹ	W	w
AE	æ	dcl	sil	g	g	m	m	riy	i	y	j
ah	ə	dh	ð	GCL	sil	M	m	rtcl	sil	Y	j
AH	ə	DH	ð	gcl	sil	n	n	s	s	z	z
ao	ɔ	dx	ɾ	h#	sil	N	n	S	s	Z	z
AO	ɔ	DX	ɾ	hh	h	ng	ŋ	sh	ʃ	zh	ʒ
aw	a_ʊ	eh	ɛ	HH	h	NG	ŋ	SH	ʃ	ZH	ʒ
AW	a_ʊ	EH	ɛ	hv	h	NOI	sil	si	sil		
ax	ə	el	l	HV	h	noi	sil	sil	sil		
AX	ə	EL	l	ih	ɪ	nx	n	t	t		
ax-h	ə	em	m	IH	ɪ	NX	n	T	t		
AX-H	ə	EM	m	ix	ɪ	ow	o_ʊ	tcl	sil		
axr	ɝ	en	n	IX	ɪ	OW	o_ʊ	TCL	sil		
AXR	ɝ	EN	n	iy	i	oy	ɔ_ɪ	th	θ		
ay	a_ɪ	eng	ŋ	IY	i	OY	ɔ_ɪ	TH	θ		
AY	a_ɪ	ENG	ŋ	j	ʤ	p	p	uh	ʊ		
b	b	epi	sil	J	ʤ	P	p	UH	ʊ		
B	b	EPI	sil	jh	ʤ	pau	sil	uw	u		
BCL	sil	er	ɝ	JH	ʤ	PAU	sil	UW	u		
bcl	sil	ER	ɝ	k	k	PCL	sil	ux	u		
ch	ʧ	ey	e_ɪ	K	k	pcl	sil	UX	u		
CH	ʧ	EY	e_ɪ	kcl	sil	q	sil	v	v		

Table 4: The conversion and folding mechanism that was used in our study to relabel the transcriptions for the TIMIT and Torgo datasets.

The end result of this is that we have 41 phones in our dataset instead of 39 phones. We still have all the 39-phones from Lee & Hon 1989, but we also have ɔ (“ao”) and ʒ (“zh”). We kept these around because there are enough occurrences of these forms in the Talkbank dataset and also we believe the reason for these sounds being folded in is due to the rules of English phonology, which isn’t necessarily what we are prioritizing for the purposes of early child language and dysarthric speech.

6.2 The Torgo Dataset

We were able to download the publicly available version of the the Torgo Dataset (Rudzicz et al., 2012) from the Torgo project website (http://www.cs.toronto.edu/~complingweb/data/TORGO/torgo.html). The entire dataset is transcribed in the TIMIT-style. All adjustments we made to the TIMIT dataset were also made to the Torgo dataset. There are were two microphones used in the recordings, an array-microphone and a head-mounted microphone. We ignored the recordings from the array-microphone and focused on recordings from the head-mounted which was an electret microphone which recorded audio at 16 kHz. The paper (Rudzicz et al., 2012) provides no further information on this microphone. There are a total of eight participants with dysarthria in this dataset, seven of them with dysarthria caused by cerebral palsy (4 male, 3 female) and 1 male participant who has dysarthria caused by amyotrophic lateral sclerosis (ALS). All the particpants have a wide-range of intelligibility as assessed by the Torgo team, live in the Toronto-area and are between the ages of 16 and 50 at the time of recording.

6.3 The Talkbank Dataset

Using recommendations from our thesis committee chairperson, we visited the Talkbank Project for speech data (https://talkbank.org) and immediately the material on the CHILDES database (MacWhinney 2000) was something we recognized as very relevant for our work, since it contained corpora containing the speech of children. Many of these corpora did not have phone-level transcriptions, opting instead to provide orthographic transcription of intelligible words made by children who were older than our target age range for pre-verbal children, which is age 3 and younger. Further searching led us to PhonBank (Rose &.MacWhinney 2014) which had datasets that provided phonetic transcriptions. Not every dataset we found there was useful to us. Many of them were for verbal child speech from older children that definitely were not babbles, which pushed us to look for earlier ages for the children. Some of the datasets that contained phonetically-transcribed babbles of younger children did not have wav-linked audio or span indications, which meant that we couldn't map an IPA transcription to a span of time in a wav file. One dataset had to be removed from consideration because the linkages were to mp3-files, which we know lead to about 10% to 15% greater matching error when used in speech-recognition training (Pollack & Behunek, 2011).

6.3.1 Source Corpora for our TalkBank dataset

On the PhonBank section of CHILDES in the TalkBank database, we selected four datasets that satisfied our needs, these are:

1. The Davis Corpus (Davis & MacNeilage 1995, Davis et al., 2002). It contains 702 sound files, 602 transcription files and 429 GB of data, generated from 21 children in an English language environment. Weblink: https://phonbank.talkbank.org/access/Eng-NA/Davis.html
2. The Kern French Corpus (Kern et al.. 2009, Kern & Davis 2009). It contains 133 sound files, 130 transcription files and 69 GB of data, generated from 4 children in a French language environment. Weblink: https://phonbank.talkbank.org/access/French/KernFrench.html
3. The Stuttgart Corpus (Lintfert 2009). It contains 138 sound files, 138 transcription files and 3.6 GB of data, generated from 8 children in a German language environment. Weblink: https://phonbank.talkbank.org/access/German/Stuttgart.html
4. The TAKI Corpus (Lintfert 2009). It contains 65 sound files, 65 transcription files and 1.8 GB of data, generated from 5 children in a German language environment, who were also part of the Stuttgart Corpus. Weblink: https://phonbank.talkbank.org/access/German/TAKI.html

These datasets contain timing information for phonetic transcriptions provided via the CLAN transcription software (MacWhinney 2000) but none of these transcriptions provide phone-aligned timing information. The best that is provided via the CLAN transcription system is the marking of 3-second spans of speech with a corresponding IPA transcription of what was uttered in that 3-second span of speech. The transcription files were CHA files, containing IPA transcriptions of child and caregiver utterances, prepared using the CLAN transcription software, following the CHAT transcription method described by Talkbank on their website. The sound files in each of these datasets were WAV files

containing several minutes of recorded speech from children, stored with 16-bit signed integer coding saved, with either a 48k sample rate or a 41k sample rate. A single wav file from any of these datasets can generate multiple transcribed utterances. Our definition of “utterances” were based on how individual spans of speech were divided up in the CHAT transcription files, encapsulating a 2 to 3 second batch of time where the researcher identified an utterance from a babbling child and transcribed it as a single event.

The following table shows information on how the recordings were made for each corpus we use:

Corpus	Recording Approach
Davis	Sessions were audiotaped once weekly in the normal home environment, using an Audio-Technika ATW1031 remote microphone clipped at the shoulder.
KernFrench	Sessions were audiotaped for an hour every two weeks, from 8 months of age to 25 months of age, in the infant’s homes, as parents followed normal activities with their children. No microphone information provided.
Stuttgart	Sessions were made in the children’s homes as they played with their parents. Recordings were done with a Sony DAT TCD-D100 and a high-quality wireless microphone NADY LT-4 (Lavalier) E-701 (600 Ohm). Recorded data was down-sampled to 16 kHz.
TAKI	

Table 5a: The recording approaches that were done by the researchers that created each of the Talkbank Corpora we made use of in our study.

The following information shows information on every child whose speech was a part of our dataset.

6.3.2 Analysis of our Talkbank Dataset

The next step was writing a Python script for reconciling which WAV files did not have corresponding CHA files, as well as which CHA files did not have corresponding WAV files. Such files were discarded so that we only kept CHA-WAV pairs that were verified to go together. There were 900 of these files across

all four corpora. The next step was to prepare a Python script that could open each of these 900 CHA files and parse through it, extracting key information about the child and each individual child-utterance transcription, ignoring utterances from caregivers and researchers. For each of these child utterances, we extracted the start and end time of the span of time in the WAV file that corresponds to the CHA file where the utterance was found. It is important to note that not all of the 900 CHA-WAV pairs we reconciled even had child speech. There were 180 of these files that contained caregiver utterances only and did not end up contributing anything to our dataset of child utterances. The remaining 720 WAV files yielded 145089 utterances from 29 children, each utterances having a length of 2 to 3 seconds. The transcription for each one was verified as existing. It is important to note that some transcriptions contained syllable stress IPA markers for primary and secondary syllable stress. Since we are excluding the analysis of syllable stress from our speech-recognition goals, we scrubbed these markers out from the transcriptions. We then wrote a slicing script using Python and the SoX library, that sequentially opened all these 720 WAV files and copied out each slice of time that corresponded to an utterance found in the transcription, storing them in separate WAV files. We obtained 145089 individual WAV files this way, each containing one transcribed sample that we are interested in. This is useful because the 720 WAV files were actually quite large and memory intensive, containing long amounts of silence and sounds that are not child speech that we really can do away with. This led to us only having to hold on to 57 GB of WAV file data instead of 503 GB of WAV file data. We then wrote an analysis script to analyze the 145089 transcriptions to provide various statistics and analytics, to explore how viable this material is to

our goals. Table 5a shows what we knew going into this analysis while Tables 5b, 5c, 5d and 5e show further information on this material and Table 5f shows our inferences about this dataset from analysis.

Total Number of Transcribed .wav Files	720
Total Number of Sample Spans	145089
Total Number of Genders	2 (15 female, 14 male)
Total Number of Children	29
Total CHILDES Corpora Sourced	4
Total Number of Language Environments	3 (English, French, German)

Table 5b: A summary of what we know about all the data we extracted from CHILDES, which we are calling the TalkBank dataset for the purposes of this study.

Total Duration of Utterance Spans (in milliseconds)	346694397
Average Duration per Utterance Span (in milliseconds)	2389.52916486
Total Number of Unique Transcriptions	45176
Total Number of Unique IPA phones Found	405
Total Number of Unique Diphones Found	6047
Total Number of Unique Triphones Found	34283
Total Number of Unique Child Ages	482
Total Number of Unique Recording Dates	534

Table 5c: Further details discovered about our TalkBank dataset using scripting tools to analyze.

Name	**Age Range (years;months.days)**	**Corpus**	**Sex**
Baptiste	0;07.19 – 2;00.12	KernFrench	M
Ben	0;11.21 – 2;04.02	Davis	M
BW	1;02.20 – 2;11.14	Stuttgart	M
Cameron	0;07.11 – 2;11.24	Davis	F
Charlotte	0;10.12 – 2;11.22	Davis	F

ED	2;03.10 – 4;03.	Stuttgart, TAKI	F
EL	0;05.04 - 2;02.08	Stuttgart, TAKI	F
Emma	0;08.16 – 2;01.08	KernFrench	F
Esteban	0;07.16 – 2;00.25	KernFrench	M
FZ	1;00.21 – 2;06.13	Stuttgart	F
Georgia	0;08.25 – 2;11.05	Davis	F
Hannah	0;11.14 – 2;04.24	Davis	F
HH	0;05.23 – 3;00.19	Stuttgart	F
Jules	0;09.14 – 2;00.29	KernFrench	M
Kaeley	1;00.24 – 2;01.23	Davis	F
LL	3;08.03 – 7;07.06	Stuttgart, TAKI	F
Martin	1;05.19 – 2;02.09	Davis	M
Micah	0;08.01 – 1;06.19	Davis	M
Nate	0;10.07 – 2;09.07	Davis	M
NB	0;07.17 – 2;08.12	Stuttgart	M
Nick	0;10.20 – 3;01.02	Davis	M
Paxton	0;08.02 – 2;00.02	Davis	M
Rachel	0;08.04 – 1;10.02	Davis	F
Rebecca	1;01.24 – 1;07.17	Davis	F
RL	3;01. - 4;06.13	Stuttgart, TAKI	M
Rowan	0;10.23 – 2;10.19	Davis	M
Sadie	0;07.11 – 1;07.03	Davis	F
Sam	0;10.07 – 2;01.03	Davis	M
Willie	0;07.29 – 1;04.13	Davis	M

Table 5d: Age and gender information on every participant whose speech data we use in our Talkbank dataset.

Language Environment	Corpus	Milliseconds of speech	Number of utterances
English	Davis	220,171,487 ms	75544
French	Kern French	115,708,342 ms	49364
German	Stuttgart	6,150,759 ms	10232
	TAKI	4,663,809 ms	9949

Table 5e: Amounts of speech contributed to the Talkbank dataset by each source corpus.

Inference	Facts that Support the Inference
The timing-information of the phonetic transcriptions are not phone-aligned, for any corpus.	Each utterance has only a single time-stamp marking the beginning of the transcriber-defined utterance event.
Utterances are 2 or 3 seconds long, on average.	Average duration for utterances is around 2389.5 ms
Many utterances share the same phonetic transcription.	45176 unique transcriptions were found among 145089 utterances.
Many of the transcriptions contain IPA marks that involve combinations and modifications (such as nasalization, lengthening, devoicing, and so on).	405 unique IPA marks were found (excluding the 2 syllable stress marks we scrubbed out earlier) across 3 languages that share several phones in common. Specific example shown in Table 10 below.
Some phones tend to have an affinity for each other, while avoiding other phones. The children already prefer to not connect some phones with others.	With 405 unique IPA marks in the dataset, we can raise this number to the power of 2, giving us 164025 possibilities for diphones, but there were only 6,047 diphones found in the dataset. 405 raised to the power of 3 leads to 66,430,125 possibilities for triphones, but there were only 34,283 triphones found in the dataset,
Most of the children are from an English language environment, but the non-English environments are still significantly represented.	The following data was brought into our dataset: Utterances from 17 out of 21 English-environment children in the Davis corpus. Utterances from 4 out of 4 French-environment children in the Kern French corpus. Utterances from 8 out of 8 German-environment

	children in the Stuttgart corpus. Utterances from 4 out of 5 German-environment children in the TAKI corpus. All four of them were also in the Stuttgart corpus. Also see Table 5d.
The children were assigned binary genders and there is an even-split between them.	15 of the 29 children across the corpora were described as female while 14 of them were described as male. None of the children were described as non-binary gender or transgender. Out of the 145089 samples identified, 74114 of them are from female voices while 70975 of them are from male voices. That's a 51:49 female-male ratio by number of samples. When we consider total duration of samples, there are 181,816,466 milliseconds of female voice utterances and 164878931 milliseconds of male voice utterances, giving a 52:48 female-male ratio by total sample duration.
We are able to have samples of older and younger vocal tracts for the same child in most cases, which increases the variation on vocal tract sources even further.	There are 482 unique “ages” in the dataset, tracked by exact days alive, across 29 children.
Most of the recordings come from pre-kindergarten children	Most of the dataset contains speech from children between 1 and 3 years of age.
We have 7 years and 2 months of spread across the ages of the children that were recorded.	The earliest age of a recorded child in the dataset is 5 months and 4 days, the oldest age of a recorded child in the dataset is 7 years and 7 months.
The recordings come from a 33 year span of time.	The dates of the recordings vary for each child and each corpus, based on when various studies were taking place. The earliest recording is from Nov 1984 and the latest recording is from May 2017, with a majority falling in the early 2000s decade. Hopefully these provide variability on the audio quality of the

	recordings for our training data, since different recording technologies will have been in place during these various decades.
Some children provide more utterances than others, by number.	The per-child sample frequency ranges from a high of 13225 to a low of 801. Average is 5,003. Median is 2,597.
Some children provide more utterances than others, by duration.	The highest is 52,410,356 milliseconds from one child (14.5 hours). The lowest is 896,249 milliseconds from another child (15 minutes).

Table 5f: Inferences made about our Talkbank dataset from our analysis

6.3.3 Phone-level statistics on the Talkbank dataset

We ran further analysis scripts to figure out what the most common phones in the dataset were. Table 6 shows the 25 most common phones in the dataset among all 29 the children across all four corpora, sorted in descending order of frequency of occurrence. No phone-folding was applied yet. The transcription did not provide tie-bars to indicate the existence of diphthongs, so we are assuming that diphthongs are in the transcriptions. This is merely the 25 most common symbols used in transcriptions out of the 405 that were employed.

We can see that the most common sounds that these young children are producing are vowels and frontal consonants, which are transcribed with no modifiers or combining symbols, except for the "long vowel" symbol. This information is agrees with findings from the research literature on childhood language acquisition (Paul & Norbury, 2012).

Rank	Phone	Frequency
1	a	42723
2	d	30347
3	b	24583
4	ɪ	22986
5	m	22494
6	t	21027
7	n	20105
8	ʌ	17937
9	ɛ	14869
10	o	14710
11	i	14573
12	ə	14284
13	s	12396

Rank	Phone	Frequency
14	aː	10907
15	p	10782
16	w	10464
17	ʊ	10343
18	h	10307
19	k	10277
20	u	10183
21	æ	9790
22	l	9414
23	e	9359
24	j	9064
25	g	9035

Table 6: The 25 most popular IPA symbols (and monophones) in our Talkbank dataset.

Table 7 below shows the 25 most common diphones in the dataset. This information was obtained by considering bigram pairs of phones that appear in sequence at any point in the transcriptions and counting every instance of them individually.

Rank	Diphone	Frequency
1	aɪ	6541
2	ba	4577
3	ma	4529
4	da	4344
5	dɪ	3176
6	aʊ	3171
7	am	2821
8	dæ	2642
9	bʌ	2480
10	ɪs	2334
11	dɛ	2261
12	no	2243
13	pa	2220

Rank	Diphone	Frequency
14	ɪn	2186
15	ta	2118
16	at	2067
17	ɪt	2057
18	la	2053
19	jɛ	1944
20	ab	1909
21	ɪd	1893
22	di	1859
23	ap	1814
24	ha	1769
25	maː	1681

Table 7: The 25 most common diphones in our Talkbank dataset.

Table 8 below shows the 25 most common triphones in the dataset. This information was obtained by considering trigram pairs of phones that appear in sequence at any point in the transcriptions and counting every instance of them individually.

Rank	Triphone	Frequency
1	mam	2169
2	aɪn	1239
3	dæd	1176
4	bʌb	1003
5	bab	978
6	dɪd	920
7	baɪ	860
8	pap	779
9	ama	682
10	aba	668
11	tat	667
12	haɪ	649
13	dɪs	641

Rank	Triphone	Frequency
14	ædæ	631
15	beb	596
16	dad	586
17	ʌbʌ	570
18	aɪs	554
19	mʌm	536
20	daʊ	528
21	ɪdɪ	526
22	maɪ	499
23	maːm	473
24	dɛd	462
25	ami	437

Table 8: The 25 most common triphones in our Talkbank dataset.

Table 9 below shows the most common transcriptions in the dataset, giving us an idea of what kinds of data our automatic phone-recognizer will be asked to transcribe. From Tables 7, 8 and 9, we see that the most common triphones also end up being the ingredients for the most well-known pre-verbal child babbles across most languages, such as <dada>, <mama>, <baba> and other variants of parent-referents that various languages have landed upon, most likely inspired by child babbles in the first place. Due to a large percentage of the data coming from an English-environment, we can understand the higher prevalence of triphones that are used to produce English greetings, such as [haɪ] and [baɪ] and typical pronouncements that young children are known to make as early words (Paul & Norbury, 2012),

such as [maɪ], [dɪd] and [beb]. We also see from the most common transcriptions that many times these pre-verbal children are only uttering single syllables or just single vowels. All of these observations agree well or are in accord with prevailing theories on child language-acquisition (Paul & Norbury, 2012).

Rank	Transcription	Freq.
1	mː	2190
2	m	2109
3	no	1661
4	œ	1404
5	aː	1341
6	a	1299
7	œː	1288
8	ʌ	1184
9	o	924
10	ɪt	815
11	da	680
12	jɛ	664
13	æ	629
14	ɛ	576
15	ba	571
16	nɔː	568
17	dɪs	541
18	ɪç	509

Rank	Transcription	Freq.
19	ɔː	493
20	bʌ	481
21	ɛ̃	481
22	uː	477
23	la	474
24	u	467
25	ɪ	465
26	oː	451
27	dʌ	417
28	ʊ	384
29	jæ	380
30	ɛ̃ː	379
31	e	366
32	mami	365
33	ɔ	363
34	haɪ	362
35	daː	361
36	jaː	360

Rank	Transcription	Freq.
37	i	356
38	nɔ	352
39	dæ	348
40	bʊ	335
41	dɪ	333
42	wi	323
43	aɪn	322
44	deːɐ	321
45	dɛs	320
46	eː	315
47	ma	311
48	do	301
49	mama	295
50	aɪ	294
51	go	293
52	təkə	292
53	diː	285
54	mamaː	281
55	wʌn	277

Table 9: The 55 most common transcriptions in our Talkbank dataset.

6.3.4 Processing transcriptions into HTK Label files

The next step was to define HTK .lab label files for each utterance .wav file that we had produced. We wrote a Python script that utilized the per-utterance IPA transcription information that we had already extracted for each utterance, to create text files encoded in UTF-8 named after each .wav file we had, but giving them a .lab extension. The contents for each file were the IPA transcriptions that we had on file for the corresponding .wav file, organized in the one-phone-per-line style that is expected from HTK. The "sil" symbol was padded at the beginning and end of each utterance. Extra care was taken to preserve combination symbols and modifier symbols so as not to clobber the original IPA transcriptions.

6.3.5 Managing the Talkbank dataset phone count

The 405 phones detected across all four corpora include combination symbols and modification symbols from the IPA symbol set. Many of these modifiers are subtle and probably can be folded into other phone labels. We noticed that the Davis corpus featured a very narrow phonetic transcription style, while the other three corpora have a broad transcription style, causing a large portion of the phone symbols in our dataset to come from the Davis corpus. Table 10 below shows an example of how the unrounded near-low front vowel [æ] presents itself in the Davis corpus in 11 different ways. Situations like these give us good clues for phone folding procedures to reduce the complexity of the phone recognition task. Other clues come from the kinds of phones that were selected in the transcriptions for the Kern French, Stuttgart and TAKI corpora and at least bringing the scope of the Davis corpus transcriptions to the level of the other three corpora. We know we must not try using allophone information to fold phones into

each other because we are working with pre-verbal child speech from three different cultures, not all of whom will have aligned to phonemic expectations for these languages.

Phone	**Unicode Description of IPA Symbols**	**Phonetic description of IPA symbols**
[æ]	LATIN SMALL LETTER AE	Unrounded near-low front vowel
[æ^{j}]	LATIN SMALL LETTER AE + MODIFIER LETTER SMALL J	Palatalized unrounded near-low front vowel
[æ^{w}]	LATIN SMALL LETTER AE + MODIFIER LETTER SMALL W	Labialized unrounded near-low front vowel
[æː]	LATIN SMALL LETTER AE + MODIFIER LETTER TRIANGULAR COLON	Long unrounded near-low front vowel
[æː^{j}]	LATIN SMALL LETTER AE + MODIFIER LETTER TRIANGULAR COLON + MODIFIER LETTER SMALL J	Palatalized long unrounded near-low front vowel
[æ̃]	'LATIN SMALL LETTER AE + COMBINING TILDE'	Nasalized unrounded near-low front vowel
[æ̃ː]	LATIN SMALL LETTER AE + COMBINING TILDE + MODIFIER LETTER TRIANGULAR COLON	Nasalized long unrounded near-low front vowel
[æ̥]	LATIN SMALL LETTER AE + COMBINING RING BELOW	Voiceless unrounded near-low front vowel
[æ̥̃]	LATIN SMALL LETTER AE + COMBINING TILDE + COMBINING RING BELOW	Voiceless nasalized unrounded near-low front vowel
[æ̰]	LATIN SMALL LETTER AE + COMBINING TILDE BELOW	Laryngealized unrounded near-low front vowel
[æ̰ː]	LATIN SMALL LETTER AE + COMBINING TILDE BELOW + MODIFIER LETTER TRIANGULAR COLON	Long laryngealized unrounded near-low front vowel

Table 10: The 11 different ways that the unrounded near-low front vowel, [æ] presents itself in the Davis corpus, with each one counting as a unique phone.

Further analysis of the frequency of each phone in the dataset yielded a useful observation. Out of 405 phones, 296 of them occur less than 50 times and many of these are the ones with subtle modifiers and combination symbols (an example of which we see in Table 10). These are prime candidates for folding into the base-form. We then decided to take the bold step of folding all modified and combination forms of a phone into its base form. This would mean that all the phones in Table 10 would be considered under a single exemplar form of [æ]. We repeated this step for all modified and combination forms until our phone-count reduced from 405 to 83 base forms. We decided to call this the “integrated” phone set to imply that all modified and combination phones had been integrated into their base forms. Table 11 shows us the IPA symbols of these 83 “integrated” phones:

a	m	y	ɑ	ɪ	ɻ	ʒ
b	n	z	ɔ	ɬ	ɽ	ʔ
c	o	æ	ɖ	ɬ	ɾ	⊙
d	p	ç	ə	ɭ	ʀ	ʙ
e	q	ð	ɚ	ɮ	ʁ	ʤ
f	r	ø	ɛ	ɯ	ʃ	ʦ
g	s	ħ	ɟ	ɱ	ʉ	ʧ
h	t	ŋ	ɣ	ɲ	ʊ	β
i	u	œ	ɤ	ɳ	ʋ	θ
j	v	‖	ɥ	ɵ	ʌ	χ
k	w	ǂ	ɦ	ɸ	ʍ	ʏ
l	x	ɐ	ɨ	ɹ	ʏ	

Table 11: The “integrated” phone set for our Talkbank dataset containing 83 base phones.

We decided to look at the frequency of occurrence for each integrated phone to see if any of them occurred less than 50 times, with 50 being a number we decided as an acceptable threshold for having enough examples of a phone for statistical machine learning to be reliably performed, even if some of them were taken away to put in a testing set. It turns out that ten phones ended up being hapax legomena in the first place, and removing them from consideration ended up being prudent. Another 16 phones fell under our 50 count threshold in their frequency of incidence. And two phones that occurred more than 50 times were discovered to be highly skewed towards male samples with much fewer female samples, which wouldn't be a representative sample set to provide for training in the first place. Table 12 below summarizes which phones were rejected for which reasons.

Rejected integrated phones	**Reason for rejection**
ʋ , ǁ , ɖ , ɟ , ɤ , ɭ , ɱ , ɦ ,ɯ , ɻ	Hapax legomena: only one incidence of each of these phones in the entire dataset
ʊ , ɵ , ɽ , ʘ , ħ , ǂ , ɦ , ɮ , ɯ , ɻ , ɋ , ɨ , ɬ , ɬ , ɲ , ɳ , ɾ , ʍ , ʦ	Too rare: Fewer than 50 incidences of each of these phones in the entire dataset.
χ	Not gender representative: 69 male samples but only 1 female sample
ɥ	Not gender representative: 111 male samples but only 6 female samples

Table 12: The 28 rejected phones from the integrated phone set, with rejection reasons.

Removing these 28 phones from the 83 phones in the integrated phone set leaves us with 55 phones remaining. We knew that we needed a symbol to represent silence and we also needed a symbol to represent all of these removed phones, since we didn't want to mislabel the actual concept of silence in

our dataset by applying it to these rejected phones, which are not acoustically examples of silence. So we added two symbols: “sil” for silence and “other” for any of these rejected phones. The result gave us 57 phones in total, which were defined by us as the “filtered” phone set. Table 13 shows us which phones these are.

sil	h	p	y	ɐ	ɸ	ʏ
a	i	r	z	ɑ	ɹ	ʒ
b	j	s	æ	ɔ	ʁ	ʔ
c	k	t	ç	ə	ʁ̞	ʙ
d	l	u	ð	ɚ	ʃ	ʤ
e	m	v	ø	ɛ	ʉ̞	ʧ
f	n	w	ŋ	ɣ	ʊ	β
g	o	x	œ	ɪ	ʌ	θ
						other

Table 13: The filtered phone set for our Talkbank dataset, containing 57 symbols.

6.4 Converting from WAV to MFCC format

Every .wav file we worked with from the Timit, Torgo and Talkbank datasets were subjected to mel-frequency cepstral coefficient (MFCC) extraction procedures using the HCopy tool in HTK. The configuration information that was used to perform this work is provided in Appendix A of this thesis. The result of this operation was a single .mfcc file being created for each .wav file for every

configuration we cared about, with each of these .mfcc files containing MFCC coefficients in the desired configuration specified. It would be with these .mfcc files that we would create HTK script-files that would define our various training and testing datasets for our experiments. We go into more detail on this later. During this process, it was discovered that 3,361 of our Talkbank wav files were too short to be converted to any MFCC format meaningfully due to too few samples existing. Checking their timespans in the .cha files they came from shows that some mistake must have occurred in defining their timespan when the original transcribers were defining them with CLAN. Some of them even had timespans that were too short for the Hamming window to create a meaningful first vector frame. They were identified and dropped from the dataset, and noted down as such for future reference. This caused us to exclude 2.37% of our original Talkbank dataset.

6.5 Splitting: Training sets vs. Testing sets

For the TIMIT dataset, we used the default split between training and testing for our work. For the Torgo dataset, we did not use this data to train any models due to the small number of speakers available. All of this dataset was used for testing. For the Talkbank dataset, we followed a very simple procedure to split the dataset into a training set and a testing set. To get a 75%-25% split, we used a script to randomly assign 113,000 of our Talkbank utterances to be considered our training set while the remaining 28727 utterances were considered our testing set. A few regex searches were done to verify that all 57 identification targets were well-represented in the test set and this was confirmed. The exact dataset manifest information can be made available to readers upon request.

6.6 Preparing the TalkBank data to be recognized by a TIMIT-trained model

As mentioned earlier, some phones in the TIMIT phone set were folded into other phones, removing their individual existence from the revised TIMIT phone set. Table 4 shows how we folded our own TIMIT phones, with inspiration from Lee & Hon 1989. However, some of these TIMIT phones that were assimilated still exist in the TalkBank dataset. Moreover, there are also other phones in the TalkBank dataset that are not among the original 61-phone TIMIT phoneset. It became necessary to explicitly handle these situations in our Talkbank dataset, in a process we called “timiting” the dataset. We did this by performing the following explicit phone replacements, as shown in Table 14. This way, when we run a TIMIT-trained phone-recognizer on the TalkBank test data (or even on the entire TalkBank dataset), we are able to not have unknown phones showing up and are also able to make fair-enough comparisons between the generated transcription and the human-created transcription. The selections made in Table 14 either reflect what was done in our TIMIT phone set folding criteria or reflect our search for a closest TIMIT phone-set substitute that shares phonetic features with the phone that is being replaced.

Before	After
c	ʧ
r	ɹ
x	k
y	i
ɐ	æ
ɚ	ɝ
ɣ	g
ɸ	f
ʁ	g
ʁ	g

Before	After
ʉ	u
ʌ	ə
ʏ	ɪ
ʔ	sil
ʙ	b
β	b
ø	e
ç	k
œ	ɛ
other	sil

Table 14: The replacement criteria for “timiting” phones that existed in the TalkBank dataset but not in the TIMIT dataset, so that fair-enough comparisons can be made with the result of a TIMIT-trained phone-recognizer working on TalkBank data.

6.6.1 Handling TIMIT diphthongs

The only phones in our folded IPA-converted TIMIT phoneset that don’t exist in our “timited”, integrated, filtered TalkBank phoneset are the five TIMIT diphthongs, shown below in Table 15. We left the ability to detect these diphthongs intact in all our TIMIT-trained models. However, we added a functionality to “dedipthongify” any output transcriptions coming from these models in situations where comparisons with human-transcribed TalkBank transcriptions were happening, since dipthong transcription was not consistently provided by the human transcriptions in the TalkBank corpora we are

using, so we removed any diphthong notations we came across. Table 15 explains how we merely split every TIMIT diphthong to become transcribed as its base phone sounds in sequence.

Before	After
a_ʊ	a ʊ
a_ɪ	a ɪ
e_ɪ	e ɪ
o_ʊ	o ʊ
ɔ_ɪ	ɔ ɪ

Table 15: Criteria for reconciling diphthongs between datasets.

6.7 Vocal Tract Length Normalization

Any TIMIT-trained phone-recognizer would be biased towards the adult voices which the TIMIT corpus is based upon. Because there are 630 of them with equal amounts of contributing utterances, no single adult voice ends up gaining a bias, but the concept of adulthood in the voice does gain bias, particularly when it comes to the assumption of how long a speaker's vocal tracts is. If we are only presenting adult voices in the test set for a TIMIT-trained phone-recognizer, this doesn't present itself as a problem. This is indeed what we do when we have TIMIT and TORGO data as our test set. However, whenever TalkBank data ends up being our test set for a TIMIT-trained phone-recognizer, we need to be mindful that vocal-tract length normalization will need to be conducted, whereby the test data is processed to warp the frequency of the speech to mimic the output from a longer vocal tract. HTK provides a

mechanism for accomplishing this using its configuration file settings (Young et al., 2006), which we have demonstrated the use of in Appendix A below. The concept of “VTLN Warp Factor” becomes a value that gains importance towards influencing a more accurate transcription for the child utterances we are working with. Previous work in this context has been done by Azizi et al., (Azizi et al., 2012) where they explored ideal HTK VTLN Warp Factor values for Persian-speaking children around the age of 8. They discovered that an ideal value for 8 year old children is 0.83, since the recordings processed with that warp factor had the lowest error rates in phone recognition. We go into extensive detail on our experimentation of this in a future section of this thesis.

6.8 Phone-recognition grammar and pronunciation dictionary

Because HTK is a full-service speech-recognition package designed to recognize words in sentences, we still need to define a “grammar” and a “pronunciation dictionary” for our project. In HTK, a pronunciation dictionary is defined in a textfile containing one line for every word that is expected to be recognized. Following each word is a single space, after which the pronunciation sequence for the word is provided, using the phones from the phoneset being used for the phone recognition task. Multiple entries are allowed for each word to take care of alternate pronunciations. Since this is a phone recognition project and not a word-recognition project, our pronunciation dictionary is trivial. It would just be each phone in our phoneset taking the role of “word” and then the pronunciation sequence is just the phone itself. The end result just becomes a textfile containing each phone placed twice on its own

line, delimited by a single space. Two such dictionaries will be required, one for our folded TIMIT phone set and one for our integrated, filtered Talkbank phone set. Both of these are provided in Appendix E. Our experiment descriptions imply that our Talkbank pronunciation dictionary will be used in Pipeline A and our TIMIT pronunciation dictionary will be used in Pipeline B.

To define an HTK grammar, one needs write a text-file containing the rules of the grammar for the task we are working in, defining where words will be found in relation to other words, in a special Backus-Naur Form (BNF) notation that defines a context-free grammar to use for sentence formation. This text-file is then processed into a machine-readable lattice using the HParse tool.

Because we are doing phone recognition, the BNF notation of our grammar will merely specify that every utterance starts with silence, followed by any number of phones from our phone set (including silence), and then ends with silence. With pronunciation dictionaries, we require two grammars for our experiments, one to define sequencing with our folded TIMIT phoneset and one to define sequencing with our integrated, filtered Talkbank phoneset. These grammar descriptions are shown in Appendix F.

7. Experimental variables

We now have everything we need to conduct our experiments. We can see from the descriptions of Pipelines A and B that a few variables need to have their values selected before we conduct any experiment. There are some HTK settings that we left constant throughout our experiments and these can be investigated in Appendix B, which shows examples of the HTK commands we used to run our experiments.

7.1 Choice of training data

We have identified two main sources of training data, the TalkBank dataset and the TIMIT dataset. We have described above how each of these has had its own training subset identified. So the only options we have for this variable are:

- TimitTrain: The training subset of the TIMIT dataset.
- TalkbankTrain: The training subset of our TalkBank dataset
- TalkbankAll: For situations where we utilize our entire Talkbank dataset for training.

7.2 Choice of testing data

We have identified three main sources of testing data: The TIMIT dataset, our TalkBank dataset and the TORGO dataset.

For the TORGO dataset, we would assume that the entire dataset can be used for testing purposes. Within this dataset, there are two kinds of people producing the speech utterances: patients and controls. We have ignored the controls from our TORGO testing dataset, since we are more interested in the speech of the dysarthria patients. This is also prudent because the publicly available TORGO dataset only contains control speech from male control subjects.

7.3 Age At Recording

For the TalkBank dataset, we are able to subdivide the test subset further by age and by language environment. This becomes relevant when it comes to exploring and setting VTLN Warp Factor values with respect to age. It also becomes relevant when thinking about how the influence of a common language-environment between the training-set and the testing-set can lead to different accuracy outcomes for our phone-recognizer. Knowing that we had children across several years of age, we decided to set age as a sub-variable that selects for specific testing subsets. Our age variable presents as an age-span called “Age At Recording”, defining six-month buckets of age from 0 years to 8 years. We therefore have 17 values for this variable, shown in this array:

[0.0 – 0.5, 0.5 – 1.0, 1.0 – 1.5, 1.5 – 2.0, 2.0 – 2.5, 2.5 – 3.0,

3.0 – 3.5, 3.5 – 4.0, 4.0 – 4.5, 4.5 – 5.0, 5.0 – 5.5, 5.5 – 6.0,

6.0 – 6.5, 6.5 – 7.0, 7.0 – 7.5, 7.5 – 8.0, 8.0 – 8.5]

This value gives us information on the age of the child in years at the time of the recording but collects age-similar clusters together by spans of six months. The way we have set this variable up is that the

lower bound is “inclusive” while the upper bound is “exclusive”. An utterance recorded from a child who was 7 months old would therefore have an “Age At Recording” value of “0.5 – 1.0”. An utterance recorded from a child who was 3 years and 9 months old would therefore have an “Age At Recording” value of “3.5 – 4.0”. A child who is exactly 3 years old would have an “Age At Recording” value of “3.0 – 3.5”. A child who is exactly 3 years and 6 months old would have an “Age At Recording” value of “3.5 – 4.0”. Across the dataset, we see examples of every child being recorded at different ages in life. Some values for “Age At Recording” have no recordings while others have a cluster of recordings. Further information on this variable is provided in the VTLN experiment descriptions to follow.

7.4 Language Environment

Regarding the language-based splitting of the TalkBank testing subset, we have only three values for this subvariable of “language environment”:

[“eng”, “fra”, “deu”]

Where “eng” represents an English language environment, “fra” represents a French language environment and “deu” represents a German language environment.

For the TIMIT dataset, we decided not to subdivide its test set whenever we used TIMIT test data.

7.5 Cepstrum

This variable focuses on the exact arrangement of mel-frequency cepstral coefficients (MFCCs) that will be used in the training-data, testing-data and HMM configurations. We decided to use the HTK codes for various MFCC configurations as our variable values, for a variable that we are calling “Cepstrum”. The different values that this variable takes in our experiment are shown in Table 16 below, with explanations. We based our MFCC configuration files on the specification shown in Appendix A, where there are 12 base MFCC co-efficients, which is the default in HTK.

Cepstrum Code	**Meaning in HTK**	**Number of coefficients**
MFCC	The 12 base MFCC coefficients.	12
MFCC_D	The 12 base MFCC coefficients, plus their first derivatives (deltas)	24
MFCC_D_A	The 12 base MFCC coefficients, plus their first and second derivatives (deltas and delta-deltas)	36
MFCC_D_A_T	The 12 base MFCC coefficients, plus their first, second and third derivatives (deltas, delta-deltas and delta-delta-deltas)	48
MFCC_E_D_A	The 12 base MFCC coefficients, the log energy coefficient, plus their first and second derivatives	39
MFCC_E_D_A_T	The 12 base MFCC coefficients, the log energy coefficient, plus their first, second and third derivatives	52

Table 16: The different MFCC configurations used in our experiments.

7.6 HMM States

The number of states in an HMM give us some control over how distinct states within a phone can be modelled, particularly when diphthongs and coarticulations are occurring. In the research literature we

have seen (Lopes & Perdigão, 2011), 5 states, with the middle three of them emitting, is the norm. We followed this approach but also experimented with HMMs that have 7 states, with the middle five of them emitting. Therefore, we define an “HMM States” experimental variable and give it two values:

[5, 7]

7.7 Gaussian Configuration

When creating HMMs in HTK, one can define the number of Gaussian functions involved to model every state in an HMM, while also managing the weight distributions between these Gaussians. We decided to make sure that all states share the same Gaussian function arrangements. This allows us to specify a specific Gaussian arrangement for an HMM that applies to all the states in the HMM. When it came to multiple Gaussian HMMs, we experimented with equally weighted and non-equally weighted Gaussians, to see if that would influence our results. When defining our “Gaussian configuration” variable, we present the number of Gaussian functions used as the identifier of every value, plus an optional specifier in parentheses describing non-equal weighting arrangements, using weight numbers that add up to 1. If this suffix is not present, it implies that the Gaussians were all equally weighted. In the course of our experiments, we end up having the following values for this variable, shown in the array below:

[1, 2, 2 (0.75, 0.25), 3 (0.5, 0.25x2), 4, 8, 16, 32]

7.8 VTLN Warp Factor

Earlier we have discussed the importance of using Vocal Tract Length Normalization in adjusting child speech data for phone recognition by an adult-biased TIMIT-trained recognizer. This involves re-extracting MFCC features from the child speech .wav files, with a VTLN Warp Factor set to something less than 1.0, according to The HTK Book (Young et al., 2006). Azizi et al., 2012, demonstrated that for 8-year olds, a VTLN Warp factor of about 0.83 gives the lowest error rate. Knowing that the children in our Talkbank dataset are all under 8 years of age, we can make a strong guess that their vocal tracts will be even shorter than that of a typical 8 year old. This means we should explore ideal values below 0.83, which will probably need to be further adjusted across years of age. We also know that too low a value will cause over-warping, causing a negative impact to the performance of the phone-recognizer. We performed further experimentation to figure out ideal values for each age-group, and this is described later in this thesis. The array below shows the different VTLN Warp Factor values we used in our experiments. Note that a factor of 1.0 means that no VTLN procedure was performed on the data:

[0.55, 0.58, 0.63, 0.65, 0.68, 0.72, 0.78, 0.83, 1.0]

8. Experiment Planning

Combining our methodology described in Section 5 with our experimental variables described in Section 7, we now have a clear plan for how we will perform our experiments in each phase.

8.1 Phase 1: Train on Talkbank, Test on Talkbank

Create phone-recognizers via Pipeline A using the Talkbank training data, building out HMMs using all combinations of our defined experimental values for the Cepstrum, Gaussian Configuration and HMM Configuration experimental variables. Test these against the Talkbank test set. This approach allows us to compare the results for all these variables to pick out the best possible combination. We would compare this with what we find in Phase 2.

8.2 Phase 2: Train on TIMIT, Test on TIMIT

Create phone-recognizers via Pipeline B using the TIMIT training data, building out HMMs using all combinations of our defined experimental values for the Cepstrum, Gaussian Configuration and HMM Configuration experimental variables. Test these against the TIMIT test set. Compare the best-performing recognizer here in Phase 2 with the best-performing one we made in Phase 1. This might involve making a few extra recognizers that follow the Phase 1 approach. We expect the TIMIT-trained recognizer to perform better on its own data, since the TIMIT dataset has richer transcription timing information.

8.3 Phase 3: Train on TIMIT, Test on Talkbank

Use the best TIMIT-trained recognizer created in Phase 2 to test across the Talkbank dataset. We do this to see how far we can get with adult-created data and to explore how performance changes across various subsets of the Talkbank data. We would want to see if English-environment children's speech is better recognized, since the TIMIT corpus is of English speech. We would also want to see if results from testing on the Talkbank test sets match the results from testing on the entire Talkbank dataset (training and test material). This would tell us if the Talkbank test sets we have extracted are indeed representative of the entire dataset. We would want to repeat this across languages to see if the language-specific test sets are also good representatives of the entire languages-specific material in our dataset.

8.4 Phase 4: Train on TIMIT, Test on Torgo

Use the best TIMIT-trained recognizer created in Phase 2 to test on the Torgo dataset. We do this to see how similar the performance of this recognizer is with respect to what happened in Phase 3. If we get similar performance numbers, it tells us that the speech of young children shares characteristics with the speech of dysarthria patients, when it comes to recognizing the phones that are produced by both groups, while ignoring the entire concept of language-modeling and task-achievement. We work with what was produced and not with what was intended to be produced. Due to the different lengths in vocal tract for both groups, we expect results to be quite dissimilar at this phase, which necessitates Phase 5.

8.5 Phase 5: Train on Timit, Test on Talkbank data for optimal VTLN

For every value in our VTLN Warp Factor variable, regenerate the MFCC data for all the Talkbank sound files. Then use the best phone-recognizer we made in Phase 2 to repeat our Phase 3 experiments, but this time, repeat the experiments for every combination of VTLN Warp Factor, Language Environment and Age At Recording. These multiples lead to a very large number of experiments that need to be done, which can be handled by judicious use of automation scripts. The idea here is to figure out what optimal value of VTLN Warp Factor works for each age group and to see how any of these concepts change across Language Environment values.

9. Results

We performed several experiments with the values of our experimental variables done in different combinations, following either Pipeline A or Pipeline B for our work. Further details on the commands used to perform the experiments are shown in Appendix C.

When it comes to defining the performance of the phone-recognizers we created, we define the following concepts for this thesis, in line with the values produced by the HTK HResults tool (Young et al., 2006):

Correctness: This is the percentage of phones in the test dataset that were correctly recognized, which is in line with the "%corr" definition used by HTK.

Hits (H): This is the number of phones that were correctly recognized.

Insertions (I): This is the number of phones in a generated transcription for an utterance that were assessed as erroneously inserted, when compared to the human transcription.

Deletions (D): This is the number of phones in a generated transcription for an utterance that were assessed as erroneously deleted, when compared to the human transcription.

Substitutions (S): This is the number of phones in a generated transcription for an utterance that were assessed as erroneously substituted for another phone, when compared to the human transcription.

Number of phones (N): This is the total number of phones in the human-transcribed test set that correspond to material for which a transcription was attempted. This number can change across experiments, as certain utterances fail to be processed due to runtime errors that emerge due to bad samples and bad conversions to MFCC features.

Accuracy: This is formally defined by HTK to be:

Accuracy = ((Number of phones – insertions – deletions – substitutions) / Number of phones) × 100%

9.1 Phase 1: Training on Talkbank, Testing on Talkbank

For Phase 1, we focused on trying to get the best Talkbank-trained HMM that we could make, using our Talkbank testing dataset to assess performance and our Talkbank training dataset to learn phone models. We used Pipeline A to perform the experiments with various arrangements of Cepstrum and Gaussian Configuration. Every experiment ensures that the training and testing datasets had their MFCC features extracted in accordance with the desired Cepstrum configuration. Each experiment creates a different model, based on the configurations specified. It is important to note that we used a grammar that automatically inserts a “sil” symbol at the beginning and end of an utterance on its own. Since we set up the entire dataset to have utterances padded with the “sil” symbol in this way, it is important to remove these symbols before we run HResults to tally the results. The “sil” padding would roughly be 30% of the entire transcription set in the first place, and getting them automatically correct by default would inflate our correctness and accuracy values, leading to an outcome that doesn’t really reflect our goal for phone recognition via a machine learning process. So it was important to remove the “sil” padding from both the generated transcriptions and the human-transcriptions to get a fair calculation for correctness and accuracy. Table 17 below gives us the results for our work on Phase 1, where all experiments shown here involved the use of five HMM states. It is important to note that both the training and the testing datasets underwent corresponding MFCC extraction configurations.

Experiment	Cepstrum	Gaussian and HMM Configuration	Correctness	Accuracy	H	N	I	D	S
1	MFCC	2 [5 states]	10.47	-82.58	10233	97702	90920	15753	71716
2	MFCC	2 (0.75,0.25) [5 states]	10.46	-82.70	10224	97702	91023	15704	71774
3	MFCC	4 [5 states]	10.48	-82.59	10238	97702	90934	15750	71714
4	MFCC	1 [5 states]	10.47	-82.58	10231	97702	90909	15753	71718
5	MFCC	3 (0.5,0.25 x2) [5 states]	10.48	-82.59	10237	97702	90928	15752	71713
6	MFCC_D	2 [5 states]	14.11	-149.38	13789	97702	159735	9965	73948
7	MFCC_D	2 (0.75,0.25) [5 states]	14.02	-149.06	13700	97702	159337	10043	73959
8	MFCC_D	4 [5 states]	14.12	-149.31	13794	97702	159675	9968	73940
9	MFCC_D	1 [5 states]	14.12	-149.39	13795	97702	159752	9969	73938
10	MFCC_D	3 (0.5,0.25 x2) [5 states]	14.12	-149.32	13791	97702	159681	9966	73945
11	MFCC_D_A	2 [5 states]	14.29	-164.06	13957	97695	174238	9359	74379
12	MFCC_D_A	2 (0.75,0.25) [5 states]	14.24	-163.48	13910	97698	173630	9467	74321
13	MFCC_D_A	4 [5 states]	14.29	-164.06	13964	97695	174241	9358	74373
14	MFCC_D_A	1 [5 states]	14.29	-164.04	13963	97695	174225	9359	74373
15	MFCC_D_A	3 (0.5,0.25 x2) [5 states]	14.29	-164.06	13962	97695	174239	9359	74374
16	MFCC_D_A_T	2 [5 states]	14.87	-189.69	14519	97644	199736	8593	74532
17	MFCC_D_A_T	2 (0.75,0.25) [5 states]	14.87	-188.42	14526	97678	198571	8740	74412
18	MFCC_D_A_T	4 [5 states]	14.87	-189.73	14518	97644	199778	8595	74531
19	MFCC_D_A_T	1 [5 states]	14.86	-189.69	14513	97644	199734	8591	74540
20	MFCC_D_A_T	3 (0.5,0.25 x2) [5 states]	14.87	-189.70	14519	97644	199749	8595	74530

Table 17: The results from our Phase 1 experiments, where 5-state HMMs with various Cepstrum and Gaussian configurations were prepared using corresponding MFCC-extractions of the Talkbank training data and tested using corresponding MFCC-extractions of the Talkbank testing data.

9.2 Phase 2: Training on TIMIT, Testing on TIMIT

In this phase of our work, we're trying to make the best TIMIT-based phone-recognizer we can, using HMMs and only HMMs, keeping things simple by following Pipeline B. We decided not to use advanced techniques such as tied-state phones and MAP speaker adaptation (Young et al., 2012), opting instead to get as close as the benchmark for HMM-only speech recognition on the TIMIT corpus, which we discussed earlier. This is a correctness of 73.80% and an accuracy of 66.08% and it is documented in Lee & Hon 1989, which they obtained using tied-state phones, their folding mechanism for their 39-phone phoneset, an MFCC configuration that we can describe with our notation as MFCC_E_D_A and a 5-state HMM with 16 Gaussians. Table 18 shows our Phase 2 experiments, where we attempt to get the best possible result, in the ballpark of the Lee & Hon's best result, using our 41-phone TIMIT phoneset described earlier, instead of the 39-phones used by Lee & Hon, while imitating their other settings.

The phone-recognizer we built in Experiment 28 comes closest to the results that Lee & Hon obtained in Lee & Hon 1989. It's correctness is lower, but by only 1.77%. It's accuracy is lower, but only by 3.81%. We decided to commit to this model of our phone-recognizer as the best TIMIT-trained one we could come up with. We will call it "Model 28" for the future reference. We also identified the phone-recognizer we created in Experiment 22 as a simplified form of our TIMIT-based phone-recognizer that would have the fastest processing speed due to having the least amount of Gaussian functions and HMM

states, sacrificing correctness and accuracy in the process. We decided to name this recognizer “Model 22” and use it for future comparison work against Model 28.

Experiment	Cepstrum	Gaussian and HMM Configuration	Correctness	Accuracy	H	N	I	D	S
21	MFCC_E_D_A_T	1 [5 states]	55.20%	40.75%	35407	64145	9268	4632	24106
22	MFCC_E_D_A	1 [5 states]	55.57%	42.31%	35646	64145	8507	4680	23819
23	MFCC_E_D_A	2 [5 states]	60.94%	48.33%	39091	64145	8087	4017	21037
24	MFCC_E_D_A	4 [5 states]	64.06%	52.46%	41093	64145	7441	3795	19257
25	MFCC_E_D_A	6 [5 states]	69.43%	59.84%	44535	64145	6148	4230	15380
26	MFCC_E_D_A	16 [5 states]	71.85%	62.98%	46089	64145	5693	4080	13976
27	MFCC_E_D_A	32 [5 states]	71.85%	62.98%	46089	64145	5693	4080	13976
28	MFCC_E_D_A	16 [7 states]	72.03%	62.27%	46202	64145	6260	3998	13945

Table 18: The results from our Phase 2 experiments, where we try to create a phone-recognizer trained on the TIMIT training dataset, that performs phone recognition on the TIMIT testing set at the same level as what was documented in Lee & Hon 1989.

Using the results we already have for Experiment 28, we can explore the confusion-matrix to get a phone-by-phone correctness value for Model 28. This information is valuable to us when it comes to figuring out which phones are better recognized by Model 28 and which phones are not. Table 19 below shows us this information.

TIMIT Phone	% Correct
b	75.0
d	69.2
g	78.2
p	74.8
t	66.9
k	83.8
ɾ	91.4
ʤ	74.5
ʧ	78.1
s	84.8
ʃ	89.0
z	73.3
ʒ	55.1
f	87.8
θ	63.1
v	74.2
ð	71.3
m	82.9
n	71.2
ŋ	85.9
l	75.4

TIMIT Phone	% Correct
ɹ	75.7
w	85.8
j	84.1
h	84.3
sil	87.4
i	82.4
ɪ	58.7
ɛ	57.4
e_ɪ	79.4
æ	75.3
ɑ	65.4
a_ʊ	67.3
a_ɪ	79.9
ɔ	72.0
ɔ_ɪ	82.6
o_ʊ	67.9
ʊ	38.2
u	66.2
ɝ	77.6
ə	57.6

Table 19: Correctness values by phone, for the results of Experiment 28, which gave us the best results for TIMIT-trained phone recognition against the TIMIT testing dataset.

9.3 Phase 3: Training on Timit, Testing on Talkbank

With Model 28 as our best candidate for phone recognition and Model 22 as our simple-version candidate, we proceeded to point both of them at the Talkbank testing dataset, as well as the entire Talkbank dataset, in order to see if it can be better than what we came up with in our Phase 1 experiments. This involved repeating our Phase 1 experiments, but this time only using the MFCC_E_D_A cepstrum configuration to match the operating parameters of Model 22 and Model 28. Table 20 shows our results.

Experiment	Model	Testing Dataset	Cepstrum	Gaussian and HMM Config.	% Correct	Acc.	H	N	I	D	S
29	22	Talkbank Test	MFCC_E_D_A	1 [5 states]	14.09	-378.80	13782	97839	384394	2289	81768
30	22	Talkbank All	MFCC_E_D_A	1 [5 states]	14.14	-382.33	67924	480283	1904173	11681	400678
31	28	Talkbank Test	MFCC_E_D_A	16 [7 states]	30.48	-484.79	29820	97839	504138	2126	65893
32	28	Talkbank All	MFCC_E_D_A	16 [7 states]	30.45	-488.06	146224	480283	2490284	10678	323381

Table 20: The results for the first set of our Phase 3 experiments, where we use Model 22 and Model 28 of our phone-recognizer to test against our Talkbank testing dataset as well as our entire Talkbank dataset.

At this point in our work, we abandoned further testing with Model 22, seeing how strongly the difference in performance it had with Model 28. All further work for Phase 3 and Phase 4 was conducted with Model 28. We realized that we should probably subdivide our Talkbank dataset by language environment in order to explore the influence Model 28's English-bias on its performance across the three language environments in which we have the recorded utterances of young children. We divided

the Talkbank testing dataset into three divisions organized by language-environment, which we called "TalkbankEngTest", "TalkbankFraTest" and "TalkbankDeuTest". We also divided the entire Talkbank dataset into three similarly-organized divisions that we called "TalkbankEngAll", "TalkbankFraAll" and "TalkbankDeuTest". We then repeated our experiment, using Model 28 to run phone recognition on all six of these language-clustered sets. Table 21a shows our results, which clearly show that there is a significant English-bias occurring, where children from the English language environment enjoy a higher accuracy rate in phone recognition when Model 28 is applied to their utterances.

Experiment	Model	Testing Dataset	Cepstrum	Gaussian and HMM Config.	% Correct	Acc.	H	N	I	D	S
33	28	Talkbank EngTest	MFCC_E_D_A	16 [7 states]	36.22	-543.76	19340	53389	309649	724	33325
34	28	Talkbank EngAll	MFCC_E_D_A	16 [7 states]	36.27	-548.14	94963	261791	1529955	3641	163187
35	28	Talkbank FraTest	MFCC_E_D_A	16 [7 states]	23.10	-615.94	6504	28152	179903	304	21344
36	28	Talkbank FraAll	MFCC_E_D_A	16 [7 states]	23.10	-619.79	32002	138521	890540	1466	105053
37	28	Talkbank DeuTest	MFCC_E_D_A	16 [7 states]	24.40	-65.10	3976	16298	14586	1098	11224
38	28	Talkbank DeuAll	MFCC_E_D_A	16 [7 states]	24.08	-63.19	19259	79971	69789	5571	55141

Table 21a: The results for the second set of our Phase 3 experiments, where we explore the performance of Model 28 by language environment, for our Talkbank testing dataset and our entire Talkbank dataset.

In order to compare the efficacy of our different approaches, we decided to revisit the Phase 1 experiments, using the MFCC_E_D_A configuration (which were not used in that phase) as well as a 7-state 16-Gaussian model (which also were not used in that phase). The table below shows the result of this reference experiment.

Experiment	Training Dataset	Testing Dataset	Cepstrum	Gaussian and HMM Config.	% Correct	Acc.	H	N	I	D	S
39	Talkbank Train	Talkbank Test	MFCC_E_D_A	16 [7 states]	12.87	-92.64	-92.64	12388	101569	1428	69596

Table 21b: Results for a Model 28 style HMM, trained and tested on Talkbank, via Pipeline A.

In Experiment 39, we train a model on the Talkbank training set and test it on the Talkbank testing set, using Cepstrum, Gaussian and HMM parameters similar to Model 28, but using Pipeline A for the workflow. We can see that the results are quite disappointing. We see here that just having the same Cepstrum, Gaussian and HMM parameters as Model 28 is not enough to boost perfomance. It gives us a deeper insight into how the model initialization and training activities of Pipeline B allows us to best utilize the Cepstrum, Gaussian and HMM parameters that make Model 28 effective.

We noticed, in Table 20, that the German language environment data has a markedly lower rate of insertion errors compared to the data from the English and French language environments. Upon investigation, we discovered an oversight on our part, where we had left the English-environment and French-environment Talkbank data at the higher sample rate of 44k instead of downsampling this data to 16k, to match that of the German-environment Talkbank data as well as the data from the TIMIT corpus and the Torgo corpus. In order to set the results up for Experiments 31 to 38 on a more equal footing for all the language-environments, we decided to repeat these experiments again with all utterances forcibly downsampled to 16k using SoX. We denote these re-done experiments with the same number but with the added suffix of “-R” to indicate the revised version of the experiment. Our results are shown below in Table 21c:

Experiment	Model	Testing Dataset	Cepstrum	Gaussian and HMM Config.	% Correct	Acc.	H	N	I	D	S
31-R	28	Talkbank Test	MFCC_E_D_A	16 [7 states]	34.81%	-625.92	34054	97839	646446	2051	61734
32-R	28	Talkbank All	MFCC_E_D_A	16 [7 states]	34.66%	-630.29	166443	480283	3193595	10204	303636
33-R	28	Talkbank EngTest	MFCC_E_D_A	16 [7 states]	38.67%	-695.53	20646	53389	391985	628	32115
34-R	28	Talkbank EngAll	MFCC_E_D_A	16 [7 states]	38.47%	-699.58	100704	261791	1932135	3085	158002
35-R	28	Talkbank FraTest	MFCC_E_D_A	16 [7 states]	33.51%	-818.53	9434	28152	239866	319	18399
36-R	28	Talkbank FraAll	MFCC_E_D_A	16 [7 states]	33.57%	-826.67	46504	138521	1191610	1508	90509
37-R	28	Talkbank DeuTest	MFCC_E_D_A	16 [7 states]	24.38%	-65.17	3974	16298	14595	1104	11220
38-R	28	Talkbank DeuAll	MFCC_E_D_A	16 [7 states]	24.05%	-63.29	19235	79971	69850	5611	55125

Table 21c: Results of revised experiments where forced downsampling was performed on all speech samples.

The results here are surprising. Downsampling did not lead to a drop in insertion error as we expected. Instead it lead to an increase in the insertion error in all cases. It also lead to the correctness levels for English to rise slightly and for French to rise dramatically, with the German correctness levels staying mostly the same. This result is peculiar and we will stop here with the downsampled experiments due to the unexpected nature of these results.

9.4 Phase 4: Training on Timit, Testing on Torgo

We had already divided all the feasible data in our Torgo dataset into two subsets: the “Patients” subset and the “Controls” subset. With Model 28 as our best TIMIT-trained candidate, we proceeded to point it towards speech data to both subsets and to the entire dataset. Table 22 shows the results.

Experiment	Model	Testing Dataset	Cepstrum	Gaussian and HMM Config.	% Correct	Acc.	H	N	I	D	S
40	28	Torgo All	MFCC_E_D_A	16 [7 states]	42.91	-15.37	26932	62767	36578	3519	32316
41	28	Torgo Patients	MFCC_E_D_A	16 [7 states]	48.28	-37.16	13790	28562	24403	1030	13742
42	28	Torgo Controls	MFCC_E_D_A	16 [7 states]	38.42	2.83	13142	34205	12175	2489	18574

Table 22: The results for our Phase 4 experiments, where we explore the performance of Model 28 across our entire Torgo dataset, as well as its “Patients” and “Controls” subsets.

9.5 Phase 5: Exploring optimal VTLN settings for Model 28

For this phase of our work, we chose VTLN to be the key method we will use to accomplish age-sensitive phone recognition. We do not know what optimal “VTLN Warp Factor” variable value is appropriate for each of the 17 age buckets we identified for our “Age At Recording” variable. We also want to be mindful of the language environments for each of these children, which influences the results because each language environment has a different age cluster of children. So for the first phase of our work, we identified 9 values for “VTLN Warp Factor” that we want to test with. Please see Appendix A for HTK configuration information that shows where we would plug-in the VTLN Warp Factor value for each case. The idea was to re-extract all our Talkbank MFCC files with the MFCC_E_D_A configuration such that the MFCC features extracted are warped accordingly with HTK’s VTLN feature (Young et al., 2006).

With 17 values for “Age At Recording” and 9 values for “VTLN Warp Factor”, we ended up conducting 153 experiments to cover all tuple pairings of these values. At this stage of our work, we had become so familiar with this process that were able to write scripts to completely automate the preparation and execution of these experiments, running Model 28 against our entire Talkbank dataset. The results of these experiments are below in Table 23. The entries marked N/A are situations where no children existed with an “Age At Recording” of that value. There are 4 out of 17 values for Age At Recording for which no children exist. These are 0, 5, 7 and 8. The “VTLN Warp Factor” and “Age At Recording” tuples for these situations are still considered experiments for the purpose of referencing. They are just experiments with no applicable results.

Note: The English and French environment data we used here was not downsampled. We used the original data available for all three language environments, focusing on relative correctness as our guide to finding the best VTLN value.

Note: Table 23 has 153 entries and goes on for several pages.

Experiment	**VTLN Factor**	**Age At Recording**	**% Correct**	**Accuracy**	**H**	**N**	**I**	**D**	**S**
43 - 51	All	0.0 – 0.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
52	0.55	0.5 – 1.0	31.44	-641.57	19652	62514	420726	1074	41788
53	0.58	0.5 – 1.0	32.69	-626.73	20436	62514	412230	1067	41011
54	0.63	0.5 – 1.0	34.81	-619.01	21759	62514	408727	1080	39675
55	0.63	0.5 – 1.0	34.81	-619.01	21759	62514	408727	1080	39675
56	0.65	0.5 – 1.0	35.41	-616.1	22134	62514	407281	1038	39342

57	0.68	0.5 – 1.0	36.17	-612.03	22609	62514	405213	926	38979
58	0.72	0.5 – 1.0	36.43	-607.03	22774	62514	402255	1043	38697
59	0.78	0.5 – 1.0	35.99	-601.59	22499	62514	398580	1200	38815
60	0.83	0.5 – 1.0	35.43	-592.79	22148	62514	392725	1152	39214
61	0.55	1.0 – 1.5	40.6	-933.04	3043	7495	72974	39	4413
62	0.58	1.0 – 1.5	42.68	-911.59	3199	7495	71523	39	4257
63	0.63	1.0 – 1.5	45.2	-898.63	3388	7495	70740	61	4046
64	0.63	1.0 – 1.5	45.2	-898.63	3388	7495	70740	61	4046
65	0.65	1.0 – 1.5	45.99	-896.78	3447	7495	70661	73	3975
66	0.68	1.0 – 1.5	46.66	-888.38	3497	7495	70081	63	3935
67	0.72	1.0 – 1.5	47.79	-877.51	3582	7495	69351	77	3836
68	0.78	1.0 – 1.5	47.79	-857.13	3582	7495	67824	72	3841
69	0.83	1.0 – 1.5	47.15	-842.66	3534	7495	66691	98	3863
70	0.55	1.5 – 2.0	31.6	-752.52	92710	293355	2300253	2916	197729
71	0.58	1.5 – 2.0	34.3	-748.34	100610	293355	2295917	2953	189792
72	0.63	1.5 – 2.0	37.27	-742.96	109330	293355	2288852	3063	180962
73	0.63	1.5 – 2.0	37.27	-742.96	109330	293355	2288852	3063	180962
74	0.65	1.5 – 2.0	38.19	-740.24	112038	293355	2283556	3091	178226
75	0.68	1.5 – 2.0	38.86	-736.91	113995	293355	2275743	3172	176188
76	0.72	1.5 – 2.0	39.42	-733.54	115651	293355	2267515	3149	174555
77	0.78	1.5 – 2.0	39.96	-729.02	117212	293355	2255839	3317	172826
78	0.83	1.5 – 2.0	39.26	-725.31	115159	293355	2242901	3710	174486
79	0.55	2.0 – 2.5	33.08	-874.86	1885	5699	51743	47	3767
80	0.58	2.0 – 2.5	35.48	-862.68	2022	5699	51186	60	3617
81	0.63	2.0 – 2.5	38.23	-848.13	2179	5699	50514	60	3460
82	0.63	2.0 – 2.5	38.23	-848.13	2179	5699	50514	60	3460
83	0.65	2.0 – 2.5	39.39	-842.48	2245	5699	50258	67	3387

84	0.68	2.0 – 2.5	40.22	-835.37	2292	5699	49900	71	3336
85	0.72	2.0 – 2.5	41.01	-828.57	2337	5699	49557	75	3287
86	0.78	2.0 – 2.5	41.29	-817.49	2353	5699	48942	79	3267
87	0.83	2.0 – 2.5	40.29	-821.57	2296	5699	49117	84	3319
88	0.55	2.5 – 3.0	30.77	-575.15	21369	69450	420811	1519	46562
89	0.58	2.5 – 3.0	33.18	-567.17	23044	69450	416947	1465	44941
90	0.63	2.5 – 3.0	36.76	-560.59	25529	69450	414862	1439	42482
91	0.63	2.5 – 3.0	36.76	-560.59	25529	69450	414862	1439	42482
92	0.65	2.5 – 3.0	37.88	-558.46	26310	69450	414163	1452	41688
93	0.68	2.5 – 3.0	39.28	-554.02	27280	69450	412045	1507	40663
94	0.72	2.5 – 3.0	40.27	-549.33	27968	69450	409475	1507	39975
95	0.78	2.5 – 3.0	40.52	-546.37	28139	69450	407593	1526	39785
96	0.83	2.5 – 3.0	39.95	-544.72	27742	69450	406051	1559	40149
97	0.55	3.0 – 3.5	21.39	-345.79	404	1889	6936	76	1409
98	0.58	3.0 – 3.5	24.83	-346.85	469	1889	7021	52	1368
99	0.63	3.0 – 3.5	28.69	-347.86	542	1889	7113	57	1290
100	0.63	3.0 – 3.5	28.69	-347.86	542	1889	7113	57	1290
101	0.65	3.0 – 3.5	30.65	-343.99	579	1889	7077	49	1261
102	0.68	3.0 – 3.5	33.19	-346.85	627	1889	7179	52	1210
103	0.72	3.0 – 3.5	35.42	-350.03	669	1889	7281	55	1165
104	0.78	3.0 – 3.5	36.79	-353.41	695	1889	7371	66	1128
105	0.83	3.0 – 3.5	36.16	-352.41	683	1889	7340	60	1146
106	0.55	3.5 – 4.0	14.61	-53.66	2829	19357	13215	1498	15030
107	0.58	3.5 – 4.0	18.21	-56.75	3525	19357	14511	1414	14418
108	0.63	3.5 – 4.0	22.72	-54.26	4397	19357	14901	1398	13562
109	0.63	3.5 – 4.0	22.72	-54.26	4397	19357	14901	1398	13562
110	0.65	3.5 – 4.0	24.09	-53.05	4664	19357	14933	1409	13284

111	0.68	3.5 – 4.0	25.64	-52.67	4963	19357	15158	1364	13030
112	0.72	3.5 – 4.0	27.35	-51.92	5295	19357	15346	1420	12642
113	0.78	3.5 – 4.0	27.86	-50.7	5392	19357	15206	1466	12499
114	0.83	3.5 – 4.0	27.08	-51.84	5241	19357	15276	1459	12657
115	0.55	4.0 – 4.5	17.43	-70.9	499	2863	2529	153	2211
116	0.58	4.0 – 4.5	21.34	-69.23	611	2863	2593	159	2093
117	0.63	4.0 – 4.5	27.31	-63.95	782	2863	2613	159	1922
118	0.63	4.0 – 4.5	27.31	-63.95	782	2863	2613	159	1922
119	0.65	4.0 – 4.5	28.08	-65.28	804	2863	2673	162	1897
120	0.68	4.0 – 4.5	30.28	-64.69	867	2863	2719	172	1824
121	0.72	4.0 – 4.5	30.98	-64.06	887	2863	2721	175	1801
122	0.78	4.0 – 4.5	29.93	-63.57	857	2863	2677	195	1811
123	0.83	4.0 – 4.5	28.36	-65.77	812	2863	2695	184	1867
124	0.55	4.5 – 5.0	12.92	-47.26	1399	10829	6517	900	8530
125	0.58	4.5 – 5.0	16.08	-53.17	1741	10829	7499	753	8335
126	0.63	4.5 – 5.0	21.23	-52.46	2299	10829	7980	730	7800
127	0.63	4.5 – 5.0	21.23	-52.46	2299	10829	7980	730	7800
128	0.65	4.5 – 5.0	22.63	-52.29	2451	10829	8113	760	7618
129	0.68	4.5 – 5.0	24.28	-51.71	2629	10829	8229	777	7423
130	0.72	4.5 – 5.0	25.39	-49.93	2749	10829	8156	756	7324
131	0.78	4.5 – 5.0	27.1	-48.81	2935	10829	8221	806	7088
132	0.83	4.5 – 5.0	27.11	-48.73	2936	10829	8213	826	7067
133 - 141	All	5.0 – 5.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
142	0.55	5.0 – 5.5	13.73	-58.81	187	1362	988	105	1070
143	0.58	5.0 – 5.5	16.89	-60.65	230	1362	1056	88	1044
144	0.63	5.0 – 5.5	20.85	-63.44	284	1362	1148	76	1002
145	0.63	5.0 – 5.5	20.85	-63.44	284	1362	1148	76	1002

146	0.65	5.0 – 5.5	22.03	-65.93	300	1362	1198	82	980
147	0.68	5.0 – 5.5	22.91	-65.57	312	1362	1205	72	978
148	0.72	5.0 – 5.5	25.92	-67.18	353	1362	1268	78	931
149	0.78	5.0 – 5.5	25.62	-67.47	349	1362	1268	82	931
150	0.83	5.0 – 5.5	24.08	-67.62	328	1362	1249	88	946
151	0.55	6.0 – 6.5	11.24	-46.93	75	667	388	51	541
152	0.58	6.0 – 6.5	16.94	-41.68	113	667	391	39	515
153	0.63	6.0 – 6.5	20.69	-38.68	138	667	396	44	485
154	0.63	6.0 – 6.5	20.69	-38.68	138	667	396	44	485
155	0.65	6.0 – 6.5	22.34	-38.08	149	667	403	54	464
156	0.68	6.0 – 6.5	25.19	-38.83	168	667	427	51	448
157	0.72	6.0 – 6.5	28.79	-39.13	192	667	453	45	430
158	0.78	6.0 – 6.5	29.99	-38.53	200	667	457	39	428
159	0.83	6.0 – 6.5	28.79	-40.48	192	667	462	35	440
160	0.55	6.5 – 7.0	12.42	-46.16	259	2086	1222	126	1701
161	0.58	6.5 – 7.0	16.2	-47.08	338	2086	1320	123	1625
162	0.63	6.5 – 7.0	20.57	-45.64	429	2086	1381	133	1524
163	0.63	6.5 – 7.0	20.57	-45.64	429	2086	1381	133	1524
164	0.65	6.5 – 7.0	23.25	-43.05	485	2086	1383	118	1483
165	0.68	6.5 – 7.0	25.84	-42.67	539	2086	1429	116	1431
166	0.72	6.5 – 7.0	28.04	-42.14	585	2086	1464	120	1381
167	0.78	6.5 – 7.0	30.58	-39.17	638	2086	1455	125	1323
168	0.83	6.5 – 7.0	30.2	-42.62	630	2086	1519	136	1320
169 - 177	All	7.0 – 7.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
178	0.55	7.5 – 8.0	14.57	-54.36	396	2717	1873	169	2152
179	0.58	7.5 – 8.0	18.15	-52.48	493	2717	1919	177	2047
180	0.63	7.5 – 8.0	23.22	-48.14	631	2717	1939	171	1915

181	0.63	7.5 – 8.0	23.22	-48.14	631	2717	1939	171	1915
182	0.65	7.5 – 8.0	25.32	-45.56	688	2717	1926	171	1858
183	0.68	7.5 – 8.0	28.63	-43.8	778	2717	1968	182	1757
184	0.72	7.5 – 8.0	30.36	-44.09	825	2717	2023	162	1730
185	0.78	7.5 – 8.0	31.54	-41.37	857	2717	1981	166	1694
186	0.83	7.5 – 8.0	33.05	-40.96	898	2717	2011	152	1667
187 - 195	All	8.0 – 8.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 23: The results for the first set of experiments we performed for Phase 5. We explored 9 values for VTLN Warp Factor and 17 values for Age At Recording for the entire Talkbank dataset, containing data from all three language environments, running Model 28 on all scenarios that were applicable.

This massive table is difficult to absorb visually. See Figure 1 below for a useful visualization of the 117 values for correctness that were applicable.

Our next exploration for Phase 5 was repeating the 153 experiments shown in Table 23, but now isolating the results for each language environment along with each age cluster. Table 24 shows how many phones exist for each language environment, clustered by age.

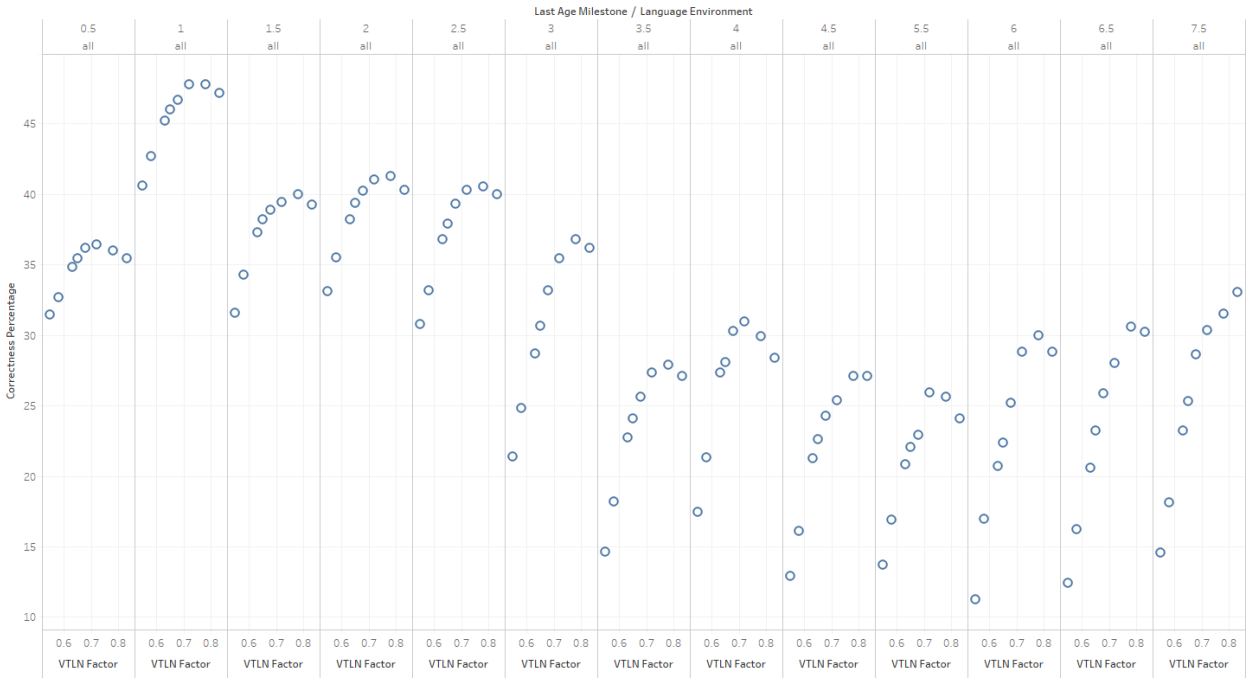

Figure 2: Visualization of results for the first set of our Phase 5 experiments, showing how correctness changes across values for VTLN Warp Factor and the lower bound of “Age At Recording”, defined as “Last Age Milestone”, across all language environments.

Age At Recording	Language Env.	N
0.0 – 0.5	deu	4169
	eng	40701
	fra	17644
0.5 – 1.0	deu	0
	eng	0
	fra	0
1.0 – 1.5	deu	0
	eng	5414
	fra	2081
1.5 – 2.0	deu	15823
	eng	174300
	fra	103232
2.0 – 2.5	deu	0
	eng	1905
	fra	3794
2.5 – 3.0	deu	18942
	eng	38738
	fra	11770
3.0 – 3.5	deu	1156
	eng	733
	fra	0
3.5 – 4.0	deu	19357
	eng	0
	fra	0
4.0 – 4.5	deu	2863
	eng	0
	fra	0
4.5 – 5.0	deu	10829
	eng	0
	fra	0
5.0 – 5.5	deu	0
	eng	0
	fra	0
5.5 – 6.0	deu	1362
	eng	0
	fra	0
6.0 – 6.5	deu	667
	eng	0
	fra	0
6.5 – 7.0	deu	2086
	eng	0
	fra	0
7.0 – 7.5	deu	0
	eng	0
	fra	0
7.5 – 8.0	deu	2717
	eng	0
	fra	0
8.0 – 8.5	deu	0
	eng	0
	fra	0

Table 24: The number of human-transcribed phones clustered by Age At Recording and Language Environment.

Figure 2 shows a visualization of the information in Table 24.

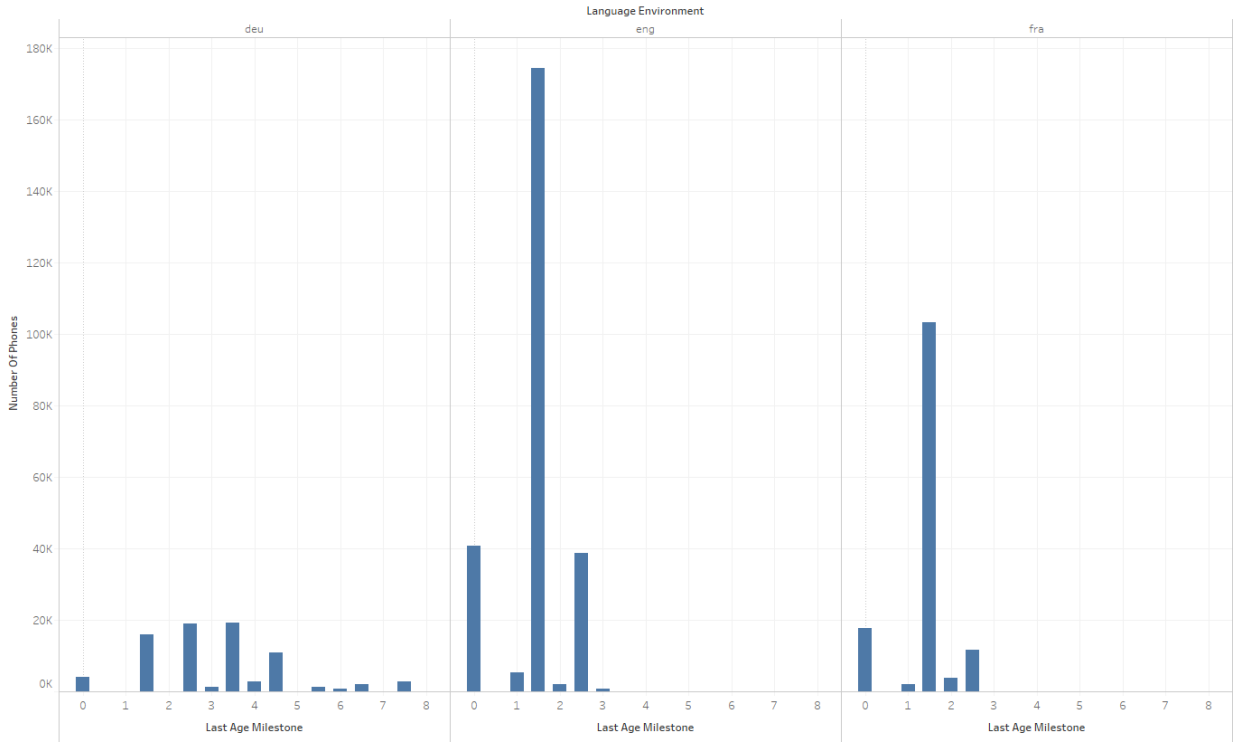

Figure 3: Bar graph showing how phones in the Talkbank dataset cluster by language-environment and the lower-bound of Age At Recording, defined here as “Last Age Milestone”.

Tables 25, 26 and 27 show the results of the Phase 5 experiments that we ran for the English, French and German language environments respectively, using their respective data subsets from the entire Talkbank dataset. These tables will take up several of the following pages.

Experiment	VTLN Factor	Age At Recording	Correctness	Accuracy	H	N	I	D	S
196	0.55	0.0 – 0.5	35.1	-516.16	14287	40701	224370	491	25923
197	0.58	0.0 – 0.5	36.7	-506.69	14938	40701	221167	519	25244
198	0.63	0.0 – 0.5	38.66	-502.03	15733	40701	220065	533	24435
199	0.63	0.0 – 0.5	38.66	-502.03	15733	40701	220065	533	24435
200	0.65	0.0 – 0.5	39.32	-501.14	16005	40701	219976	517	24179
201	0.68	0.0 – 0.5	39.74	-501.34	16176	40701	220228	488	24037
202	0.72	0.0 – 0.5	40.56	-499.56	16509	40701	219834	539	23653
203	0.78	0.0 – 0.5	39.72	-495.86	16165	40701	217985	609	23927
204	0.83	0.0 – 0.5	38.88	-487.69	15826	40701	214320	581	24294
205 – 213	All	0.5 – 1.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
214	0.55	1.0 – 1.5	42.43	-805.19	2297	5414	45890	25	3092
215	0.58	1.0 – 1.5	44.38	-788.23	2403	5414	45078	29	2982
216	0.63	1.0 – 1.5	47.12	-783.58	2551	5414	44974	50	2813
217	0.63	1.0 – 1.5	47.12	-783.58	2551	5414	44974	50	2813
218	0.65	1.0 – 1.5	48.08	-784.37	2603	5414	45069	52	2759
219	0.68	1.0 – 1.5	48.78	-777.28	2641	5414	44723	41	2732
220	0.72	1.0 – 1.5	50.65	-768.43	2742	5414	44345	48	2624
221	0.78	1.0 – 1.5	50.37	-745.33	2727	5414	43079	58	2629
222	0.83	1.0 – 1.5	49.32	-728.98	2670	5414	42137	85	2659
223	0.55	1.5 – 2.0	35.65	-765.23	62144	174300	1395939	1473	110683
224	0.58	1.5 – 2.0	38.57	-760	67230	174300	1391905	1450	105620
225	0.63	1.5 – 2.0	41.6	-755.6	72504	174300	1389508	1497	100299
226	0.63	1.5 – 2.0	41.6	-755.6	72504	174300	1389508	1497	100299
227	0.65	1.5 – 2.0	42.64	-752.98	74322	174300	1386758	1456	98522
228	0.68	1.5 – 2.0	43.15	-750.66	75206	174300	1383614	1515	97579

229	0.72	1.5 – 2.0	43.37	-748.99	75586	174300	1381077	1448	97266
230	0.78	1.5 – 2.0	43.49	-744.45	75801	174300	1373384	1582	96917
231	0.83	1.5 – 2.0	42.1	-739.51	73376	174300	1362344	1964	98960
232	0.55	2.0 – 2.5	46.93	-987.24	894	1905	19701	14	997
233	0.58	2.0 – 2.5	49.5	-964.46	943	1905	19316	16	946
234	0.63	2.0 – 2.5	52.49	-945.3	1000	1905	19008	15	890
235	0.63	2.0 – 2.5	52.49	-945.3	1000	1905	19008	15	890
236	0.65	2.0 – 2.5	53.81	-933.86	1025	1905	18815	18	862
237	0.68	2.0 – 2.5	54.38	-921.47	1036	1905	18590	12	857
238	0.72	2.0 – 2.5	54.8	-915.49	1044	1905	18484	18	843
239	0.78	2.0 – 2.5	53.33	-906.88	1016	1905	18292	17	872
240	0.83	2.0 – 2.5	50.55	-913.28	963	1905	18361	19	923
241	0.55	2.5 – 3.0	40.83	-780.49	15818	38738	318164	347	22573
242	0.58	2.5 – 3.0	42.83	-767.62	16590	38738	313950	333	21815
243	0.63	2.5 – 3.0	45.91	-757.95	17786	38738	311402	300	20652
244	0.63	2.5 – 3.0	45.91	-757.95	17786	38738	311402	300	20652
245	0.65	2.5 – 3.0	46.76	-755.75	18115	38738	310877	292	20331
246	0.68	2.5 – 3.0	47.64	-749.81	18455	38738	308916	319	19964
247	0.72	2.5 – 3.0	47.86	-743.12	18539	38738	306410	363	19836
248	0.78	2.5 – 3.0	47.13	-739.61	18257	38738	304768	403	20078
249	0.83	2.5 – 3.0	45.89	-736.8	17776	38738	303197	398	20564
250	0.55	3.0 – 3.5	36.15	-806.96	265	733	6180	10	458
251	0.58	3.0 – 3.5	39.43	-802.32	289	733	6170	5	439
252	0.63	3.0 – 3.5	42.16	-800.68	309	733	6178	5	419
253	0.63	3.0 – 3.5	42.16	-800.68	309	733	6178	5	419
254	0.65	3.0 – 3.5	43.66	-799.18	320	733	6178	1	412
255	0.68	3.0 – 3.5	45.16	-807.64	331	733	6251	1	401

256	0.72	3.0 – 3.5	46.11	-815.55	338	733	6316	3	392
257	0.78	3.0 – 3.5	46.38	-832.61	340	733	6443	4	389
258	0.83	3.0 – 3.5	44.88	-836.7	329	733	6462	3	401
259 – 267	All	3.5 – 4.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
268 – 276	All	4.0 – 4.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
277 – 285	All	4.5 – 5.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
286 – 294	All	5.0 – 5.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
295 – 303	All	5.5 – 6.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
304 – 312	All	6.0 – 6.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
310 – 321	All	6.5 – 7.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
322 – 330	All	7.0 – 7.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
331 – 339	All	7.5 – 8.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
340 – 348	All	8.0 – 8.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 25: The results for the second set of experiments we performed for Phase 5. We explored 9 values for VTLN Warp Factor and 17 values for Age At Recording for the English-environment subset of the entire Talkbank dataset, running Model 28 on all scenarios that were applicable.

Experiment	**VTLN Factor**	**Age at Recording**	**Correctness**	**Accuracy**	**H**	**N**	**I**	**D**	**S**
349	0.55	0.0 – 0.5	25.35	-1060.5	4472	17644	191586	111	13061
350	0.58	0.0 – 0.5	25.61	-1029.1	4519	17644	186093	113	13012
351	0.63	0.0 – 0.5	28.11	-1013.28	4960	17644	183743	133	12551
352	0.63	0.0 – 0.5	28.11	-1013.28	4960	17644	183743	133	12551
353	0.65	0.0 – 0.5	28.64	-1005.06	5053	17644	182386	145	12446
354	0.68	0.0 – 0.5	30.21	-989.81	5330	17644	179972	160	12154
355	0.72	0.0 – 0.5	29.21	-976.65	5153	17644	177473	153	12338
356	0.78	0.0 – 0.5	29.98	-967.96	5289	17644	176076	159	12196
357	0.83	0.0 – 0.5	30.16	-955.62	5321	17644	173930	157	12166
358 – 366	All	0.5 – 1.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A

367	0.55	1.0 – 1.5	35.85	-1265.64	746	2081	27084	14	1321
368	0.58	1.0 – 1.5	38.25	-1232.53	796	2081	26445	10	1275
369	0.63	1.0 – 1.5	40.22	-1197.93	837	2081	25766	11	1233
370	0.63	1.0 – 1.5	40.22	-1197.93	837	2081	25766	11	1233
371	0.65	1.0 – 1.5	40.56	-1189.24	844	2081	25592	21	1216
372	0.68	1.0 – 1.5	41.13	-1177.41	856	2081	25358	22	1203
373	0.72	1.0 – 1.5	40.37	-1161.27	840	2081	25006	29	1212
374	0.78	1.0 – 1.5	41.09	-1148.01	855	2081	24745	14	1212
375	0.83	1.0 – 1.5	41.52	-1138.4	864	2081	24554	13	1204
376	0.55	1.5 – 2.0	26.73	-835.67	27593	103232	890271	783	74856
377	0.58	1.5 – 2.0	28.88	-831.94	29818	103232	888647	828	72586
378	0.63	1.5 – 2.0	31.58	-823.95	32603	103232	883180	897	69732
379	0.63	1.5 – 2.0	31.58	-823.95	32603	103232	883180	897	69732
380	0.65	1.5 – 2.0	32.31	-820.7	33357	103232	880581	946	68929
381	0.68	1.5 – 2.0	33.18	-815.27	34249	103232	875867	931	68052
382	0.72	1.5 – 2.0	34.27	-808.56	35377	103232	870070	1007	66848
383	0.78	1.5 – 2.0	35.44	-803.26	36590	103232	865816	981	65661
384	0.83	1.5 – 2.0	35.82	-800.97	36979	103232	863841	1030	65223
385	0.55	2.0 – 2.5	26.12	-818.42	991	3794	32042	33	2770
386	0.58	2.0 – 2.5	28.44	-811.57	1079	3794	31870	44	2671
387	0.63	2.0 – 2.5	31.08	-799.34	1179	3794	31506	45	2570
388	0.63	2.0 – 2.5	31.08	-799.34	1179	3794	31506	45	2570
389	0.65	2.0 – 2.5	32.16	-796.6	1220	3794	31443	49	2525
390	0.68	2.0 – 2.5	33.1	-792.15	1256	3794	31310	59	2479
391	0.72	2.0 – 2.5	34.08	-784.92	1293	3794	31073	57	2444
392	0.78	2.0 – 2.5	35.24	-772.61	1337	3794	30650	62	2395
393	0.83	2.0 – 2.5	35.13	-775.51	1333	3794	30756	65	2396

394	0.55	2.5 – 3.0	22.83	-722.51	2687	11770	87727	70	9013
395	0.58	2.5 – 3.0	25.06	-714.6	2950	11770	87059	91	8729
396	0.63	2.5 – 3.0	28.37	-711.1	3339	11770	87035	76	8355
397	0.63	2.5 – 3.0	28.37	-711.1	3339	11770	87035	76	8355
398	0.65	2.5 – 3.0	29.36	-707.73	3456	11770	86756	84	8230
399	0.68	2.5 – 3.0	31.37	-703.47	3692	11770	86490	106	7972
400	0.72	2.5 – 3.0	32.99	-699.96	3883	11770	86268	115	7772
401	0.78	2.5 – 3.0	35.19	-697.35	4142	11770	86220	124	7504
402	0.83	2.5 – 3.0	35.74	-695.96	4207	11770	86122	127	7436
403 – 411	All	3.0 – 3.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
412 – 420	All	3.5 – 4.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
421 – 429	All	4.0 – 4.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
430 – 438	All	4.5 – 5.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
439 – 447	All	5.0 – 5.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
448 – 456	All	5.5 – 6.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
457 – 465	All	6.0 – 6.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
466 – 474	All	6.5 – 7.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
475 – 483	All	7.0 – 7.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
484 – 492	All	7.5 – 8.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
493 – 501	All	8.0 – 8.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 26: The results for the third set of experiments we performed for Phase 5. We explored 9 values for VTLN Warp Factor and 17 values for Age At Recording for the French-environment subset of the entire Talkbank dataset, running Model 28 on all scenarios that were applicable.

Experiment	**VTLN Factor**	**Age at Recording**	**Correctness**	**Accuracy**	**H**	**N**	**I**	**D**	**S**
502	0.55	0.0 – 0.5	21.42	-93.00	893	4169	4770	472	2804
503	0.58	0.0 – 0.5	23.48	-95.73	979	4169	4970	435	2755
504	0.63	0.0 – 0.5	25.57	-92.42	1066	4169	4919	414	2689

505	0.63	0.0 – 0.5	25.57	-92.42	1066	4169	4919	414	2689
506	0.65	0.0 – 0.5	25.81	-92.18	1076	4169	4919	376	2717
507	0.68	0.0 – 0.5	26.46	-93.79	1103	4169	5013	278	2788
508	0.72	0.0 – 0.5	26.67	-92.01	1112	4169	4948	351	2706
509	0.78	0.0 – 0.5	25.07	-83.33	1045	4169	4519	432	2692
510	0.83	0.0 – 0.5	24.01	-83.33	1001	4169	4475	414	2754
511 – 519	All	0.5 – 1.0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
520 – 528	All	1.0 – 1.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
529	0.55	1.5 – 2.0	18.79	-69.96	2973	15823	14043	660	12190
530	0.58	1.5 – 2.0	22.51	-74.59	3562	15823	15365	675	11586
531	0.63	1.5 – 2.0	26.69	-75.47	4223	15823	16164	669	10931
532	0.63	1.5 – 2.0	26.69	-75.47	4223	15823	16164	669	10931
533	0.65	1.5 – 2.0	27.55	-74.94	4359	15823	16217	689	10775
534	0.68	1.5 – 2.0	28.69	-74.08	4540	15823	16262	726	10557
535	0.72	1.5 – 2.0	29.63	-73.82	4688	15823	16368	694	10441
536	0.78	1.5 – 2.0	30.47	-74.69	4821	15823	16639	754	10248
537	0.83	1.5 – 2.0	30.36	-75.28	4804	15823	16716	716	10303
538 – 546	All	2.0 – 2.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
547	0.55	2.5 – 3.0	15.12	-63.65	2864	18942	14920	1102	14976
548	0.58	2.5 – 3.0	18.5	-65.64	3504	18942	15938	1041	14397
549	0.63	2.5 – 3.0	23.25	-63.46	4404	18942	16425	1063	13475
550	0.63	2.5 – 3.0	23.25	-63.46	4404	18942	16425	1063	13475
551	0.65	2.5 – 3.0	25.02	-62.25	4739	18942	16530	1076	13127
552	0.68	2.5 – 3.0	27.1	-60.74	5133	18942	16639	1082	12727
553	0.72	2.5 – 3.0	29.28	-59.4	5546	18942	16797	1029	12367
554	0.78	2.5 – 3.0	30.3	-57.36	5740	18942	16605	999	12203
555	0.83	2.5 – 3.0	30.4	-57.93	5759	18942	16732	1034	12149

556	0.55	3.0 – 3.5	12.02	-53.37	139	1156	756	66	951
557	0.58	3.0 – 3.5	15.57	-58.04	180	1156	851	47	929
558	0.63	3.0 – 3.5	20.16	-60.73	233	1156	935	52	871
559	0.63	3.0 – 3.5	20.16	-60.73	233	1156	935	52	871
560	0.65	3.0 – 3.5	22.4	-55.36	259	1156	899	48	849
561	0.68	3.0 – 3.5	25.61	-54.67	296	1156	928	51	809
562	0.72	3.0 – 3.5	28.63	-54.84	331	1156	965	52	773
563	0.78	3.0 – 3.5	30.71	-49.57	355	1156	928	62	739
564	0.83	3.0 – 3.5	30.62	-45.33	354	1156	878	57	745
565	0.55	3.5 – 4.0	14.61	-53.66	2829	19357	13215	1498	15030
566	0.58	3.5 – 4.0	18.21	-56.75	3525	19357	14511	1414	14418
567	0.63	3.5 – 4.0	22.72	-54.26	4397	19357	14901	1398	13562
568	0.63	3.5 – 4.0	22.72	-54.26	4397	19357	14901	1398	13562
569	0.65	3.5 – 4.0	24.09	-53.05	4664	19357	14933	1409	13284
570	0.68	3.5 – 4.0	25.64	-52.67	4963	19357	15158	1364	13030
571	0.72	3.5 – 4.0	27.35	-51.92	5295	19357	15346	1420	12642
572	0.78	3.5 – 4.0	27.86	-50.7	5392	19357	15206	1466	12499
573	0.83	3.5 – 4.0	27.08	-51.84	5241	19357	15276	1459	12657
574	0.55	4.0 – 4.5	17.43	-70.9	499	2863	2529	153	2211
575	0.58	4.0 – 4.5	21.34	-69.23	611	2863	2593	159	2093
576	0.63	4.0 – 4.5	27.31	-63.95	782	2863	2613	159	1922
577	0.63	4.0 – 4.5	27.31	-63.95	782	2863	2613	159	1922
578	0.65	4.0 – 4.5	28.08	-65.28	804	2863	2673	162	1897
579	0.68	4.0 – 4.5	30.28	-64.69	867	2863	2719	172	1824
580	0.72	4.0 – 4.5	30.98	-64.06	887	2863	2721	175	1801
581	0.78	4.0 – 4.5	29.93	-63.57	857	2863	2677	195	1811
582	0.83	4.0 – 4.5	28.36	-65.77	812	2863	2695	184	1867

583	0.55	4.5 – 5.0	12.92	-47.26	1399	10829	6517	900	8530
584	0.58	4.5 – 5.0	16.08	-53.17	1741	10829	7499	753	8335
585	0.63	4.5 – 5.0	21.23	-52.46	2299	10829	7980	730	7800
586	0.63	4.5 – 5.0	21.23	-52.46	2299	10829	7980	730	7800
587	0.65	4.5 – 5.0	22.63	-52.29	2451	10829	8113	760	7618
588	0.68	4.5 – 5.0	24.28	-51.71	2629	10829	8229	777	7423
589	0.72	4.5 – 5.0	25.39	-49.93	2749	10829	8156	756	7324
590	0.78	4.5 – 5.0	27.1	-48.81	2935	10829	8221	806	7088
591	0.83	4.5 – 5.0	27.11	-48.73	2936	10829	8213	826	7067
592 – 600	All	5.0 – 5.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
601	0.55	5.5 – 6.0	13.73	-58.81	187	1362	988	105	1070
602	0.58	5.5 – 6.0	16.89	-60.65	230	1362	1056	88	1044
603	0.63	5.5 – 6.0	20.85	-63.44	284	1362	1148	76	1002
604	0.63	5.5 – 6.0	20.85	-63.44	284	1362	1148	76	1002
605	0.65	5.5 – 6.0	22.03	-65.93	300	1362	1198	82	980
606	0.68	5.5 – 6.0	22.91	-65.57	312	1362	1205	72	978
607	0.72	5.5 – 6.0	25.92	-67.18	353	1362	1268	78	931
608	0.78	5.5 – 6.0	25.62	-67.47	349	1362	1268	82	931
609	0.83	5.5 – 6.0	24.08	-67.62	328	1362	1249	88	946
610	0.55	6.0 – 6.5	11.24	-46.93	75	667	388	51	541
611	0.58	6.0 – 6.5	16.94	-41.68	113	667	391	39	515
612	0.63	6.0 – 6.5	20.69	-38.68	138	667	396	44	485
613	0.63	6.0 – 6.5	20.69	-38.68	138	667	396	44	485
614	0.65	6.0 – 6.5	22.34	-38.08	149	667	403	54	464
615	0.68	6.0 – 6.5	25.19	-38.83	168	667	427	51	448
616	0.72	6.0 – 6.5	28.79	-39.13	192	667	453	45	430
617	0.78	6.0 – 6.5	29.99	-38.53	200	667	457	39	428

618	0.83	6.0 – 6.5	28.79	-40.48	192	667	462	35	440
619	0.55	6.5 – 7.0	12.42	-46.16	259	2086	1222	126	1701
620	0.58	6.5 – 7.0	16.2	-47.08	338	2086	1320	123	1625
621	0.63	6.5 – 7.0	20.57	-45.64	429	2086	1381	133	1524
622	0.63	6.5 – 7.0	20.57	-45.64	429	2086	1381	133	1524
623	0.65	6.5 – 7.0	23.25	-43.05	485	2086	1383	118	1483
624	0.68	6.5 – 7.0	25.84	-42.67	539	2086	1429	116	1431
625	0.72	6.5 – 7.0	28.04	-42.14	585	2086	1464	120	1381
626	0.78	6.5 – 7.0	30.58	-39.17	638	2086	1455	125	1323
627	0.83	6.5 – 7.0	30.2	-42.62	630	2086	1519	136	1320
628 – 636	All	7.0 – 7.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A
637	0.55	7.5 – 8.0	14.57	-54.36	396	2717	1873	169	2152
638	0.58	7.5 – 8.0	18.15	-52.48	493	2717	1919	177	2047
639	0.63	7.5 – 8.0	23.22	-48.14	631	2717	1939	171	1915
640	0.63	7.5 – 8.0	23.22	-48.14	631	2717	1939	171	1915
641	0.65	7.5 – 8.0	25.32	-45.56	688	2717	1926	171	1858
642	0.68	7.5 – 8.0	28.63	-43.8	778	2717	1968	182	1757
643	0.72	7.5 – 8.0	30.36	-44.09	825	2717	2023	162	1730
644	0.78	7.5 – 8.0	31.54	-41.37	857	2717	1981	166	1694
645	0.83	7.5 – 8.0	33.05	-40.96	898	2717	2011	152	1667
646 – 654	All	8.0 – 8.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Table 27: The results for the third set of experiments we performed for Phase 5. We explored 9 values for VTLN Warp Factor and 17 values for Age At Recording for the German-environment subset of the entire Talkbank dataset, running Model 28 on all scenarios that were applicable.

The material in Tables 25, 26 and 27 are difficult to digest. So we created a visualization of this data in

Figures 3, 4 and 5 below to give us more insight into what the data is telling us.

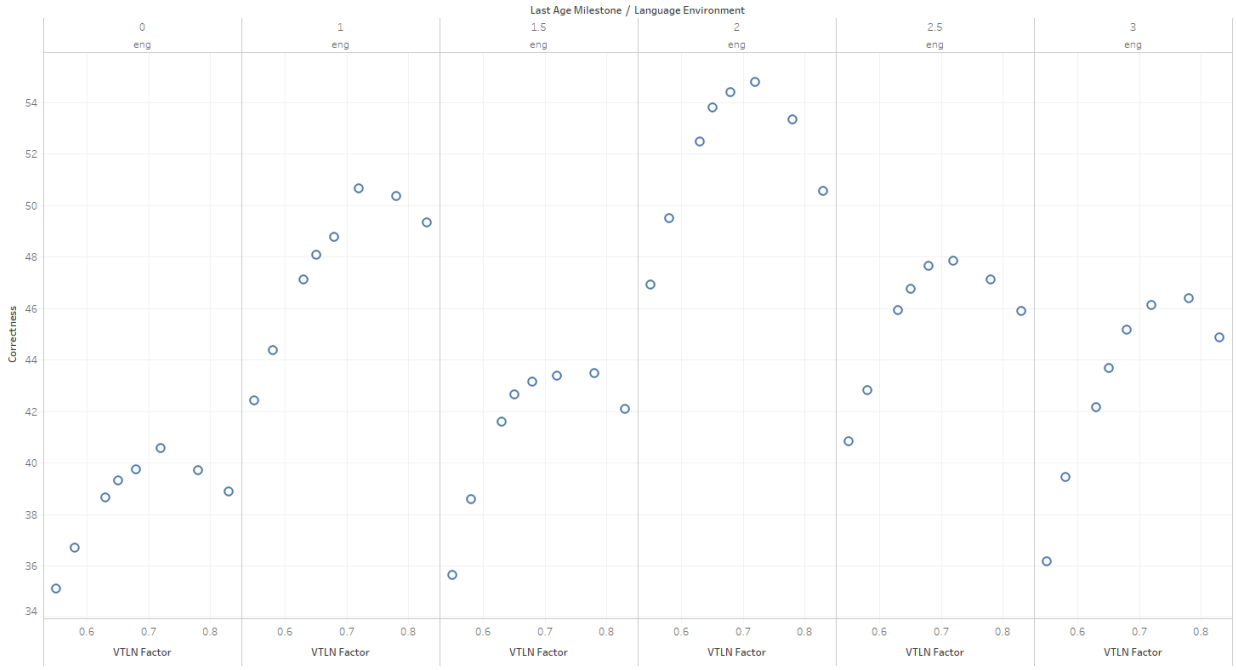

Figure 4: Visualization of results for the second set of our Phase 5 experiments, in the English language environment, showing how correctness changes across values for VTLN Warp Factor and the lower-bound of Age At Recording, defined here as “Last Age Milestone”.

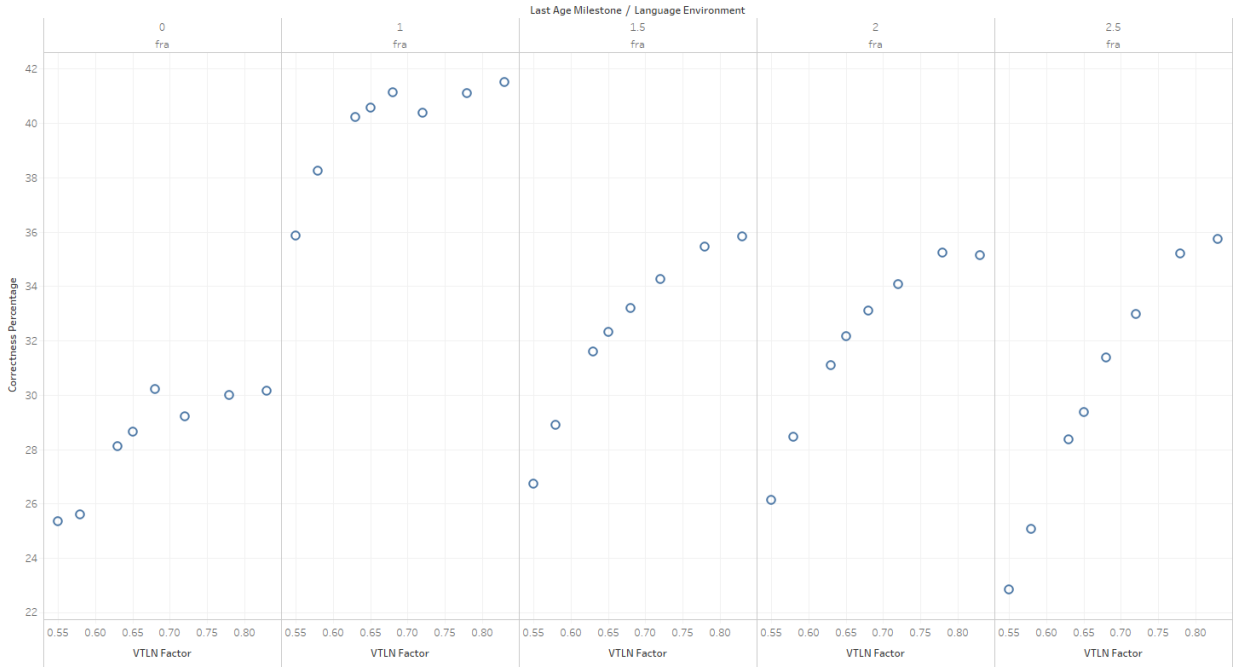

Figure 5: Visualization of results for the third set of our Phase 5 experiments, for the French environment, showing how correctness changes across values for VTLN Warp Factor and the lower-bound of Age At Recording, defined here as “Last Age Milestone”.

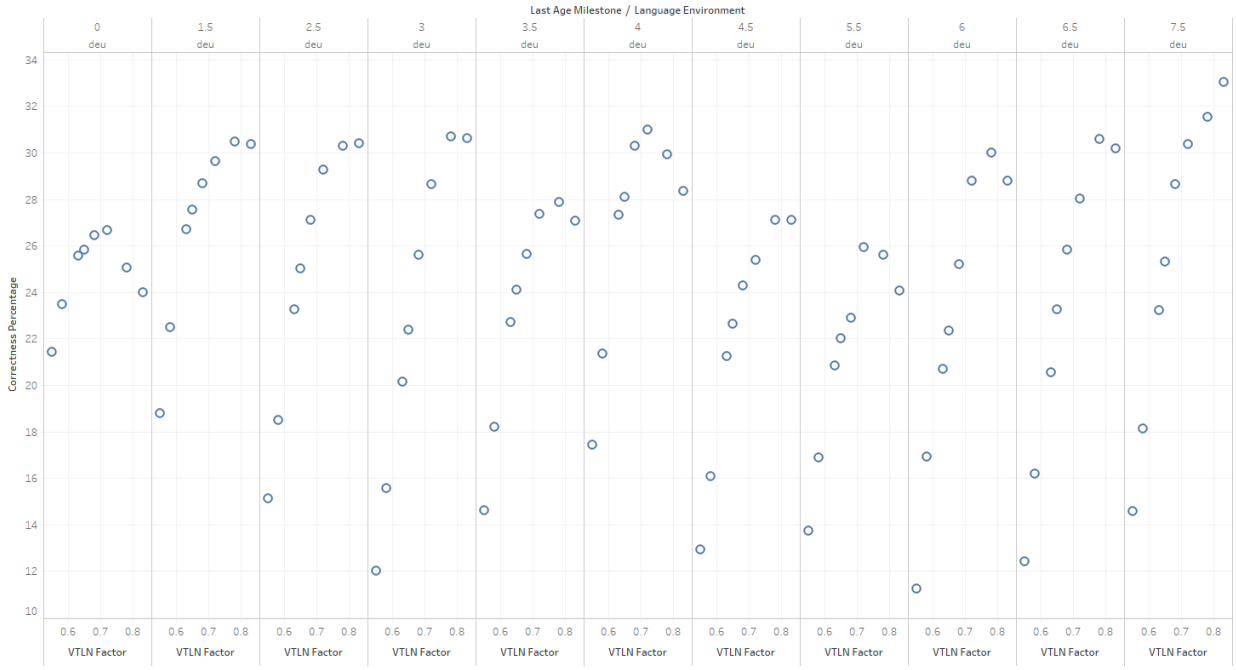

Figure 6: Visualization of results for the fourth set of our Phase 5 experiments, for the German environment, showing how correctness changes across values for VTLN Warp Factor and the lower-bound of Age At Recording, defined here as “Last Age Milestone”.

We can see that there are indeed optimal values of VTLN Factor for each age group, within each language-environment grouping.

10. Analysis of Results

In Phase 1, we attempted to use phonetic transcription data from our Talkbank dataset that was not time-aligned at the phone-level, to create a phone-recognizer without the use of the HInit and Hrest tools in HTK to properly initialize the model. The results were not impressive, with the correctness value failing to even cross the 15% mark. We should expect this kind of result because of the low-quality of the training data. This kind of outcome provides us with a motivation to find a way to time-align the Talkbank dataset at the phone-level in some way. This could be done manually by hand, by trained human transcribers, which would involve a lot of funding, coordination and management. It could also be done using a forced analyzer powered by a model that is more powerful than Model 28. This is a task that we decided was out-of-scope for our project. We do talk about it in the future research section of this thesis.

In Phase 2, we attempted to replicate the results of Lee & Hon 1989, creating a TIMIT-trained phone-recognizer that could perform on-par with their model, using an HMM-only approach. We believe we were able to come very close to accomplishing this, without the use of tied-state phones and with 2 extra phones in our phone set. The model that helped us accomplish this, Model 28, became the key tool for us to move forward with the next phases of our project.

We also explored the recognizability of our 41-phone TIMIT-oriented dataset, against the TIMIT testing data. We discovered a few phones such as [ʊ], [ɛ], [ə], [ɪ], [ʒ] having less than 60% correctness. This

gives us a clue that extra work will be required, perhaps through hybrid methods, to improve the recognizability of these phones. A lot of these phones seem to have been confused for other phones that share many articulation properties, which gives us a clue about what kind of disambiguation effort needs to happen, perhaps as a correction mechanism for a second-pass. We discuss this further in the next section on future research.

In Phase 3, we explored how Model 28 performed on the Talkbank dataset. We immediately see a significant improvement in the performance here, compared to what we saw in Phase 1, crossing the 30% mark for correctness. However, when we explored the performance on a per-language-environment basis, we discovered that the utterances from English-environment children had a higher correctness value than utterances from the French-environment and German-environment children. This showed us that a signficant English-bias is active, enough to account for a noticeable difference. This makes sense intuitively because the TIMIT corpus is in the English language. The phones and the phonology of the utterances are English-aligned. Children align quite early with the phonology of their language environment (Paul & Norbury, 2012), and perhaps the numbers we are seeing here are a reflection of this phenomenon.

Another interesting result we noticed in our Phase 3 work is that the correctness and accuracy values for the Talkbank testing dataset were very similar to those of the entire Talkbank dataset. This similarity followed through even when we cluster the data by language-environment. For each language environment, the subset of the data that belongs to the testing dataset had similar correctness and

accuracy to the entire dataset under that language-environment. This result gives us confidence in how representative the testing dataset we created really is. This test set truly does contains data that represents the entire dataset well, while only being a quarter of the size. This makes it a good choice for faster experiments in the future.

We noticed in these experiments is the much higher substitution and insertion rates for the English and French data, which was not downsampled to a 16k rate like the German data was. This probably explains why the German data has less substitution and insertion happening with a TIMIT-trained phone-recognizer, which was trained on TIMIT data that was also sampled at 16k. We tried to make a fairer comparison happen by redoing some of our experiments with downsampled English and French data. The results of these revised experiments were unexpected and puzzling, with insertion errors remaining high and correctness getting a boost as well. We feel that something about the downsampling process removed confounding factors towards some predictions. We also feel that some of the insertion errors might be a result of microphone quality differences, since many of the English and French environment recordings are older than the German ones by several years.

Another peculiar result is how adding more derivatives in the MFCC configuration leads to a higher insertion rate. We believe that this happens because the higher sample rates are able to capture more nuanced changes in higher-order derivatives, convincing a lower-sample-rate trained phone-recognizer that a brand-new phone must be identified, when this need not be the case in truth.

In Phase 4, we explored how Model 28 performed on the Torgo dataset, which contains speech utterances from patients who have dysarthria as well as from control subjects. The data involving the control subjects is not relevant to us. But we did see a correctness of 48.20% when it came to phone recognition on the Torgo patient dataset with Model 28. This is significantly lower than what we had for the TIMIT testing dataset but it is significantly higher than the best English-environment result we got from Phase 3. The acoustic impact of dysarthria is probably the influencing factor here, leading to variation in acoustic productions of the phones, impacting the results. If more data from more patients was available, we could create a Torgo-trained phone-recognizer to revisit this experiment, giving us more information on creating disability-aware phone recognition systems. Since the Torgo patients were all adults, so VTLN did not need to be carried out for these experiments. However, the Phase 3 results are from children with shorter vocal tracts, and discrepancy will impact the Phase 3 results negatively. A fairer comparison between the Torgo patients and the Talkbank children would occur using the results from Phase 5, which takes VTLN into consideration.

In Phase 5, we explored optimal values for the VTLN Warp Factor, that is used on the Talkbank data to prepare it for phone recognition using Model 28. We contextualized this exploration with respect to the Age At Recording experimental variable and the Language Environment experimental variable. This created multipliers that necessitated a high experiment count, which we achieved with the use of carefully constructed automation scripts. We ended up considering 612 scenarios (17 age-milestones vs.

9 warp factors vs. 4 language environment arrangements [all, eng, fra, deu]). Some of these scenarios had zero children, and therefore, no results. Figure 2 shows us that the English-environment children and the French-environment children have a primary cluster around the 1.5 years age milestone, with secondary clusters in nearby milestones as well. The German-environment children are more spread out across the age milestones, with most of their clusters being smaller in proportion and occuring between the 1.5 years and 4.5 years age milestones.

We know from Azizi et al., 2012 that a warp factor of 0.83 is a good value to use for HTK-based phone recognition for children who are 8 years of age (Azizi et al., 2012). We knew from the HTK Book (Young et al., 2006) that lower VTLN Warp Factors denote shorter vocal tracts. We know intuitively from phonetics first principles that younger children have shorter vocal tracts. We also know that having VTLN Warp Factors that are too low can lead to an increase in errors. Therefore, we expect to see some kind of parabolic graph when we plot correctness vs. VTLN factor, for each language environment and age grouping. Figures 1, 3, 4 and 5 show us visually where the ideal VTLN Warp Factor for every grouping lies. We can see intuitively that for children around the age of 1.5, a Warp Factor of 0.78 or 0.75 (or something in between) will usually give the best results. This remains relevant for a few more years of age. But as the children approach age 8, higher Warp Factors around the 0.83 mark end up bringing the best results, which is in-line with the results from Azizi et al., 2012.

The best result we get in Phase 5 is in Experiment 234, where we used Model 28 to perform phone recognition on the the English-environment Talkbank utterances for children whose Age At Recording

was 2.0 to 2.5 years, using a VTLN Warp Factor of 0.72. For 1905 phones that needed to be recognized, 56.8% of them were correctly recognized. This result is comparable to the correctness value of 48.2% obtained in Phase 4, when we used Model 28 to perform phone recognition on the Torgo patients. This comparison gives us a picture on how TIMIT-biased Model 28 really is. It goes from a peak performance of 72.03% on its own TIMIT data, down to a peak performance of 56.8% on a subset of a subset of the Talkbank data, down to a peak performance of 48.2% on the Torgo patients data. This motivates us to explore speaker adaptation methods to ensure that Model 28 is not being overfitted to the voices in the TIMIT corpus. We did not explore further speaker adaptation methods beyond VTLN in this project. We talk about this further in the further studies section below.

11. Threats to Validity and Opportunities for Further Study

The most obvious threat to validity we can see is in the fact that the Talkbank dataset is not time-aligned at the phonetic level, which means that we are not working with the best kind of data to accomplish machine-learning for phone recognition. We considered using forced-alignment to acquire time-aligned phone-level transcription but this requires a reliable HMM model to begin with. The only one we had was Model 28, which would not produce phonetic transcription symbols for all the 405 Talkbank phones, due to our folding mechanism. It would also try to match for symbols that aren't in the Talkbank dataset, namely the five diphthongs as well as [ɾ] and [ɝ]. Moreover, the correctness and accuracy values of Model 28 did not seem high enough to attempt a forced-alignment procedure on the entire dataset. This activity would be better suited to employing a model that has higher correctness and accuracy, such as the ones that use hybrid methods with neural networks or CRF.

Another consequence of the Talkbank-data not being time-aligned at the phonetic-level is that any recognizer attempting to make predictions will be likely to over-insert phones into the transcriptions where they don't belong, perhaps even repeating a correct recognition multiple times. One way to remedy this is to create a phone-segmentation system that identifies explicit time-boundaries for each phone. This can be a possible future task when it comes to enriching this dataset and making a more robust phone-recognizer, that attempts a phone-segmentation task before the actual phone recognition.

Another threat to validity occurs in the granularity and range of the VTLN Warp Factors we used in our study. With more experiments using more precise fine-grained values, we probably will discover more about how we can set optimal VTLN Warp Factor values by age.

For Phase 5, we should have made time to re-do all the experiments with downsampled speech data from the Davis and KernFrench corpora to 16k in order to make a fairer comparison with the Stuttgart and TAKI corpora which was downsampled already by 16k when we acquired it. We ran out of time to include these experiments.

The number of children in the English, French and German subsets of the Talkbank datasets were 17, 4 and 8 respectively. These are not a large enough amount of children to make statistically relevant observations involving vocal tract. Each child may have followed an individual growth pathway, leading to different vocal tract length milestones being reached at different times. It isn't wise to make decisions about an ideal VTLN Warp Factor by age, when so much variability exists in such a small dataset.

We never created a Torgo-trained phone-recognizer. This was because we didn't believe we had enough data from enough patients to make this viable. However, for future research, we should probably create this Torgo-trained phone-recognizer anyway and see how well it performs on recognizing phones produced by the Torgo patients, as well as other patients from other studies that may have speech and transcription data that is publicly available.

From a phone-recognizer commercialization point of view, the accuracy numbers on Model 28 are still not high enough to reliably make phone recognition with this model to a level of trustworthiness where clinical decisions can be made. We limited ourselves to creating models that were HMM-only, without the use of hybrid methods, without the use of tied-state phones, without the use of MAP speaker adaptation or MLLR speaker adaptation. Seeing that some of the low-correctness "problematic"phones in Experiment 28 were confused for phones that had similar articulation features, it opens the door for researching a two-pass system that uses binary classifiers trained using Support Vector Machines (SVMs) or some other well-known machine-learning algorithms for classification tasks, that kick in after the first HMM-based pass is complete, revisiting phone transcriptions involving the problematic phones, trying to match each of them against specific two-way classification systems versus the phone they are most confused for. This would be expected to boost the correctness numbers for these problematic phones. These advanced approaches are definitely avenues for further study in this topic, where better models than Model 28 can be prepared and brought in to explore how the story of phone recognition can be carried out more reliably for the children in our Talkbank dataset and the dysarthria patients in the Torgo dataset.

We do not have information on the race, culture, socioeconomic background, family size or dialect environment of the children or patients in our dataset. This is important information to keep in mind when we think of which phones get attention and how rich the learning-environment is for the speaker

when it comes to absorbing language stimuli. It is important to gather this information so we can study how things change across these variables, so we can be sensitive about our inferences and findings.

There currently is no publicly available dataset based on non-English speech utterances that has been prepared to the same level of richness as the TIMIT corpus. If a French and German edition of the TIMIT corpus existed and was publicly available to us, we could run experiments to avoid the impact of the English-bias that we discovered in Phase 3. Furthermore, we should not just focus on these three languages when it comes to making phone-recognizers. Further effort needs to be made to make similar speech corpora that have phonetic transcriptions that are time-aligned at the phonetic level, for many languages, not just the languages that are associated to first-world countries. We also need to consider creating language models that are specialized for specific multi-lingual environments, being trained on data from more than one language-environment. An exciting prospect would be choosing particular multilingual phone recognition models that are trained on data from the same monolingual and/or bilingual language environments that a child or patient has been exposed to. That level of customization and culturally-sensitive model-selection can possibly improve the resulting performance of the phone recognition task.

Knowing that the current state-of-the-art performance, as of December 2018, is still in the low-to-mid 80s when it comes to correctness percentage, we feel that future milestones in phone recognition research can also definitely open new research pathways from the work we have done here.

12. Conclusion

We conclude that the speech of pre-verbal children and the speech of patients with dysarthria are indeed quite recognizable and analyzable in a phone recognition context. The use of HMM-only phone-recognizers is not the most modern approach and some level of hybridization of the approach with other machine-learning paradigms will be required to make a system that is commercially viable for the purposes required by speech-language pathologists. Further effort would be required to introduce multiple speaker adaptation methods to ensure optimal performance. We also believe that more datasets should be collected from non-English language environments to create more diverse models rooted in more languages. With this modern complex globalized world we live in today, where humans migrate to new language environments and become multilingual, even pre-verbal children and people with speech disorders end up being exposed to multilingual language environments. Having culturally-sensitive phone recognition systems customized and re-trained from multilingual corpora on a per-patient basis ends up being an exciting direction where the commercialization of this technology can lead to very valuable improvements in the services that speech-language pathologists can provide to their patients.

References

Azizi, S., Towhidkhah, F., and Almasganj, F., "Study of VTLN method to recognize common speech disorders in speech therapy of Persian children," *2012 19th Iranian Conference of Biomedical Engineering (ICBME)*, Tehran, 2012, pp. 246-249. doi: 10.1109/ICBME.2012.6519690

Badino, L. (2016) “The ArtiPhon Task at Evalita 2016.” In: *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* 1749, 0-5

Bondy, A. S., & Frost, L. A. (1994). The Picture Exchange Communication System. *Focus on Autistic Behavior*, 9(3), 1–19. https://doi.org/10.1177/108835769400900301

Cosi, P. (2016) Phone Recognition Experiments on ArtiPhon with KALDI. In: *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian* 1749, 0-5

Davis B. L. & MacNeilage P. F. (1990). Acquisition of correct vowel production. A quantitative case study. *J Speech Lang Hear Res.* 1990;33:16–27.

Davis, Barbara L. & Peter F. MacNeilage (1995). The articulatory basis of babbling. *Journal of Speech and Hearing Research*, 38, 1199-1211.

Davis, Barbara L., Peter F. MacNeilage & Christine L. Matyear (2002). *Acquisition of serial complexity in speech production: A Comparison of Phonetic and Phonological Approaches to First Word Production. Phonetica*, 59, 75-107

Finch, E.; Rumbach, A. F.; Park, S. (2018) Speech pathology management of non-progressive dysarthria: a systematic review of the literature, *Disability and Rehabilitation*, 2018 Oct 4:1-11 DOI: 10.1080/09638288.2018.1497714

Fotuhi, M. & Yadegari, F. & Teymouri, R. (2016). Vowels Development in Babbling of typically developing 6-to-12-month old Persian-learning Infants. *Logopedics, phoniatrics, vocology*. 42. 1-8. 10.1080/14015439.2016.1221446.

Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; Pallett, D. S.; Dahlgren, N. L., Zue, V. (1993) *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Web Download. Philadelphia: Linguistic Data Consortium.

Goldsmith, J.A. (1990) *Autosegmental and metrical phonology*. Blackwell, Oxford.

Hoge, H.; Tropf, H. S.; Winski, R.; van den Heuvel, H.; Haeb-Umbach, R.; Choukri, K. (1997) European speech databases for telephone applications, *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, pp. 1771-1774 vol.3.

Hosom, J.-P., Kain, A.B., Mishra, T., van Santen, J.P.H., Fried-Oken, M., Staehely, J., (2003). Intelligibility of modifications to dysarthric speech. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP '03), vol. 1, pp. 924–927.

Kain, A.B., Hosom, J.-P., Niu, X., van Santen, J.P., Fried-Oken, M., Staehely, J., 2007 September. Improving the intelligibility of dysarthric speech. *Speech Communication* 49 (9), 743–759.

Kent, R.D. (1984), Psychobiology of speech development: co-emergence of language and a movement system. *Am. J. Physiol.* 246: 888–894.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, 54, 482-499.

Kent, R. D. (1990). The acoustic and physiologic characteristics of neurologically impaired speech movements. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kent, R. D., & Read, W. C. (1992). *The acoustic analysis of speech*. San Diego, CA: Singular Publishing Group.

Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., & Duffy, J. R. (1999). Acoustic studies of dysarthric speech: Methods, progress and potential. *Journal of Communication Disorders*, 32, 141-186.

Kern, Sophie, Barbara L. Davis, & Inge Zink (2009). From babbling to first words in four languages: Common trends, cross language and individual differences. In Francesco d'Errico & Jean-Marie Hombert (eds) *Becoming eloquent: Advances in the Emergence of language, human cognition and modern culture*. John Benjamins' Publishing Company.

Kern, Sophie & Barbara L. Davis (2009). Emergent complexity in early vocal acquisition: Cross-linguistic comparisons of canonical babbling. In F. Pellegrino, E. Marsico, I. Chitoran & C. Coupé (eds), *Approaches to phonological complexity*. Berlin, Mouton de Gruyter.

Köhler, J. (2001) Multilingual phone models for vocabulary-independent speech recognition tasks. *Speech Communication*, Vol. 35, Issues 1-2, Pages 21-30

Lamel, L. F.; Gauvain, J. (1993) Cross-lingual experiments with phone recognition. *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, MN, USA, pp. 507-510 vol.2.

Lamel, L. F.; Gauvain, J.; Eskenazi, M. (1991). *BREF, a Large Vocabulary Spoken Corpus for French.* EUROSPEECH 1991

Lee, K. & Hon, H. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.37 (11), November 1989, pp. 1642-1648, ISSN: 0096-3518.

Lindblom, B.,Krull, D., Stark, J. (1993) Phonetic systems and phonological development*; i*n Boysson-Bardies, de Schoen, Jusczyk, MacNeilage, Mortor, *Developmental neurocognition: speech and face processing in the first year of life.* pp. 399–409 (Kluwer, Dordrecht 1993).

Lintfert, Britta. (2009). *Phonetic and Phonological Development of Stress in German.* Universität Stuttgart Ph.D. Dissertation.

Lopes, C., & Perdigão. F. (2011). *Phoneme Recognition on the TIMIT Database*, Speech Technologies, Ivo Ipsic, IntechOpen, DOI: 10.5772/17600. Available from: https://www.intechopen.com/books/speech-technologies/phoneme-recognition-on-the-timit-database

Lohrenz, Timo; Li, Wei & Fingscheidt, Tim. (2018). A New TIMIT Benchmark for Context-Independent Phone Recognition Using Turbo Fusion. *Workshop On Spoken Language Technology, Athens, 2018.* DOI:10.13140/RG.2.2.36766.18243.

Ma, J.; Coltekin, C.; Hinrichs, E. (2016) Learning Phone Embeddings for Word Segmentation of Child-Directed Speech. *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, Berlin, Germany*, pages 53–63.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mohamed, A.; Sainath, T. N.; Dahl, G.; Ramabhadran B.; Hinton, G. E.; Picheny, M. A. (2011) Deep Belief Networks Using Discriminative Features for Phone Recognition. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5060-063.

Morris, J. & Fosler-Lussier, E. (2008). Conditional Random Fields for Integrating Local Discriminative Classifiers. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:3, March 2008, pp 617-628. , ISSN 1558-7916.

Oller D. K. (1980) The emergence of the sounds of speech in infancy. *Child Phonol.*;1:93–112.

Oller D. K.; Eilers R. E.; Neal A.R.; Schwartz H.K. (1999) Precursors to speech in infancy: the prediction of speech and language disorders. *J Commun Disord.* 1999;32:223–4

Paul, D. B., Baker, J. M. (1992) The design for the Wall Street Journal-based CSR corpus. *HLT '91 Proceedings of the workshop on Speech and Natural Language*, Pages 357-362

Paul, R. & Norbury, C. F., (2012) *Language Disorders from Infancy through Adolescence: Listening, Speaking, Reading, Writing, and Communicating. Fourth Edition.* Elsevier Mosby.

Pollak, P., and Behunek, M., (2011) "Accuracy of MP3 speech recognition under real-word conditions: Experimental study," *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, Seville 2011, pp. 1-6.

Potamianos, A.; Narayanan, S.; Lee S. (1997), Automatic speech recognition for children, in *Proc. Eurospeech, vol. 5, Rhodes, Greece*, Sept. 1997, pp. 2371–2374.

Potamianos, A. & Narayanan, S. (2003), Robust recognition of children's speech, in *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603-616, Nov. 2003.

Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L; Glembek, O.; Goel, N.; Hannemann, M.; Motlíček, P.; Qian, Y.; Schwarz, P.; Silovský, J.; Stemmer, G.; Vesel, K. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.*

Rose, R. & Momayyez, P. (2007). Integration Of Multiple Feature Sets For Reducing Ambiguity In ASR". *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007 (ICASSP), Hawaii, April 2007.

Rose, Y., & MacWhinney, B. (2014). The PhonBank Project: Data and software-assistend methods for the study of phonology and phonological development. In J. Durand, U. Gut & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 380-401). Oxford, UK: Oxford University Press.

Rodman, R., Moody, T., Price, J., (1985). Speech recognizer performance with dysarthric speakers: a comparison of two training procedures. *Speech Technology* 1, 65–71

Roy, N.; Leeper, H.A.; Blomgren, M.; Cameron, R. M. (2001) A description of phonetic, acoustic, and physiological changes associated with improved intelligibility in a speaker with spastic dysarthria. *American Journal of Speech-Language Pathology*, 10(3), 274–290.

Rudzicz, F., Namasivayam, A.K. & Wolff, T. (2012) The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Lang Resources & Evaluation 46: 523*. https://doi.org/10.1007/s10579-011-9145-0.

Rudzicz, F. (2013) Adjusting dysarthric speech signals to be more intelligible, *Computer Speech & Language*, Volume 27, Issue 6, Pages 1163-1177.

Russell, M.; Brown, B.; Skilling, A.; Series, R., Wallace, J.; Bonham, B.; Barker, P. (1996), Applications of automatic speech recognition to speech and language development in young children, in *Proc. ICSLP, Philadelphia, PA, Oct. 1996*

Saffran, J. R., Aslin, R. N., Newport, E. L. (1996) Statistical Learning by 8-Month-Old Infants. *Science*. 274(5294):1926-8.

Scanlon, P.; Ellis, D. & Reilly, R. (2007). Using Broad Phonetic Group Experts for Improved Speech Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, vol.15 (3) , pp 803-812, March 2007, ISSN 1558-7916.

Selby, J. C.; Robb, M. P.; Gilbert, H. R. (2000) Normal vowel articulations between 15 and 36 months of age, *Clinical Linguistics & Phonetics*, 14:4, 255-265

Sharma, H. V.; Hasegawa-Johnson, M. (2013) Acoustic model adaptation using in-domain background models for dysarthric speech recognition, *Computer Speech & Language*, Volume 27, Issue 6, Pages 1147-1162,

Siniscalchi, S. M.; Schwarz, P. & Lee, C.-H.; (2007). High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007 (ICASSP), Hawaii, April 2007.

Taussig, K., & Bernstein, J. (1994) *MACROPHONE: An American English Telephone Speech Corpus.* HLT.

Tolba, H.; Torgoman, A. S. E., (2009) Towards the improvement of automatic recognition of dysarthric speech. In: *International Conference on Computer Science and Information Technology, IEEE Computer Society*, Los Alamitos, CA, USA, pp. 277–281.

Teixeira, E. R. & David, B. L. (2002) Early Sound Patterns in the Speech of Two Brazilian Portuguese Speakers. *Language and Speech,* 45 (2) 179-204

Young, S. J. (1992). The general use of tying in phone-based HMM speech recognizers. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1992 (ICASSP), USA.

Vihman, M.M. (1996) *Phonological development: the origins of language in the infant.* Blackwell, Oxford.

Wilpon, J. G. & Jacobsen, C. N. (1996) A study of speech recognition for children and the elderly, IEEE *International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA,* 1996, pp. 349-352 vol. 1.

Young, S. J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V.,, Woodland, P. (2006). *The HTK Book Version 3.4*. Cambridge University Press.

Zlatić, L.; MacNeilage, P. F.; Matyear, C. L.; Davis, B. L. (1997) Babbling of twins in a bilingual environment. Applied Psycholinguistics. Vol. 8, Issue 4, pp. 453-469

Appendices

Appendix A: MFCC Configuration

Below are the MFCC configuration settings in HTK that we used for extracting MFCC features from each of the speech utterance .wav files we worked with for our study.

```
SOURCEFORMAT = <varies>
TARGETKIND = <varies>
TARGETRATE = 100000.0
SAVECOMPRESSED = F
SAVEWITHCRC = F
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
LOFREQ = 0
HIFREQ = 8000.0
WARPLCUTOFF = 300.0
WARPUCUTOFF = 5500.0
WARPFREQ = <varies>
```

When it came to the SOURCEFORMAT value, we set it to “WAV” for all Talkbank and Torgo data and we set it to “NIST” for all TIMIT data.

When it came to the TARGETKIND and WARPFREQ values, these were our main experimentation variables. Please read the section on our experiment descriptions to see what we set these values to be in our various experiments.

Appendix B: HTK Commands Used

Here are examples of the various HTK commands that we used to accomplish our goals. This information is provided in order to allow for reproducibility of our results. It can also be useful for people who want to learn how to use HTK.

HCopy

HCopy -T 1 -C <configFilePath> -S <scriptFilePath>

HCompV

HCompV -T 2 -f 0.01 -m -S <scriptFilePath> -L <pathToLabelFiles> -M
<pathToDirectoryforOutputHMMFile> -o <fileNameForOutputHMMFile>
<pathToPrototypeHMMFile>

HVite

HVite -A -D -T 2 -o SWT -b sil -C <pathToConfigFile> -H <pathToInputHMMGlobalsFile -H
<pathToHMMListFile> -t 250.0 150.0 1000.0 -y rec -i <nameOfOutputTranscriptionFile> -w
<pathToWordnetFile> -S <pathToMFCCScriptfile> <pathToDictionaryFile>
<pathToPhoneListFile>

HInit

HInit -i 100 -I <pathToMasterLabelFile> -S <pathToScriptFile> -H <pathToHMMGlobalsFile> -C
<pathToConfigFile> -T 1 -M <pathToOutputHMM> -l <phoneSymbol> <pathToPhoneHMM>

HRest

HRest -i 100 -l <phoneSymbol> -H <pathToInputHMMGlobalsFile> -I <pathToMasterLabelFile> -M
<pathToOutputHMMDirectory> -S <pathToScriptFile> <pathToPhoneHMM>

HERest

HERest -T 2 -C <pathToConfigFile> -I <pathToMasterLabelFile> -t 250.0 150.0 1000.0 -S <pathToScriptFile> -H <pathToInputHMMGlobalsFile> -H <pathToInputHMMList> -M <pathToOutputHMMDirectory> <pathToListOfPhones>

HResults

HResults -a UTTERANCE -b PHONE -p -L <pathToPhoneListFile> <pathToFileWithTranscriptions> > <pathToOutputResultsFile>

Appendix C: HMM Prototype

The following material represents an example of how we prototyped our HTK HMMs for every experiment. The example below represents an HMM with 5 states, 3 of them emitting, that feed into each other in sequence, with each state represented with a single-Gaussian function and 12 MFCC features. All other prototypes we used in our experiments were modified from this form. VecSize changes with number of coefficients and the lengths of the mean and variance vectors change accordingly as well, with extra padded dummy 0.0 and 1.0 values to set us up for training. For seven-state HMMs, the NumStates entry changes, two more State entries are provided and the TransP matrix becomes 7x7.

```
~o <VecSize> 12 <MFCC>
<BeginHMM>
<NumStates> 5
<State> 2
<Mean> 12
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 12
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<State> 3
<Mean> 12
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 12
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<State> 4
<Mean> 12
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 12
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<TransP> 5
0.0 1.0 0.0 0.0 0.0
0.0 0.6 0.4 0.0 0.0
0.0 0.0 0.6 0.4 0.0
0.0 0.0 0.0 0.7 0.3
0.0 0.0 0.0 0.0 0.0
<EndHMM>
```

Appendix D: HTK Grammars

Grammar for performing phone recognition in Pipeline A, with training on the Talkbank dataset.

```
$phone = a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | r | s | t | u | v | w | x | y | z | æ | ç | ð | ø | ŋ
| œ | ɐ | ɑ | ɔ | ə | ɚ | ɛ | ɣ | ɪ | ɸ | ɹ | ʁ | ʁ | ʃ | ʉ | ʊ | ʌ | ʏ | ʒ | ʔ | ʙ | ʤ | ʧ | β | θ | other;
( sil <$phone> sil )
```

Grammar for performing phone recognition in Pipeline B, with training on the TIMIT dataset.

```
$phone = b | d | g | p | t | k | ɾ | ʤ | ʧ | s | ʃ | z | ʒ | f | θ | v | ð | m | n | ŋ | l | ɹ | w | j | h | sil | i | ɪ | ɛ | e_ɪ | æ | ɑ | a_ʊ | a_ɪ | ɔ | ɔ_ɪ | o_ʊ | ʊ | u | ɝ | ə;
( <$phone> )
```

Appendix E: Pronunciation dictionaries

For Talkbank data in Pipeline A	**For TIMIT data in Pipeline B**
sil sil	b b
other other	d d
a a	g g
b b	p p
c c	t t
d d	k k
e e	ɾ ɾ
f f	ʤ ʤ
g g	ʧ ʧ
h h	s s
i i	ʃ ʃ
j j	z z
k k	ʒ ʒ
l l	f f
m m	θ θ
n n	v v
o o	ð ð
p p	m m
r r	n n

s s	ŋ ŋ
t t	l l
u u	ɹ ɹ
v v	w w
w w	j j
x x	h h
y y	sil sil
z z	i i
æ æ	ɪ ɪ
ç ç	ɛ ɛ
ð ð	e_ɪ e_ɪ
ø ø	æ æ
ŋ ŋ	ɑ ɑ
œ œ	a_ʊ a_ʊ
ɐ ɐ	a_ɪ a_ɪ
ɑ ɑ	ɔ ɔ
ɔ ɔ	ɔ_ɪ ɔ_ɪ
ə ə	o_ʊ o_ʊ
ɚ ɚ	ʊ ʊ
ɛ ɛ	u u
ɣ ɣ	ɜ ɜ
ɪ ɪ	ə ə
ɸ ɸ	
ɹ ɹ	
ʀ ʀ	
ʁ ʁ	
ʃ ʃ	
ʉ ʉ	
ʊ ʊ	
ʌ ʌ	
ʏ ʏ	
ʒ ʒ	
ʔ ʔ	
ʙ ʙ	
ʤ ʤ	
ʧ ʧ	

β β θ θ	