LINGUISTIC ANALYSIS OF SPONTANEOUS CHILDREN SPEECH IN INTERACTION WITH COMPUTER

By
Vassiliki I. Farantouri
Technical University of Crete
Chania
October 2008

TECHNICAL UNIVERSITY OF CRETE
DEPARTMENT OF
ELECTRONICS AND COMPUTER ENGINEERING

Dated: October 2008

Supervisor:

Assoc. Prof. Alexandros Potamianos

Readers:

Prof. Vasilis Digalakis

Assist. Prof. Ekaterini Mania

To my beloved parents

Στους αγαπημένους μου γονείς

Table of Contents

List of Tables

List of Figures

Abstract

Spoken dialogue systems are widely used both by adults and by children. Dialogue systems for children are used in educational and entertainment applications, as well as in health care for the diagnosis and therapy of speech-related disfluencies. However speech-technology resources designed for the adult population are not directly usable for children users since linguistic and acoustic characteristics of spontaneous children's speech differ significantly from those of adults'. Actually children's speech characteristics exhibit great variance even among ages and genders.

In this thesis, we investigated the duration, lexical and linguistic properties of children's spontaneous speech for children ages 8 to 14 interacting with animated characters in a computer game. The corpus used here was collected during CHIMP Wizard of Oz experiment based on a spoken dialogue system with multimodal input. Out data was organized in 3 age groups in order to preserve statistical significance. Age and gender trends are studied for fluency, lexical and syntactical parameters, such as number of mispronounciations and hesitations, sentence duration, phone duration, speaking rate, vocabulary size, number of words per utterance and linguistic variability measured via bigram language model perplexity. Statistical significance of the results is tested using 2-way ANOVA models. The analysis shows significant differences between read- and spontaneous children speech in terms of absolute values of acoustic and linguistic parameters, as well as linguistic variability. Variations in children linguistic characteristics according to age are also spotted. In addition, spontaneous data present clear gender-specific trends, e.g., increased 'language exploration' by girls in the 12-14 age group and a clear difference between males and females, as far as language use is concerned.

Finally, the meaning of the results acquired and the applicability of these results for acoustic and linguistic modeling and spoken dialogue systems interface design is discussed.

Περίληψη

Τα διαλογικά συστήματα χρησιμοποιούνται σήμερα ευρέως όχι μόνο από ενήλικες αλλά και από παιδιά. Στην περίπτωση των παιδιών τα συστήματα αυτά εφαρμόζονται τόσο για εκπαιδευτικούς όσο και για ψυχαγωγικούς σκοπούς, καθώς και στην ιατρική για διάγνωση και θεραπεία διαταραχών που σχετίζονται με την ομιλία και την γλώσσα. Παρολαυτά οι εφαρμογές φωνής που έχουν σχεδιαστεί για ενήλικες δεν μπορούν να χρησιμοποιηθούν άμεσα και για παιδιά, καθώς τα γλωσσολογικά και φωνητικά χαρακτηριστικά των παιδιών διαφέρουν σημαντικά από εκείνα των ενηλίκων. Μάλιστα παρατηρείται μεγάλη διαφοροποίηση των χαρακτηριστικών αυτών σε παιδιά διαφορετικών ηλικιών και φύλου.

Στην εργασία αυτή μελετήθηκαν και αναλύθηκαν γλωσσολογικά, λεκτικά και ακουστικά χαρακτηριστικά του αυθόρμητου παιδικού λόγου παιδιών ηλικίας 8-14 χρονών, κατά τη διάρκεια διάδρασης με κινούμενους χαρακτήρες παιχνιδιών. Τα δεδομένα που χρησιμοποιήθηκαν είχαν συλλεχθεί στα πλαίσια του Project CHIMP, ενός Wizard of Oz πειράματος που περιελάμβανε την χρήση ενός πολυτροπικού διαλογικού συστήματος. Τα δεδομένα μας οργανώθηκαν σε 3 ηλικιακές ομάδες, προκειμένου να διασφαλίσουμε την στατιστική σημασία των αποτελεσμάτων μας. Η διάρκεια των φωνημάτων και των προτάσεων, ο ρυθμός ομιλίας, η ευχέρεια λόγου, το μέγεθος του λεξιλογίου και η γλωσσική πολυπλοκότητα μέσω bigram γλωσσικών μοντέλων μελετήθηκαν σε συνάρτηση με την ηλικία και το φύλο. Έγινε επίσης σύγκριση μεταξύ του αυθόρμητου παιδικού λόγου και του παιδικού λόγου κατά την ανάγνωση . Τα αποτελέσματα της μελέτης αυτής ελέγχθηκαν ως προς την στατιστική τους σημασία μέσω 2-way ANOVA μοντέλων. Η ανάλυσή μας καταδεικνύει σημαντικές διαφορές μεταξύ του αυθόρμητου παιδικού λόγου και του λόγου παιδιών που διαβάζουν ένα κείμενο. Παρατηρείται επίσης έντονη διαφοροποίηση των γλωσσολογικών χαρακτηριστικών των παιδιών από ηλικία σε ηλικία, όπως επίσης και μεταξύ αγοριών-κοριτσιών. Για παράδειγμα τα κορίτσια ηλικίας 12-14 χρονών φαίνονται από την ανάλυσή μας να εξερευνούν πολύ περισσότερο την γλώσσα και τις δομές της σε σύγκριση με τα αντίστοιχης ηλικίας αγόρια, τα οποία φαίνονται να χρησιμοποιούν την γλώσσα κυρίως ως εργαλείο για να ολοκληρώσουν γρηγορότερα το παιχνίδι και όχι ως μέρος του παιχνιδιού.

Τα συμπεράσματα της ανάλυσης αυτής μελετήθηκαν επίσης ως προς την σημασία τους για τον σχεδιασμό ακουστικών και γλωσσικών μοντέλων για παιδιά καθώς και για το σχεδιασμό εξειδικευμένων διαλογικών συστημάτων .

Acknowledgements

I would like to express my gratitude to my advisor Assoc. Prof. Alexandros Potamianos for his invaluable guidance, encouragement and support and also to thank him for assigning me this particular subject. I am moreover grateful for all his help and guidance in writing the conference paper related to this analysis.

I would also like to thank professors Vasilis Digalakis and Aikaterini Mania for participating in the examining committee.

Furthermore, I would like to thank my friends and colleagues in the Telecommunications Laboratory for their comments and hints, their help and most of all for their support during the long days, nights, weekdays and weekends we spent together in the lab or elsewhere. Thank you Elia, Manoli, Orfea, Pavlo, Despina, Kelly.

Special thanks to my old friends who were by my side through all the years I have spent in Chania, and they were always there for me even if they were physically away. Eleni,Maria, Foti, Mantho, Konstantine, a big 'thank you' to all of you.

Last but certainly not least, I would like to thank my parents, for their patience, their encouragement and their unlimited and continuous support. I am incredibly lucky to have you. Mom, Dad, this thesis is dedicated to you.

Chapter 1

Spoken Dialogue Systems and Children Speech

1.1 Introduction

Spoken dialogue systems are widely used nowadays both by adults and children. With the remarkable improvements in computer performance and speech recognition, spoken dialogue systems are applied to various areas, such as voice portal services, car navigation systems, and input for controlling PC operations. These applications assume that the users are mainly adults and the adult speech databases are easily available. However creating spoken dialogue systems specifically aimed at children is of great importance since children form a crucial segment of customer population for interactive multimedia systems and they are also *eager and quick to embrace and use new technologies*. Moreover speech-technology resources designed for the adult population are not directly usable for children users since linguistic and acoustic characteristics of spontaneous children's speech differ significantly from those of adults'. Actually children's speech characteristics exhibit great variance even among ages and genders, while there are also social factors that interfere. Moreover human-human dialog and child-computer interaction exhibit different linguistic and acoustic characteristics. Other factors also interfere, such as the emotional state of the child [57], the nature of the interaction (task to be accomplished), etc. Lastly, spontaneous children speech and 'read' children speech also differ significantly.

1.2 Spoken Dialogue Systems for Children

The ease of children in adopting technology [39] has stimulated and boosted interest. The improvement of human-machine interaction (HMI) for young children has become an issue of great importance as youth have grown more comfortable with using new technologies. Recently, children's speech has been gaining growing attention from the research community and in industry. The potential contributions of automatic interactive systems for children are tremendous, especially in areas such as education and entertainment. The three main speech related fields where research on children speech has acquired considerable momentum are *health care* (aids for diagnostic and therapy),

edutainment (aids for pronunciation, understanding), and *entertainment* (computer games). Children are not just another group of users. Speech and multimedia technologies are still far from perfection for adults and children's speech poses even more compelling questions to solve. Differences in cognitive development and application-specific domain mismatches demand that children are examined separately.

Early spoken dialogue application prototypes that were specifically aimed at children included word games for pre-schoolers [50], aids for reading [36] and pronunciation tutoring [48]. Recently a number of systems have been implemented with advanced spoken dialogue interfaces, multimodal interaction capabilities and/or embodied conversational characters [39, 26, 5, 6]. Data collected from these systems as well as new available corpora [3, 56, 4] have improved our understanding of verbal child-machine interaction.

One of the reasons why research on children's speech has not been as extensive as that on adults' is that it is more difficult to collect the children's speech. Most of the databases of children recordings focus on the 6-18 age group (or a subset thereof) where collection conditions can be more easily controlled and the subjects are collaborating. Examples of corpora ('read speech') that is mostly used for acoustic analysis and modeling are the American English CID children corpus [33], the KIDS corpus [12], the CU Kids' Audio Speech Corpus [26] and the PF-STAR corpus available in the following languages: British English, Italian, German and Swedish [3]. These corpora consist of prompted speech and *monologues* where children recount stories.

As far as child-machine spontaneous speech interaction (*dialogue* data) is concerned a handful of corpora has been recently collected and analyzed. In [5], the NICE fairy-tale corpus is presented, where children use open-ended spoken dialogue to interact with animated characters in a game setting. In [4], a child-robot interaction corpus is presented; children interacted with an AIBO robot is open-ended scenarios. However, since the AIBO did not answer back, the children's utterances mostly consisted of short commands and little dialogue interaction took place.

As far as interaction with computer animated characters is concerned, in [5] a high degree of social involvement of the children with the characters was observed. In CHIMP, it was found that using animated sequences to communicate information and adding 'personality' to the interface significantly improved the user experience.

1.3 Multimodal Dialogue Systems

A limiting feature of modern interfaces that has also become increasingly evident is their reliance on a single mode of interaction-a mouse movement, key press, speech input, or hand motion. Even though it may be adequate in many cases, the use of a single interaction mode proves to be inadequate in HCI.

Recently, there has also been increasing interest in the design of *multimodal* interfaces that combine speech with a variety of other input modalities such as text, touch, mouse clicks, handwriting, and gestures [49]. Results of these investigations suggest that the use of multiple modalities, rather than a single modality, leads to more efficient and natural interaction and enhances the overall user experience (for example, [9]). Multimodality is attractive in the creation of conversational interfaces for children

in the sense of both overcoming inherent limitations in speech technology and exploiting the ubiquitous availability and/or familiarity with conventional modalities such as the computer mouse, keyboard, joystick and pen.

There are numerous potential benefits in integrating multiple modalities into HCI. As stated and analysed in [49], the reasons range from the fact that natural human interaction itself has a multimodal character to the statistical advantages of combining multiple observations.

- Practical Reasons
  Some inherent drawbacks of current advanced single-modality HCI systems undermine their effectiveness and call for multimodal HCI. Single-modality HCI lacks robustness and accuracy. However, concurrent use of two or more interaction modalities may loosen the strict restrictions needed for accurate and robust interaction with the individual modes and can help reduce the complexity and increase the naturalness of the interface for HCI.

- Biological Reasons
  Almost any natural communication among humans involves multiple, concurrent modes of communication. Thus, any HCI system that tries to have the same naturalness should be multimodal. Indeed, studies have shown that people prefer to interact multimodally with computers, since among other things, such interaction eases the need for specialized training.

- Mathematical Reasons
  The disadvantage of using a single modal system is that it may not be able adequately to reduce the uncertainty for decision making. Uncertainty arises for example when features are missing or when observations are ambiguous. On the other hand, it is well known that it is statistically advantageous to combine multiple observations from the same source because improved estimates are obtained using redundant observations.

In [39], a corpus was collected in a Wizard-of-Oz, scenario where children used speech to play an interactive computer game using voice commands or keyboard and mouse and interact with animated characters on screen. The CHildren's Interactive Multimedia Project (acronym: CHIMP) aimed at providing essential guidelines for engineering successful *multimodal-input multimodal-output* applications for children with an emphasis on the spoken dialog interface. The resulting corpus (also used in our analysis) was used to create novel language models and understanding strategies for dialogue systems aimed towards young users. The authors found that user experience was improved by adding 'personality' to the interface, allowing for *multimodal* interaction and using animated sequences to convey information. Examining the dialog strategies of the children, the belief that, although speech might not be the most efficient modality always, it is a more natural modality, was reinforced. This agrees with the observations in the NICE project [5], where most users reported that it was quite natural to use speech in games and many expected that games will be like this in the future. In fact, in the CHIMP project, it was found that the children tended to switch modalities from voice to mouse clicks either when there was repeated ASR errors or when there was

a need for dialogue disambiguation. In addition, the flexible choice of input modality (any of speech, natural language, commands or buttons) made the application easy to use even for novice users. Children enjoyed interacting with the computer using voice but also preferred combining interface modalities.

Moreover, in [56], a corpus of child-machine interaction via a multimodal voice and pen interface was collected and analyzed.

1.4 Children's Speech Specifics

Recent research using both naturalistic and experimental methods has found that the vast majority of young children's early language is organized around concrete, item based linguistic schemas. From this beginning, children then construct more abstract and adult-like linguistic constructions, but only gradually. Children imitatively learn concrete linguistic expressions from the language they hear around them, and then, using their general cognitive and social-cognitive skills, categorize, schematize and creatively combine these individually learned expressions and structures [52].

Both acoustic and linguistic characteristics of children's speech differ from those of adults: pitch, volume, formant positions, and co-articulations vary strongly due to anatomical and physiological development [33],[21]; the linguistic structure of the children's utterance is not too uniform, and lapses, short or not well-constructed sentences, repetitions, and disfluencies are generally frequent, mainly depending on age and socio-economic factors [33].

The acoustic characteristics of children for 'read speech' have been first analyzed in [11, 30] and later on in [33] for American English. Recently such studies have been carried out for other languages as well, e.g., Italian [21]. In all studies, children demonstrate larger fundamental and formant frequency, as well as, *higher acoustic variability*. In general, it is considered that variability converges to adult values around 13-14 years of age [33]. A detailed comparison of temporal features and speech segment durations for children vs adults (for 'read speech') can be found in [31, 33]. Again, distinct age-related differences were found. On average, the speaking rate of children is slower than that of adults. Further, children speakers display higher variability in speaking rate, vocal effort, and degree of spontaneity.

In [13] and [14],detailed analysis of the way different kinds of pausing strategies, such as empty and filled pauses, and phoneme lengthening are used by children to shape the discourse structure in spontaneous speech (narrations) are presented. Spontaneous speech, as well as other types of speech, is characterized by the presence of silent intervals (empty pauses) and vocalizations (filled pauses) that do not have a lexical meaning. These pausing means play several communicative functions and their occurrence is determined by several factors such as build up tension, signal anxiety, emphasis, syntactic complexity, degree of spontaneity, gender, and educational and socio-economical information. Cognitive psychologists suggest that pausing strategies reflect the complexity of neural information processing. Pauses will surface in the speech stream as the end product of a 'planning' process that cannot be carried out during speech articulation and the amount and length of pausing reflects the cognitive effort related to lexical choices and semantic difficulties for generating new information. However, pauses are

not only generated by psychological motivations but also as a linguistic mean for discourse segmentation. The reported data showed that children pause, like adults, to recover from their memory the new information they try to convey. Higher is the recovery effort, longer is the pausing time. As a linguistic mean for discourse segmentation, pauses are used by children to mark words, clause, and paragraph boundaries.

Phonological characteristics of *imitative* (repeat-after-me) and *spontaneous* children speech are compared in [34].The findings of this study suggest that the relationship between the phonological characteristics of children's imitative and spontaneous speech is not static, but varies as children proceed through the course of acquiring their language. Initially, children seem to learn words and, more incidentally, corresponding articulations. Such words are not strictly phonemically principled and variability may be seen in a child's production of different words with the same sounds. Imitations occurring at this time appear to reflect this same variability. That is, the productions of words imitated may not be in close correspondence with the child's spontaneous productions of other words with the same target sounds.

In an attempt to explain the differences between female and male children's speech [32] introduces two factors: the frequency hypothesis and the role-model hypothesis. Children of different gender speak differently either because they are usually spoken to differently or because they model on their same-sex parent or same-sex peers. The findings of this analysis support mainly the second hypothesis. Moreover in [32], it is argued that social factors like early instutionalisation of children lead to increased peer group influence and help explain why gender differences occur at an earlier age among children of different cultures.

There is no detailed analysis in the literature of the acoustic and linguistic characteristics of spontaneous children's speech due to the lack of large corpora. However, there are limited studies of child-machine spontaneous speech interaction using smaller corpora. In [5], significant differences in the duration and language usage where found in child-machine dialogue compared to human-human dialogue. Specifically children ages 8-15 communicated with fairy-tale characters in a computer game scenario, using shorter utterances, slower speaking rate and much less filled pauses, filler words and phrases, compared to human-human dialogue. In [2], politeness and frustration markers were analyzed for the CHIMP database (the database also analyzed in this thesis). Younger children used politeness markers more commonly and expressed frustration verbally more often than older children. In [56], the multimodal integration patterns of children ages 7-10 were investigated for a speech and pen interface. It was found that the modality usage was similar between children and adults, although children tend to use both input modes simultaneously rather than sequentially.

In this study, we analysed the linguistic and acoustic characteristics of spontaneous children speech while interacting with a computer animated character in a WoZ experiment (CHIMP). Duration, fluency and lexical statistics were acquired as well as linguistic variabilities. A comparison with previous acquired results for read speech was also performed. The results of the this analysis lead to interesting conclusions regarding children spontaneous speech's characteristics.

The rest of the thesis is organized as follows:

In Chapter 2 basic elements of Language Modeling theory and n-gram statistical models are presented.
In Chapter 3 Acoustic Models and their underlying characteristics are described.
In Chapter 4 we describe the CHIMP game scenario for collecting data and the corpus used in our analysis
In Chapter 5 our method for calculating the results is presented, as well as the tools we used.
In Chapter 6 the most representative results are shown and discussed.
Finally, in Chapter 7 general conclusions and future research directions are discussed.

Chapter 2

Statistical Language Modeling

2.1 Introduction

The notion of 'word' in linguistics is denoted by the term 'lexeme', the minimal unit of language, which has one or more semantic interpretations. A word exists with other words and these units build more larger comprehensive units: phrases, sentences, paragraphs, etc. It is reasonable to claim that a word preserves a kind of conceptual relationship with its neighboring words, in some way. From this point of view each word contains an amount of information about the other words of its lexical environment. We can say that the occurrence of a word is dependent to the context words . Thus, it is possible to utilize the surface statistics of language in order to proceed to a deeper level.

Modeling a word sequence can be useful for various reasons, for example for predicting the next word in a sequence. The prediction becomes applicable when it can be measured. By treating words as events and distributing to them a probability mass we can create a probabilistic language model that will be able not only to predict the next word in a sequence but also to estimate the probability even for a completely unknown word.

The simplest language probabilistic model lets any word to follow any other word with equal probability. A more complex language model uses the frequency of occurrence of a word. For example, consider a paragraph that has totally 100 words, in which the words 'example' and 'a' occur 5 and 12 times, respectively. According to the simple language model the words 'example' and 'but' have $\frac{5}{100}$ and $\frac{12}{100}$ probability, respectively, to follow any word. But we also have to consider the following: In a given phrase, for example in the phrase 'this is a nice', not all of the words of the vocabulary have the same probability to follow the word 'nice'. For example the word 'example' is more reasonable than 'a' to follow the word 'nice'. This observation shows that we have to consider the conditional probability of a word given the previous word, instead of using the relative word frequency.

2.2 Language modeling

In few words, a language model gives the probability $P(s)$ of a sentence s. Let S be a word sequence. In general, statistical language modeling estimates $P(S)$.

The majority of the language models decomposes the sentence probability, $P(s)$, into a product of conditional probabilities

$$P(s) = P(w_1 \dots w_n) = \prod_{i=1}^{N} P(w_i|h_i) \tag{2.2.1}$$

where w_i is the i^{th} word in the sentence and $h_i = \{w_1, w_2, \dots, w_{i-1}\}$ is the sequence of preceding words (the *history* of w_i).

2.3 N-gram language modeling

An n-gram is a sequence of n symbols (e.g. words, syntactic categories, etc) and an n-gram language model (LM) is used to predict each symbol in the sequence given its $n-1$ predecessors. It is built on the assumption that the probability of a specific n-gram occurring in some unknown test text can be estimated from the frequency of its occurrence in some given training text. Thus, as illustrated by the picture above, n-gram construction is a three stage process. Firstly, the training text is scanned and its n-grams are counted and stored in a database of *gram* files. In the second stage some words may be mapped to an out of vocabulary class or other class mapping may be applied, and then in the final stage the counts in the resulting gram files are used to compute n-gram probabalities which are stored in the *language model* file. Lastly, the *goodness* of a language model can be estimated by using it to compute a measure called *perplexity* on a previously unseen test set. In general, the better a language model then the lower its test-set perplexity.

Although the basic principle of an n-gram LM is very simple, in practice there are usually many more potential n-grams than can ever be collected in a training text in sufficient numbers to yield robust frequency estimates. Furthermore, for any real application such as speech recognition, the use of an essentially static and finite training text makes it difficult to generate a single LM which is well-matched to varying test material. For example, an LM trained on newspaper text would be a good predictor for dictating news reports but the same LM would be a poor predictor for personal letters or a spoken interface to a flight reservation system. A final difficulty is that the *vocabulary* of an n-gram LM is finite and fixed at construction time. Thus, if the LM is word-based, it can only predict words within its vocabulary and furthermore new words cannot be added without rebuilding the LM.

2.3.1 General

The N-gram language model considers the language as a Markov process of order $N-1$.

$$P(w_i|h_i) = P(w_i \mid w_{i-N+1}, \dots, w_{i-1}) \approx P(w_i \mid w_{i-N+1}^{i-1}) \tag{2.3.1}$$

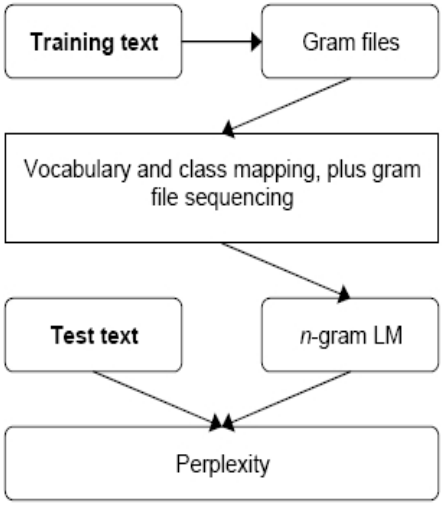


Figure 2.1: Language Modeling Diagram.

Equation 2.3.1 states that the probability of word w_i given all the previous words of the sentence can be approximated by the probability given only the previous $N-1$ words.

N-gram probabilities are computed by counting and normalizing the N-gram occurrences. For the bigram case the conditional probability of word w_{i-1} given that it is followed by word w_i is computed as

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{\sum_w C(w_{i-1}w)} = \frac{C(w_{i-1}w_i)}{C(w_{i-1})} \tag{2.3.2}$$

Equation 2.3.2 takes the *count* C of $w_{i-1}w_i$ bigram and divides it by the sum of all bigrams that have w_{i-1} as first word. Note that the latter sum is equal to the count of w_{i-1} unigram. For the general case of N-gram model the above equation is written as

$$P(w_i|w_{i-N+1}^{i-1}) = \frac{C(w_{i-N+1}^{i-1}w_i)}{C(w_{i-N+1}^{i-1})} \tag{2.3.3}$$

Equations 2.3.2 and 2.3.3 use the frequency interpretation of probability [1], applying the technique of Maximum Likelihood Estimation (*MLE*). Even with large corpora many N-grams occur only once or they have low counts, so the computation of N-gram probabilities remains a sparse estimation problem. Thus, it is preferable not to apply *MLE* of N-gram probabilities in a straightforward way, based on counts. Instead, several smoothing approaches [20] can be used in order to smooth the *ML* estimates.

2.3.2 Smoothing

The N-gram models are trained from corpora. In practice, every training corpus is of finite size, so, naturally some acceptable N-grams are bound to be absent. This intrinsic characteristic of corpora leads to zero and low counts of N-grams. Using the

MLE approach the absent N-grams are assigned zero probability, while the probabilities of low-count N-grams are underestimated.

Consider the sentence 'put language back into language modeling' If the bigram 'put language' has never occurred in the training corpus, then

$$P(language \mid put) = \frac{C(put\ language)}{\sum_w C(put\ w)} = \frac{0}{a}, \qquad a > 0 \tag{2.3.4}$$

The probability of sentence $P(put\ language \ldots modeling) = 0$. Clearly, this is an underestimate for the sentence probability, since in real life there is an 'amount' of probability by which the sentence is likely to occur.

Smoothing battle the problem of data sparseness by re-evaluating the zero- and low-probabilities and assigning them non-zero values. The name of this strategy describes what is actually happens. Smoothing techniques make the probability distributions more uniform: adjust low probabilities upward and high probabilities downward [20]. Next, we briefly survey some of the most widely-used smoothing strategies in order to outline the underlying ideas.

Additive smoothing

This is a simplistic technique of smoothing, since it pretends that an N-gram occurs δ times more than it does, where $0 < \delta \leq 1$ [22, 55, 24]. For example, for the bigram case we have [20]

$$P_{Add}(w_i \mid w_{i-1}) = \frac{\delta + C(w_{i-1}\ w_i)}{\delta \mid V \mid + \sum_{w_i} C(w_{i-1}\ w_i)} \tag{2.3.5}$$

where set V is the vocabulary of the training corpus and $|V|$ denotes the cardinality of V. In general, the additive smoothing has poor performance [53, 54].

Good-Turing estimate

The key idea of Good-Turing smoothing is the exploration of N-grams of high counts in order to re-estimate the amount of probability mass that is to be given to N-grams with zero or low counts [28]. The Good-Turing estimate feigns that for any N-gram that occurs r times we can feign that it occurs $r^\star$ times:

$$r^\star = (r+1)\frac{k_{r+1}}{k_r} \tag{2.3.6}$$

where k_{r+1} and k_r is the number of N-grams that occur exactly $r+1$ and r times, respectively. For instance, if a bigram occurs r times, the corresponding probability is

$$P_{GT}(w_i \mid w_{i-1}) = \frac{r^\star}{\sum_\infty^{r=1} r k_r} \tag{2.3.7}$$

In particular, the Good-Turing estimate is applied as stand alone N-gram smoothing approach, because does not combine high- and low-order models that obtain better performance [20].

Deleted interpolation

It is fruitful to interpolate higher-order N-gram models with lower-order N-gram models, because there are cases whereas there is no sufficient data to compute probabilities for the higher-order models [16]. So, the lower-order models is more trustworthy, providing supplementary useful information.

$$P_{DelInt}(w_i|w_{i-N+1}^{i-1}) = \lambda_1(w_{i-N+1}^{i-1})P_{ML}(w_i|w_{i-N+1}^{i-1}) + \lambda_2(w_{i-N+1}^{i-1})P_{DelInt}(w_i|w_{i-N+2}^{i-1}) \tag{2.3.8}$$

The smoothed model of Equation 2.3.8 use recursion as interpolates linearly an N^{th}-order model estimated with maximum likelihood and an N^{th-1}-order smoothed model [18]. Note that the λ weights sum to 1:

$$\sum_i \lambda_i = 1 \tag{2.3.9}$$

Each λ value is a function of the context. The optimal $\lambda_(w_{i-N+1}^{i-1})$ is different for different histories. For example, a context that have been occurred for many times should be given a high weight since its' distribution tends to be reliable. In contrast, for a history of low frequency, a lower λ weight will be reasonable. For the training of the λ parameters many approaches have been proposed in the literature [16, 47, 19, 20].

Witten-Bell smoothing

The Witten-Bell smoothing can be considered as an instance of deleted interpolation [27].

As Equation 2.3.8, the N^{th}-order maximum likelihood model is linearly interpolated with the N^{th-1}-order smoothed model:

$$P_{WB}(w_i|w_{i-N+1}^{i-1}) = \lambda(w_{i-N+1}^{i-1})P_{ML}(w_i|w_{i-N+1}^{i-1}) + (1 - \lambda(w_{i-N+1}^{i-1}))P_{WB}(w_i|w_{i-N+2}^{i-1}) \tag{2.3.10}$$

The computation of $\lambda(w_{i-N+1}^{i-1})$ parameter requires the number of unique words that follow the history w_{i-N+1}^{i-1}. This number is denoted as $U_{1+}(w_{i-N+1}^{i-1}\bullet)$. Using a more formal notation we write [20]

$$U_{1+}(w_{i-N+1}^{i-1}\bullet) =\mid \{w_i : C(w_{i-N+1}^{i-1}w_i) > 0\} \mid \tag{2.3.11}$$

The number of words that occur one or more times (1+) are denoted by U_{1+}. The symbol $\bullet$ is used for a free variable (in our case, word) that is summed over. The parameter $\lambda(w_{i-N+1}^{i-1})$ is calculated as

$$\lambda(w_{i-N+1}^{i-1}) = 1 - \frac{U_{1+}(w_{i-N+1}^{i-1}\bullet)}{U_{1+}(w_{i-N+1}^{i-1}\bullet) + \sum_{w_i} C(w_{i-N+1}^{i})} \tag{2.3.12}$$

Substituting Equation 2.3.12 into Equation 2.3.10, we have

$$P_{WB}(w_i|w_{i-N+1}^{i-1}) = \frac{C(w_{i-N+1}^{i}) + U_{1+}(w_{i-N+1}^{i-1}\bullet)P_{WB}(w_i|w_{i-N+2}^{i-1})}{\sum_{w_i} C(w_{i-N+1}^{i}) + U_{1+}(w_{i-N+1}^{i-1}\bullet)} \tag{2.3.13}$$

According to Equation 2.3.10, we use the higher-order model with probability $\lambda(w_{i-N+1}^{i-1})$, while the lower-order model is used with $1-\lambda(w_{i-N+1}^{i-1})$ probability. The probability mass that equals to the $1-\lambda(w_{i-N+1}^{i-1})$ probability, is the probability of a word that is not observed immediately after the w_{i-N+1}^{i-1} history in the training data, but appears in any later position.

Absolute smoothing

In absolute smoothing [25] a higher-order model is interpolated with a lower-order model. However, instead of multiplying the higher-order distribution by a factor $\lambda(_{i-N+1}^{i-1})$, the higher-order distribution is derived by subtracting a fixed discount $D \leqslant 1$ from each non-zero count.

$$P_{Abs}(w_i|w_{i-N+1}^{i-1}) = \frac{max\{C(w_{i-N+1}^{i}) - D, 0\}}{\sum_{w_i} C(w_{i-N+1}^{i})} + (1-\lambda(w_{i-N+1}^{i-1}))P_{Abs}(w_i|w_{i-N+2}^{i-1}) \tag{2.3.14}$$

In order to make the distribution sum to 1, the $\lambda(w_{i-N+1}^{i-1})$ factor is computed as

$$\lambda(w_{i-N+1}^{i-1}) = 1 - \frac{D}{\sum_{w_i} C(w_{i-N+1}^{i})} U_{1+}(w_{i-N+1}^{i-1}\bullet) \tag{2.3.15}$$

In [25], a suggested value for D is

$$D = \frac{n_1}{n_1 + 2n_2} \tag{2.3.16}$$

where n_1 and n_2 are the total number of N-grams of the higher-order distribution with exactly one and two counts, respectively.

Kneser-Ney smoothing

In Kneser-Ney smoothing [46] the higher-order model is interpolated with a lower-order model and the higher distribution is discounted as in absolute smoothing. The difference between absolute and Kneser-Ney smoothing is in the lower-order distribution. In Kneser-Ney method, the lower-order distribution is proportional to the number of different words that it follows. Consider for example a language model trained over a corpus about computer industry and the word "Packard". If the frequency of this word is high, then the *MLE* of the unigram probability will, also, be high. The idea of Kneser-Ney smoothing is that the unigram probability of word "Packard"must be low, since it folows only one different word, "Packard".

$$\begin{aligned} P_{KN}(w_i|w_{i-N+1}^{i-1}) = & \frac{max\{C(w_{i-N+1}^{i}) - D, 0\}}{\sum_{w_i} C(w_{i-N+1}^{i})} + \\ & + \frac{D}{\sum_{w_i} C(w_{i-N+1}^{i})} U_{1+}(w_{i-N+1}^{i-1}\bullet) P_{KN}(w_i|w_{i-N+2}^{i-1}) \end{aligned} \tag{2.3.17}$$

In order to make the distribution sum to 1, we take

$$P_{KN}(w_i|w_{i-N+2}^{i-1}) = \frac{U_{1+}(\bullet w_{i-N+2}^{i})}{U_{1+}(\bullet w_{i-N+2}^{i-1}\bullet)} = \frac{|\{w_{i-N+1} : C(w_{i-N+1}^{i}) > 0\}|}{|\{w_{i-N+1}, w_i : C(w_{i-N+1}^{i}) > 0\}|} \tag{2.3.18}$$

2.3.3 Backoff

One main contribution of the discussed smoothing methods is the solution of the problem caused by the zero-count N-grams. Moreover, there is another methodology that tackle this problem. Suppose that there are no occurrences of a particular trigram, $w_{i-2}\ w_{i-1}\ w_i$, in the training corpus. In this case we can estimate the trigram probability $P(w_i \mid w_{i-2}\ w_{i-1})$ using the bigram probability $P(w_i \mid w_{i-1})$. In the same manner, if there are no counts of the bigram $w_{i-1}\ w_i$, we can estimate $P(w_i \mid w_{i-1})$ using the unigram probability $P(w_i)$. This strategy is called *b*ackoff. According to the above description of backoff method, an amount of probability mass is taken away from the higher-order models and is distributed to the lower-order models [7, 10, 15]. Of course, the resulted probability estimation must remain valid, i.e., sums to one.

The backoff model was introduced by Katz [37] and is similar to the deleted interpolation in the sense that the construction of an N-gram model is based on an N-1 model. The difference between backoff and deleted interpolation is that in backoff, for example, if there are non-zero frequency trigarm, we use only these counts without interpolating the bigram and unigram models [10]. The "back off" step downwards to a lower-order model is followed if there are zero counts for the higher-order model

For a trigram language model, the backoff method is defined as [10]

$$P(w_i \mid w_{i-2}\ w_{i-1}) = \begin{cases} P_{ML}(w_i \mid w_{i-2}\ w_{i-1}), & \text{if} \quad C(w_{i-2}\ w_{i-1}\ w_i) > 0 \\ \alpha_1 P_{ML}(w_i \mid w_{i-1}), & \text{if} \quad C(w_{i-2}\ w_{i-1}\ w_i) = 0 \\ & \quad \text{and } C(w_{i-1}\ w_i) > 0 \\ \alpha_2 P_{ML}(w_i), & \text{otherwise} \end{cases} \tag{2.3.19}$$

Some smoothing techniques assume that the unseen N-grams are all equally probable and an amount of probability mass is distributed ti them according to an even scheme. A more neat and fair way is to combine smoothing with backoff for distributing the probability mass to the unseen events. The smoothing quantifies the total mass of probability that must be reserved for the unseen events and the backoff procedure defines how to assign the reserved probability.

Let's consider Equation 2.3.19. The presence of α parameters ensures that the computed probability is a valid probability. This can be explained as follows. If the frequency of the trigram of interest is non-zero, then the $P_{ML}(w_i \mid w_{i-2}\ w_{i-1})$ probability that is computed over relative frequencies is a true probability. Otherwise, we have to back off to a lower-order model, and, then, we will add extra probability mass, resulting to a non-true probability. So, the backoff model must be smoothed. Using these considerations, the $P_{ML}(\cdot)$ probabilities of Equation 2.3.19 must be substituted by smoothed probabilities $\tilde{P}(\cdot)$. The use of smoothing saves an amount of probability mass for the lower-order models. Moreover, the α parameters guarantee that the sum of the distributed (to the lower-order models) portions of probability mass is equal to the initially saved amount of probability [10]. In the general N-gram case, the probability

mass that must be given form an N-gram to an N-1-gram is defined as follows [10].

$$\alpha(w_{i-N+1}^{i-1}) = \frac{1 - \sum_{w_i : C(w_{i-N+1}^{i})>0} \tilde{P}(w_i \mid w_{i-N+1}^{i-1})}{1 - \sum_{w_i : C(w_{i-N+1}^{i})>0} \tilde{P}(w_i \mid w_{i-N+2}^{i-1})} \tag{2.3.20}$$

Note that the α parameter is a function of the history w_{i-N+1}^{i-1}. Also, recall that the $\tilde{P}(\cdot)$ probabilities are estimated using smoothing. In final, Equation 2.3.19 is reformulated as [10]

$$P_{Bo}(w_i \mid w_{i-2}\ w_{i-1}) = \begin{cases} \tilde{P}(w_i \mid w_{i-2}\ w_{i-1}), & \text{if} \quad C(w_{i-2}\ w_{i-1}\ w_i) > 0 \\ \alpha(w_{i-2}^{i-1})\tilde{P}(w_i \mid w_{i-1}), & \text{if} \quad C(w_{i-2}\ w_{i-1}\ w_i) = 0 \\ & \qquad \text{and } C(w_{i-1}\ w_i) > 0 \\ \alpha(wi-1)\tilde{P}(w_i), & \text{otherwise} \end{cases} \tag{2.3.21}$$

Some approaches when apply the backoff method treat N-grams of only one occurrence as zero-frequency events.

2.4 N-gram language modeling toolkits

Two widely-used toolkits for N-gram language modeling that are freely available are:

- **CMU-Cambridge Statistical Language Modeling toolkit:**a suite of UNIX software tools to facilitate the construction and testing of statistical language models. The first version was written by Roni Rosenfeld at Carnegie Mellon University[45] and
- **HTK toolkit:** originally developed at the Machine Intelligence Laboratory of the Cambridge University Engineering Department . The toolkit is primarily used for building and manipulating HMMs for speech recognition, although a component for N-gram language modeling is also included.[1]

2.5 Evaluation of language models

The field of information theory [8] provides some useful notions in order to measure the performance of a language model. *Entropy* and *perplexity* are used to evaluate a language model.

Natural language is a kind of information source and a natural language sentence can be considered as a emitted signal, being a sequence of words. The distribution of the next word is highly dependent to the previous words. There is a great deal of variability and uncertainty in natural language. Entropy is a measure of information. Alternatively, entropy can be considered as a measure of 'uncertainty' of a random

[1] In this thesis, CMU was used for acquiring Linguistic Variability Metrics and HTK for Duration Metrics.

variable. Let W be a random variable that ranges over the corpus vocabulary V and has a probability function P_w. The entropy of the random variable is

$$H(W) = -\sum_{w \in V} P(w) \log_2 P(w) \tag{2.5.1}$$

If log base 2 is used, the resulting units called binary digits. If the base 10 is used, the resulting units are expressed in decimal digits. Intuitively, entropy can be interpreted as a lower bound of bits that are required in order to encode a chunk of information according to an optimal encoding [10].

Given that a language model uses all possible vocabulary words to predict the next words, it follows that the model embodies a *per-word entropy* (*entropy rate*). The per-word entropy of a language model, L, for all possible sequences of words $w_1, w_2, \ldots, w_m$ is as follows [58]

$$H(L) = -\lim_{m \to \infty} \frac{1}{m} \sum_{w_1, w_2, \ldots, w_m} P(w_1, w_2, \ldots, w_m) \log_2 P(w_1, w_2, \ldots, w_m) \tag{2.5.2}$$

If the language being modeled is ergodic [38], the summation in Equation 2.5.2 can be omitted and $H(L)$ becomes [58]

$$H(L) = -\lim_{m \to \infty} \frac{1}{m} \log_2 P(w_1, w_2, \ldots, w_m) \tag{2.5.3}$$

It is interesting to note that if we have a long enough sequence of words (given the ergodicity), then $H(L)$ can be approximated as [58]

$$\hat{H}(L) = -\frac{1}{m} \log_2 P(w_1, w_2, \ldots, w_m) \tag{2.5.4}$$

Equation 2.5.4 has a suitable form for measuring the quality of a language model in terms of per-word entropy. This measurement is achieved by the notion of perplexity as [58]

$$PP = 2^{\hat{H}(L)} \tag{2.5.5}$$

If we substitute Equation 2.5.4 into Equation 2.5.5, we have [58]

$$PP = \hat{P}(w_1, w_2, \ldots, w_m)^{-\frac{1}{m}} \tag{2.5.6}$$

that is the perplexity of a language model. $\hat{P}(w_1, w_2, \ldots, w_m)$ denotes the estimated probability that the language model, L, assigns to the sentence $w_1, w_2, \ldots, w_m$.

Perplexity can be seen as a measurement giving the average number of the most probable words ,which can follow any word, with equal probability. It follows that more qualitative language models have lower perplexities.

2.6 Summary

In this chapter we discussed briefly the main aspects of N-gram statistical language modeling. Some of the smoothing and backoff techniques were presented and an evaluation measurement for the quality of language models was defined.

Chapter 3

Acoustic Models and Acoustic Analysis Fundamentals

3.1 Introduction

The acoustic models play an important - if not the most important - role within any speech recognizer. Their task is to allow the recognizer to derive a measure, usually a probability, of the incoming stream of feature vectors corresponding to some elementary unit of speech. Three overall choices are associated with acoustic modeling:

- The choice of basic units of speech to be modelled
- The choice of model architecture taking into account the speech units and the selected feature estimation scheme
- The choice of training- and decoding algorithms

The units of speech to be modelled are usually linguistically motivated and in most cases context dependent phonemes are chosen, in particular when large or non-fixed vocabularies are to be recognized. In the last decade the dominant model architecture has been *Hidden Markov Models (HMM)* as impressive performance has been achieved across a number of sites, tasks and languages. HMMs model the sequence of feature vectors as a piecewise stationary process in which an utterance is considered as a sequence of discrete stationary states with instantaneous transitions between them. The HMM approach defines two concurrent stochastic processes: the sequence of states (modeling the temporal structure of speech) and a set of state output processes modeling the feature distribution in each state. The HMM is called hidden because the state sequence is not observable. State of the art speech recognition typically makes use of context dependent HMMs in the form of triphones or even quinphones, i.e. phonemes defined also from their context. However, to model all existing contexts for many speakers would be impossible, and parameters are therefore shared between several context dependent models to achieve robust training for all possible contexts. Decision trees are often used to cluster model parameters with respect to context and equally important, they provide useful structures for predicting unseen triphones (for exible vocabulary applications).

3.2 System Overview

The goal of speech recognition can be formulated as follows: for a given acoustic observation $X = X_1, X_2, ...X_n$, find the corresponding sequence of words $\hat{W} = w_1, w_2, ..., w_m$ with maximum *a posteriori* probability $P(W|X)$. Using *Bayes' decision rule*, this can be expressed as:

$$\hat{W} = \arg\max_w P(W|X) = \arg\max_w \frac{P(X|W)P(W)}{P(X)} \tag{3.2.1}$$

Since the acoustic observation X is fixed, equation 3.2.1 is equal to:

$$\hat{W} = \arg\max_w P(X|W)P(W) \tag{3.2.2}$$

Probability $P(W)$ is the *a priori* probability of observing W independent of the acoustic observation and is referred to as a *language model*. Probability $P(X|W)$ is the probability of observing acoustic observation X given a specific word sequence W and is determined by an *acoustic model*. In pattern recognition theory, the probability $P(X|W)$ is referred to as the *likelihood* function. It measures how likely it is that the underlying parametric model of W will generate observation X.

In a typical speech recognition process, a word sequence W is postulated and its probability determined by the language model. Each word is then converted into a sequence of phonemes using a pronunciation dictionary, also known as the *lexicon*. For each phoneme there is a corresponding statistical model called a *hidden Markov model* (HMM). The sequence of HMMs needed to represent the utterance are concatenated to a single composite model and the probability $P(X|W)$ of this model generating observation X is calculated. This process is repeated for all word sequences and the most likely sequence is selected as the recognizer output.

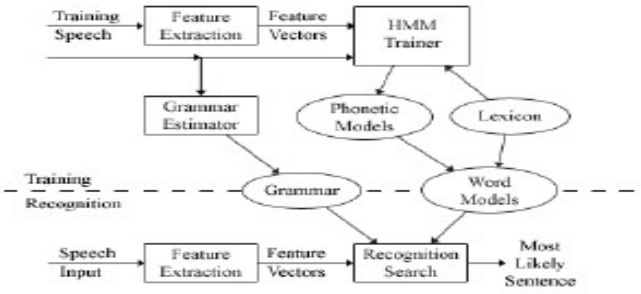


Figure 3.1: A general system for training and recognition.

Most contemporary speech recognition systems share an architecture as illustrated in Fig. 3.1. The acoustic observations are represented by *feature vectors*. Choosing

appropriate feature vectors is essential to good speech recognition. The process of extracting features from speech waveforms will be described in detail in the next section. Hidden Markov models are used almost exclusively for acoustic modeling in modern speech recognition systems.

3.3 Basic Elements of Speech Recognition and Acoustic Analysis

The acoustic analysis is the process of extracting feature vectors from input speech signals (i.e. waveforms). A feature vector is essentially a parametric representation of a speech signal, containing the most important information and stored in a compact way. In most speech recognition systems, some form of preprocessing is applied to the speech signal (i.e. applying transformations and filters), to reduce noise and correlation and extract a good set of feature vectors. In Fig. 3.2 the process of extracting feature vectors is illustrated. The speech signal is divided into analysis frames at a certain frame rate. The size of these frames is often 10 ms, the period that speech is assumed to be stationary for. Features are extracted from an analysis window. The size of this window is independent of the *frame rate*. Usually the window size is larger than the frame rate, leading to successive windows overlapping, as is illustrated in Fig. 3.2. Much work is done in the field of signal processing and several methods of speech analysis exist. Two of the most popular will be discussed: *linear predictive coding* and *Mel-frequency cepstral analysis*.

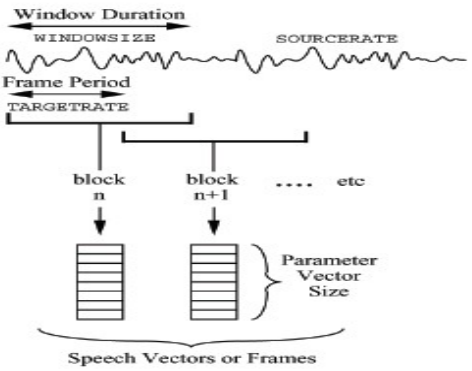


Figure 3.2: Feature Extraction.

3.3.1 Linear Predictive Coding

Linear predictive coding (LPC) is a fast, simple and effective way of estimating the main parameters of speech. In linear predictive coding the human vocal tract is modeled as an *infinite impulse response* filter system that produces the speech signal. This

modeling produces an accurate representation of vowel sounds and other voice speech segments that have a resonant structure and a high degree of similarity over time shifts that are multiples of their pitch period. The linear prediction problem can be stated as finding the coefficients ak, which result in the best prediction (that minimizes the mean-square prediction error) of speech sample $s[n]$ in terms of past samples $s[n-k]$, with $k = 1, 2, ..., P$. The predicted sample $\hat{s}[n]$ is given by:

$$\hat{s}[n] = \sum_{k=1}^{P} a_k s[n-k] \tag{3.3.1}$$

with P the required number of past sample of $s[n]$. The prediction error can be formulated as:

$$e[n] = s[n] - \hat{s}[n] = s[n] - \sum_{k=1}^{P} a_k s[n=k] \tag{3.3.2}$$

To find the predictor coefficients several methods exist, such as the Covariance method and the Autocorrelation method. In both methods the key to finding the predictor coefficients involves solving large matrix equations.

3.3.2 Mel-Frequency Cepstral Analysis

In contrast to linear predictive coding, Mel-frequency cepstral analysis is a *perceptually motivated* representation. Perceptually motivated representations include some aspect of the human auditory system in their design. In the case of Mel-frequency cepstral analysis, a nonlinear scale, referred to as the Mel-scale, is used that mimics the acoustic range of the human hearing. The Mel-scale can be approximated by:

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.3.3}$$

The process of obtaining feature vectors based on the Mel-frequency is illustrated in Fig. 3.3. First, the signal is transformed to the spectral domain by a Fourier transform. The obtained spectrum of the speech signal is then smoothed by integrating the spectral coefficients with triangular frequency bins arranged on the non-linear Mel-scale.

Next, a log compression is applied to the filter bank output, in order to make the statistics of the estimated speech power spectrum approximately Gaussian. In the final processing stage, a discrete cosine transform (DCT) is applied. It is common for feature vectors derived from Mel-frequency cepstral analysis to contain first-order and second-order differential coefficients besides the static coefficients. Sometimes a measure of the signal energy is included. A typical system usings feature vectors based on *Melfrequency cepstral coefficients* (MFCCs) can have the following configuration:

- 13th-order MFCC c_k
- 13th-order 1st-order delta MFCC computed from $\triangle c_k = c_{k+2} - c_{k-2}$
- 13th 2nd-order delta MFCC computed from $\triangle\triangle c_k = \triangle c_{k+1} - \triangle c_{k-1}$

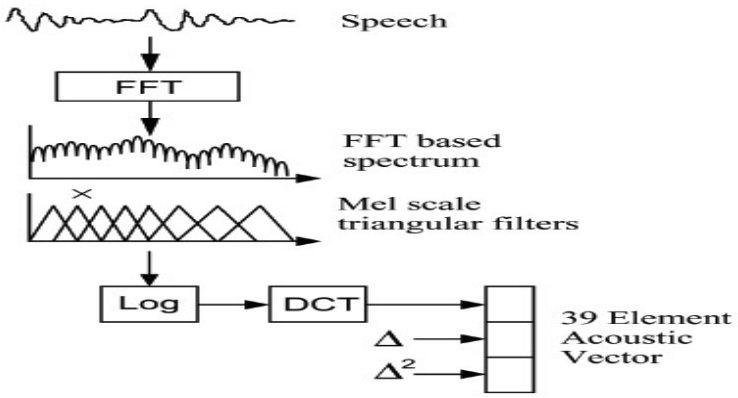


Figure 3.3: Mel-frequency cepstral coefficients.

3.3.3 Hidden Markov Models

In this section the *hidden Markov model* (HMM) will be introduced. The HMM is a powerful statistical method of characterizing data samples of a discrete time-series. Data samples can be continuously or discretely distributed and can be either scalars or vectors. The HMM has become the most popular method for modeling human speech and is used successfully in automatic speech recognition, speech synthesis, statistical language modeling and other related areas. As an introduction to hidden Markov models, the *Markov chain* will be described first.

Markov chains

A first order Markov Chain of N states is a triplet (S, A, π), S is a set of N states, A is the N x N matrix of the transition probabilities between states, π is an N-dimensional row vector of the probability to be in a state at the first time. A Markov Chain property is that the sum of each row of A is one. Another property that holds for Markov Chains is that initial probabilities π_i must sum up to one. The Markov chain is a state model of a process where one can observe the state transitions that for a first-order Markov Chain depend only on the state at the previous discrete time.

Hidden Markov Models

A Hidden Markov Model is a Markov Chain where output symbols or probabilistic functions describing output symbols are associated to the states. This result to a model with embedded stochastic process with an underlying stochastic process that is not directly observable but can be observed only through another set of stochastic processes that produce the sequence of observations.

An HMM with discrete symbol observations is characterized by:

1. The number of states in the model. Although the states of the model are hidden, there is some significance attached to the states, such as in speech signals. These states are labeled as $1, 2, .., N$.

2. The number of distinct observation symbols,M, per state, for example a set of phonemes. These symbols are denoted as $V = v_1, v_2, ..., v_M$.

3. The state-transition probability distribution $A = a_{ij}$ where

   $a_{ij} = P[q_{t+1} = j | q_t = i], 1 \leq i, j \leq N$

4. The observation symbol probability distribution $B = b_j(k)$, in which

   $b_j(k) = P[o_t = v_k | q_t = j], 1 \leq k \leq M$

   defines the symbol distribution in state j.

5. The initial distribution in state $\pi = \{\pi_t\}$, in which

   $\pi_t = P[q_1 = i], 1 \leq i \leq M$

To summarize, HMM specification requires two model parameters N and M, specification of the observation symbols and the specification of the three sets of probability measures A, B and π, described by the compact notation:

$\lambda = (A, B, \pi)$

3.3.4 Basic Algorithms on HMMs

Given the definition of an HMM above, there are three basic problems that need to be addressed:

1. **The Evaluation Problem** : Given an HMM λ and a sequence of observations $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$, what is the probability that the observations are generated by the model, $p\{O|\lambda\}$?

2. **The Decoding Problem**: Given a model λ and a sequence of observations $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$, what is the most likely state sequence in the model that produced the observations?

3. **The Learning Problem**: Given a model λ and a sequence of observations $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T$, how should we adjust the model parameters Λ, B, π in order to maximize $p\{O|\lambda\}$?

Evaluation problem can be used for isolated (word) recognition. Decoding problem is related to the continuous recognition as well as to the segmentation. Learning problem must be solved, if we want to train an HMM for the subsequent use of recognition tasks.

The Evaluation Problem and the Forward Algorithm

Let the forward probability $\alpha_j(t)$ for some model M with N states be defined as

$$\alpha_j(t) = P(\mathbf{o}_1, \ldots, \mathbf{o}_t, x(t) = j|M). \tag{3.3.4}$$

That is, $\alpha_j(t)$ is the joint probability of observing the first t speech vectors and being in state j at time t. This forward probability can be efficiently calculated by the following recursion

$$\alpha_j(t) = \left[\sum_{i=2}^{N-1} \alpha_i(t-1)a_{ij}\right] b_j(\mathbf{o}_t). \tag{3.3.5}$$

This recursion depends on the fact that the probability of being in state j at time t and seeing observation $\mathbf{o}_t$ can be deduced by summing the forward probabilities for all possible predecessor states i weighted by the transition probability a_{ij}. The slightly odd limits are caused by the fact that states 1 and N are non-emitting.The initial conditions for the above recursion are

$$\alpha_1(1) = 1 \tag{3.3.6}$$

$$\alpha_j(1) = a_{1j}b_j(\mathbf{o}_1) \tag{3.3.7}$$

for $1 < j < N$ and the final condition is given by

$$\alpha_N(T) = \sum_{i=2}^{N-1} \alpha_i(T)a_{iN}. \tag{3.3.8}$$

Notice here that from the definition of $\alpha_j(t)$,

$$P(\mathbf{O}|M) = \alpha_N(T). \tag{3.3.9}$$

Hence, the calculation of the forward probability also yields the total likelihood $P(\mathbf{O}|M)$.

The backward probability $\beta_j(t)$ is defined as

$$\beta_j(t) = P(\mathbf{o}_{t+1}, \ldots, \mathbf{o}_T|x(t) = j, M). \tag{3.3.10}$$

As in the forward case, this backward probability can be computed efficiently using the following recursion

$$\beta_i(t) = \sum_{j=2}^{N-1} a_{ij}b_j(\mathbf{o}_{t+1})\beta_j(t+1) \tag{3.3.11}$$

with initial condition given by

$$\beta_i(T) = a_{iN} \tag{3.3.12}$$

for $1 < i < N$ and final condition given by

$$\beta_1(1) = \sum_{j=2}^{N-1} a_{1j}b_j(\mathbf{o}_1)\beta_j(1). \tag{3.3.13}$$

Notice that in the definitions above, the forward probability is a joint probability whereas the backward probability is a conditional probability. This somewhat asymmetric definition is deliberate since it allows the probability of state occupation to be determined by taking the product of the two probabilities. From the definitions,

$$\alpha_j(t)\beta_j(t) = P(\mathbf{O}, x(t) = j|M). \tag{3.3.14}$$

Hence the probability of state occupation $L_j(t)$ is

$$\begin{aligned} L_j(t) &= P(x(t) = j|\mathbf{O}, M) \\ &= \frac{P(\mathbf{O}, x(t) = j|M)}{P(\mathbf{O}|M)} \\ &= \frac{1}{P}\alpha_j(t)\beta_j(t) \end{aligned} \tag{3.3.15}$$

where $P = P(\mathbf{O}|M)$.

The Decoding Problem and the Viterbi Algorithm (Recognition)

The previous section has described the basic ideas underlying HMM parameter re-estimation using the Baum-Welch algorithm. In passing, it was noted that the efficient recursive algorithm for computing the forward probability also yielded as a by-product the total likelihood $P(\mathbf{O}|M)$. Thus, this algorithm could also be used to find the model which yields the maximum value of $P(\mathbf{O}|M_i)$, and hence, it could be used for recognition.

In practice, however, it is preferable to base recognition on the maximum likelihood state sequence since this generalizes easily to the continuous speech case whereas the use of the total probability does not. This likelihood is computed using essentially the same algorithm as the forward probability calculation except that the summation is replaced by a maximum operation. For a given model M, let $\phi_j(t)$ represent the maximum likelihood of observing speech vectors $\mathbf{o}_1$ to $\mathbf{o}_t$ and being in state j at time t. This partial likelihood can be computed efficiently using the following recursion (cf. equation 3.3.5)

$$\phi_j(t) = \max_i \{\phi_i(t-1)a_{ij}\}\, b_j(\mathbf{o}_t). \tag{3.3.16}$$

where

$$\phi_1(1) = 1 \tag{3.3.17}$$

$$\phi_j(1) = a_{1j}b_j(\mathbf{o}_1) \tag{3.3.18}$$

for $1 < j < N$. The maximum likelihood $\hat{P}(O|M)$ is then given by

$$\phi_N(T) = \max_i \{\phi_i(T)a_{iN}\} \tag{3.3.19}$$

As for the re-estimation case, the direct computation of likelihoods leads to underflow, hence, log likelihoods are used instead. The recursion of equation 3.3.16 then becomes

$$\psi_j(t) = \max_i \{\psi_i(t-1) + log(a_{ij})\} + log(b_j(\mathbf{o}_t)). \tag{3.3.20}$$

This recursion forms the basis of the so-called Viterbi algorithm. As shown in Fig. 3.4, this algorithm can be visualised as finding the best path through a matrix where the vertical dimension represents the states of the HMM and the horizontal dimension represents the frames of speech (i.e. time). Each large dot in the picture represents the log probability of observing that frame at that time and each arc between dots corresponds to a log transition probability. The log probability of any path is computed simply by summing the log transition probabilities and the log output probabilities along that path. The paths are grown from left-to-right column-by-column. At time t, each partial path $\psi_i(t-1)$ is known for all states i, hence equation 3.3.20 can be used to compute $\psi_j(t)$ thereby extending the partial paths by one time frame.

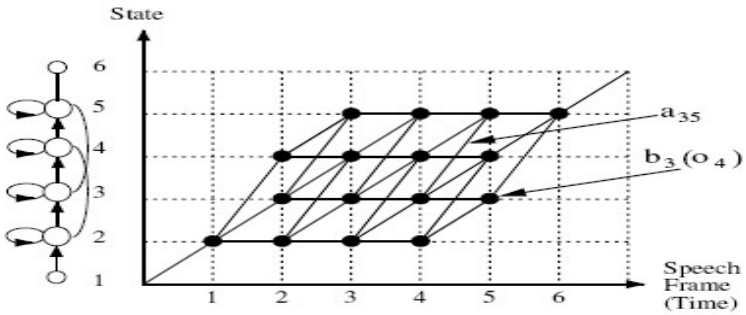


Figure 3.4: The Viterbi Algorithm for Isolated Word Recognition.

The Learning Problem and the Baum-Welch Re-Estimation

To determine the parameters of a HMM it is first necessary to make a rough guess at what they might be. Once this is done, more accurate (in the maximum likelihood sense) parameters can be found by applying the so-called Baum-Welch re-estimation formulae.

Here the basis of the formulae will be presented in a very informal way. Firstly, it should be noted that the inclusion of multiple data streams does not alter matters significantly since each stream is considered to be statistically independent. Furthermore, mixture components can be considered to be a special form of sub-state in which the transition probabilities are the mixture weights

Thus, the essential problem is to estimate the means and variances of a HMM in which each state output distribution is a single component Gaussian, that is

$$b_j(\mathbf{o}_t) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma_j}|}} e^{-\frac{1}{2}(\mathbf{o}_t-\mu_j)'\mathbf{\Sigma}_j^{-1}(\mathbf{o}_t-\mu_j)} \tag{3.3.21}$$

If there was just one state j in the HMM, this parameter estimation would be easy. The

maximum likelihood estimates of μ_j and $\mathbf{\Sigma}_j$ would be just the simple averages, that is

$$\hat{\mu}_j = \frac{1}{T}\sum_{t=1}^{T}\mathbf{o}_t \tag{3.3.22}$$

and

$$\hat{\mathbf{\Sigma}}_j = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{o}_t - \mu_j)(\mathbf{o}_t - \mu_j)' \tag{3.3.23}$$

In practice, of course, there are multiple states and there is no direct assignment of observation vectors to individual states because the underlying state sequence is unknown. Note, however, that if some approximate assignment of vectors to states could be made then equations 3.3.22 and 3.3.23 could be used to give the required initial values for the parameters.

Since the full likelihood of each observation sequence is based on the summation of all possible state sequences, each observation vector $\mathbf{o}_t$ contributes to the computation of the maximum likelihood parameter values for each state j. In other words, instead of assigning each observation vector to a specific state as in the above approximation, each observation is assigned to every state in proportion to the probability of the model being in that state when the vector was observed. Thus, if $L_j(t)$ denotes the probability of being in state j at time t then the equations 3.3.22 and 3.3.23 given above become the following weighted averages

$$\hat{\mu}_j = \frac{\sum_{t=1}^{T} L_j(t)\mathbf{o}_t}{\sum_{t=1}^{T} L_j(t)} \tag{3.3.24}$$

and

$$\hat{\mathbf{\Sigma}}_j = \frac{\sum_{t=1}^{T} L_j(t)(\mathbf{o}_t - \mu_j)(\mathbf{o}_t - \mu_j)'}{\sum_{t=1}^{T} L_j(t)} \tag{3.3.25}$$

where the summations in the denominators are included to give the required normalisation.

Equations 3.3.24 and 3.3.25 are the Baum-Welch re-estimation formulae for the means and covariances of a HMM. A similar but slightly more complex formula can be derived for the transition probabilities.

In equations 3.3.24 and 3.3.25, the probability of state occupation $L_j(t)$ is given by equation 3.3.15.

All of the information needed to perform HMM parameter re-estimation using the Baum-Welch algorithm is now in place. The steps in this algorithm may be summarised as follows

1. For every parameter vector/matrix requiring re-estimation, allocate storage for the numerator and denominator summations of the form illustrated by equations 3.3.24 and 3.3.25. These storage locations are referred to as *accumulators*

2. Calculate the forward and backward probabilities for all states j and times t.

3. For each state j and time t, use the probability $L_j(t)$ and the current observation vector $\mathbf{o}_t$ to update the accumulators for that state.

4. Use the final accumulator values to calculate new parameter values.

5. If the value of $P = P(\mathbf{O}|M)$ for this iteration is not higher than the value at the previous iteration then stop, otherwise repeat the above steps using the new re-estimated parameter values.

All of the above assumes that the parameters for a HMM are re-estimated from a single observation sequence, that is a single example of the spoken word. In practice, many examples are needed to get good parameter estimates. However, the use of multiple observation sequences adds no additional complexity to the algorithm. Steps 2 and 3 above are simply repeated for each distinct training sequence.

One final point that should be mentioned is that the computation of the forward and backward probabilities involves taking the product of a large number of probabilities. In practice, this means that the actual numbers involved become very small.

3.4 Acoustic Modeling

This section focuses on the application of hidden Markov models to modeling human speech. First, the selection of appropriate modeling units will be described, after which model topology will be discussed.

3.4.1 Selecting Model Units

When considering using hidden Markov models to model human speech, an essential question is what unit of language to use. Several possibilities exist, such as: words, syllables or phonemes. Each of these possibilities has advantages as well as disadvantages. At a high level, the following criteria need to be considered when choosing an appropriate unit:

- The unit should be *accurate* in representing the acoustic realization in different contexts.

- The unit should be *trainable*. Enough training data should exist to properly estimate unit parameters.

- The unit should be *generalizable*, so that any new word can be derived.

A natural choice to consider is using whole-word models, which have the advantage of capturing the coarticulation effects inherent within these words. When properly trained, word models in small-vocabulary recognition systems yield the best recognition results compared to other units. Word models are both accurate and trainable and there is no need to be generalizable. For large-vocabulary continuous speech recognition, however, whole word models are a poor choose. Given a fixed set of words, there is no obvious way to derive new words, making word models not generalizable. Each word needs to be trained separately and thus a lot of training data is required to properly train each unit. Only if such training data exists, are word models trainable and accurate. An alternative to using whole-word models is the use of phonemes. English and other European language typically have between 40 and 50 phonemes.

Acoustic models based on phonemes can be trained sufficiently with as little as a few hundred sentence, satisfying the trainability criterium. Phoneme models are by default generalizable as they are the principle units all vocabulary can be constructed with. Accuracy, however, is more of an issue, as the realization of phonemes is strongly affected by its neighboring phonemes, due to coarticulatory effects. Phonetic models can be made significantly more accurate by taking context into account, which usually refers to the immediate left and right neighboring phonemes. This leads to biphone and triphone models. A triphone phoneme model takes into consideration both its left and right neighbor phone thus capturing the most important coarticulatory effects. Unfortunately trainability becomes an issue when using triphone models, as there can be as many as $50x50x50 = 125000$ of them.

3.4.2 Model Topology

Speech is a non-stationary signal that evolves over time. Each state of an HMM has the ability to capture some stationary segment in a non-stationary speech signal. A left-to-right topology thus seems the natural choice to model the speech signal. Transition from left-to-right enable a natural progression of the evolving signal and self-transition can be used to model speech features belonging to the same state. Fig. 3.5 illustrates a typical 3-state HMM common to many speech recognition systems. The first state, the entry-state, and the final state, the exit-state are so called null-states. These states do not have self loops and do not generate observations. Their purpose is merely to concatenate different models.

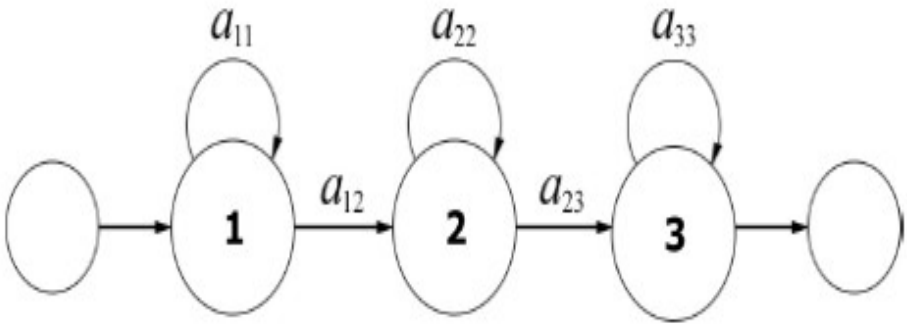


Figure 3.5: Basic structure of a phonetic HMM.

The number of internal states of an HMM can vary depending on the model unit. For HMMs representing a phoneme, three to five states are commonly used. If the HMM represents a word, a significantly larger number of internal states is required. Depending on the pronunciation and duration of the word, this can be 15 to 25 states. More complex transitions between states than the simple topology illustrated in Fig. 3.5 are also possible. If skipping states is allowed, the model becomes more flexible, but

also harder to train properly.

The choice of output probability function $b_j(\mathbf{o}_t)$ is essential to good recognizer design. Early HMM systems used discrete output probability functions in conjunction with vector quantization. Vector quantization is computationally efficient but introduces quantization noise, limiting the precision that can be obtained. Most contemporary systems use parametric continuous density output distributions. Multivariate Gaussian mixture density functions, which can approximate any continuous density function, are popular among contemporary recognition systems. Given M Gaussian mixture density functions: $\{b_j(\mathbf{o}_t)\}$

Most continuous density HMM systems, represents output distributions by Gaussian Mixture Densities. However, a further generalization can be made. Each observation vector at time t is split into a number of S independent data streams o_{st}.

The formula for computing $b_j(\mathbf{o}_t)$ is then

$$b_j(\mathbf{o}_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_s} c_{jsm} \mathcal{N}(\mathbf{o}_{st}; \mu_{jsm}, \mathbf{\Sigma}_{jsm}) \right]^{\gamma_s} \tag{3.4.1}$$

where M_s is the number of mixture components in stream s, c_{jsm} is the weight of the m'th component and $\mathcal{N}(\cdot; \mu, \mathbf{\Sigma})$ is a multivariate Gaussian with mean vector μ and covariance matrix $\mathbf{\Sigma}$, that is

$$\mathcal{N}(\mathbf{o}; \mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{o}-\mu)' \mathbf{\Sigma}^{-1} (\mathbf{o}-\mu)} \tag{3.4.2}$$

where n is the dimensionality of $\mathbf{o}$.

The exponent γ_s is a stream weightİt can be used to give a particular stream more emphasis, however, it can only be set manually.

Multiple data streams are used to enable separate modeling of multiple information sources.

3.5 Forced Alignment-Forced Segmentation

A speech recognition system uses a search engine along with an acoustic and language model which contains a set of possible words, phonemes, or some other set of data to match speech data to the correct spoken utterance. The search engine processes the features extracted from the speech data to identify occurences of the words, phonemes, or whatever set of data it is equipped to search for and returns the results.

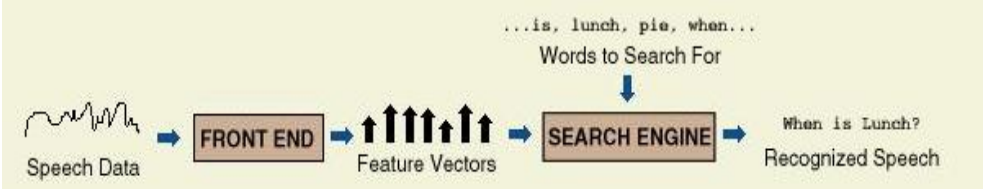


Figure 3.6: General Decoding.

Forced alignment is similar to this process, but it differs in one major aspect. Rather than being given a set of possible words to search for, the search engine is given an exact transcription of what is being spoken in the speech data. The system then aligns the transcribed data with the speech data, identifying which time segments in the speech data correspond to particular words in the transcription data. Forced alignment can also be used to align the phonemes of the transcription data to the speech data given, as shown in Fig. 3.7, although with more explicitly defined boundaries on where each phoneme begins and ends.

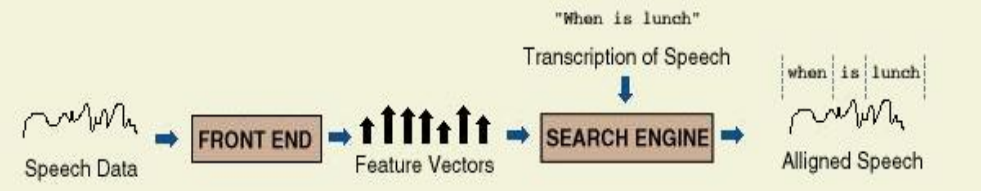


Figure 3.7: Forced Alignment.

In this case, the recognition network is constructed from a word level transcription and a dictionary. The compiled network may include optional silences between words and pronunciation variants. Forced alignment is often useful during training to automatically derive phone level transcriptions. It can also be used in automatic annotation systems.

Using the HTK speech recognition software we can generate forced alignments by computing a new network for each input utterance using the word level transcriptions and the dictionary. By default, the output transcription will just contain the words and their boundaries. With HTK however we can determine the actual pronunciations used in the utterances used to train the HMM system. Initially models are trained on the basis of one fixed pronunciation per word. Then HTK tools are used in forced alignment mode to select the best matching pronunciations. The new phone level transcriptions can then be used to retrain the HMMs.

3.6 Summary

In this chapter we discussed Hidden Markov Models and the fundamentals of speech analysis. We also presented the main design characteristics of acoustic models and the idea behind forced segmentation method.

Chapter 4

The CHIMP Experiment

4.1 Introduction

The 'CHildren's Interactive Multimedia Project' (acronym: CHIMP)aimed at providing essential guidelines for engineering successful multimodal-input multimedia-output applications for children with an emphasis on the spoken dialog interface. The 'Agent CHIMP' prototype[42, 39, 2] was developed in order to investigate how children converse with interactive systems and to collect speech data, dialog interaction and user experience data in a realistic spoken language application environment. It combines speech, keyboard and mouse input modalities and uses text, graphics, speech and animation for output presentation. The application is controlled by animated agents.

Since increased acoustic and linguistic variability are typical of spontaneous speech, a Wizard of Oz (WoZ) experiment was designed. This means that there was no actual ASR in the system just a person operating the computer. The children were not informed of the existence of a wizard and an observation room. Further, for approximately half of the experimental runs the player was alone in the game room without a moderator present.

4.2 Game Description

About 160 children, aged 6 to 14 years, participated in the study by playing an interactive computer game using voice commands, or keyboard and mouse control. The software selected for this WoZ experiment was the popular computer game 'Where in the U.S.A. is Carmen Sandiego?' (WITUICS) by Broderbund Software. WITUICS is an interactive detective game for children ages eight years and older. The game was rich in dialog subtasks including navigation and multiple queries, database entry, and database search. During the game, the children engaged in conversations with animated cartoon characters on the screen, thus the dialog is more natural and human-like. As a result spontaneous speech could be elicited.

Although the children believe that they are working with a piece of software, they are in fact just interacting with the investigator (the wizard) who is located at another terminal, carefully manipulating the computer agent and conducting the games. This makes it appear as though the child is speaking with and giving commands to the

computer and facilitates a natural child-computer agent interaction.

To successfully complete the game, i.e., arrest the appropriate suspect, two subtasks had to be completed:

1. Determine the physical characteristics of the suspect and complete a profile sketch to issue an arrest warrant, and

2. Track the suspects whereabouts and apprehend him.(by traveling through at least five of the 50 U.S. states every game)

During the game the player could talk to various characters appearing on the game screen seeking clues about the suspect's whereabouts and physical appearance. Additional aids to interpret the clues, such as geographical databases that could be queried using single or multiple word searches. Overall, the game was rich in dialog subtasks including: navigation and multiple queries, database entry, and database search.

Successful ending of the game resulted when the player travelled to the correct location and identified the suspect correctly (using the constructed profile information) from among several cartoon characters on the screen.

Game Procedure

The investigator leads the child and parent into the testing room. This room is relatively empty and contains only a couch designated for the parent of the child, two chairs, a desk, and a computer monitor that sits on top of the desk. A one-way mirror connects this room to the adjacent room that serves as the control center and viewing area for the investigators. The parent is instructed to sit on the couch while the child and investigator sit at the computer. The investigator asks the child a brief series of open-ended questions that illicit a sampling of the child's natural, person-to-person behavior. The investigator then introduces the on-screen agent and begins to play one of the games, demonstrating how the games should be played and showing the child that the computer responds to voice commands. After this demonstration, the investigator suggests that the child interacts with the machine. The investigator sits out of the way and makes sure that the child is properly positioned in front of the screen. When everything is set, the agent welcomes the child and begins a conversation based on a set of questions similar to those asked by the investigator. After the conversation section, the agent leads the child into the game section. The child works through the various exercises until all are completed. The agent thanks the child for playing and then says goodbye. Before the child leaves, the investigator asks a series of debriefing questions regarding the child's experience with the computer and agent.

4.3 Experimental Design

The Wizard of Oz (WoZ) experimental setup is shown in Fig. 4.1

In the preparatory stages of the study, short video clips of the realistically animated on-screen teenager agent were created that depict him asking questions, giving encouragement, and making other statements that could potentially arise in the conversations

with the child. The wizard simply cues these video clips whenever appropriate during the interaction. Through this process, the wizard can manipulate the agent so that it appears to the child that the agent is actually responding and commenting in real time.

All of the software runs on a single system that is located in the control room. The system that the child interacts with is actually just a second monitor stemming from the control system. The child views and reacts to the events that are put on screen and the wizard manipulates the software accordingly. The audio of the experiment is recorded by a set of microphones and stored on computer. A camcorder that is mounted above the mirror records a frontal view of the child while the on-screen events are recorded onto a separate video.Neither the loudspeaker nor the video camera was reported as being intrusive by any of the children.

4.4 Description of Database

Data from a total of 160 children and 7 adult players were collected using the speech WoZ scenario (with no recognition errors). Most players played the game twice. The total number of games played per age group and gender are shown in Tbl. 4.1:

	Age								
Gnd	8	9	10	11	12	13	14	8-14	>21
F	18	23	32	24	10	8	4	119	5
M	21	51	16	23	21	25	14	171	8

Table 4.1: Number of games per age group.

There was also a limited amount of data collected for 6 and 7 year-olds; these data were excluded from our study

A total of about 50000 utterances were collected. In order to obtain statistically significant results, our analysis focused on three age and gender groups, namely: 8-9, 10-11 and 12-14[1].

In addition, each utterance was annotated with the emotional state of the child [2]. Child-computer interaction turns were also manually categorized into a set of predefined 'dialogue states'. Dialog states roughly corresponded to one (or a group of similar) game actions taken by the wizard in response to a voice command.

The data was transcribed and annotated for disfluencies and hesitation phenomena. Child-computer interaction turns were also manually assigned to a set of predefined 'dialog states' according to the game actions they triggered [42]. Dialog states roughly corresponded to one (or a group of similar) game actions taken by the wizard in response to a user input(voice command).

For example, the dialog state 'Talk2Him' incorporated voice commands asking for a cartoon character's attention, while states 'WhereDid' and 'TellMeAbout' correspond to queries about the suspect's whereabouts and physical characteristics, respectively.

[1] However detailed results for all ages were calculated although not presented here

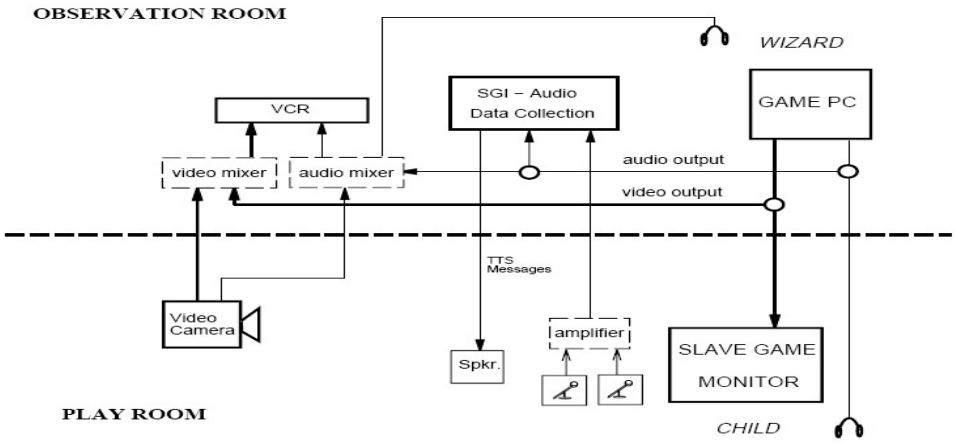


Figure 4.1: The experimental WoZ setup.

User input/System output	Dialogue State
User: Tell me about the suspect? System: She is neither long- nor short-legged	S_{t-3} : TellmeAbout
U: Her height is average S: ... [updating suspect's drawing]	S_{t-2} : EnterFeature
U: Where did the suspect go? S: She is picking peonies in Bloomington	S_{t-1} : WhereDid
U: Go to Indiana S: ... [travel theme]	S_t : GoToState

Table 4.2: Transcript of a sample interaction along with dialog state tags.

Spoken utterances were assigned to predefined dialog states by the wizard's assistant while the game was being played and were later verified by a group of human labelers. In Tbl. 4.2 a sample interaction illustrating dialog state is presented.

A total of about 30 dialog states were identified for this application. The flow through the task was characterized by a sequence of dialog state transitions. In Fig 4.2 shows the dialog flow diagram for the navigation/query subdialog.

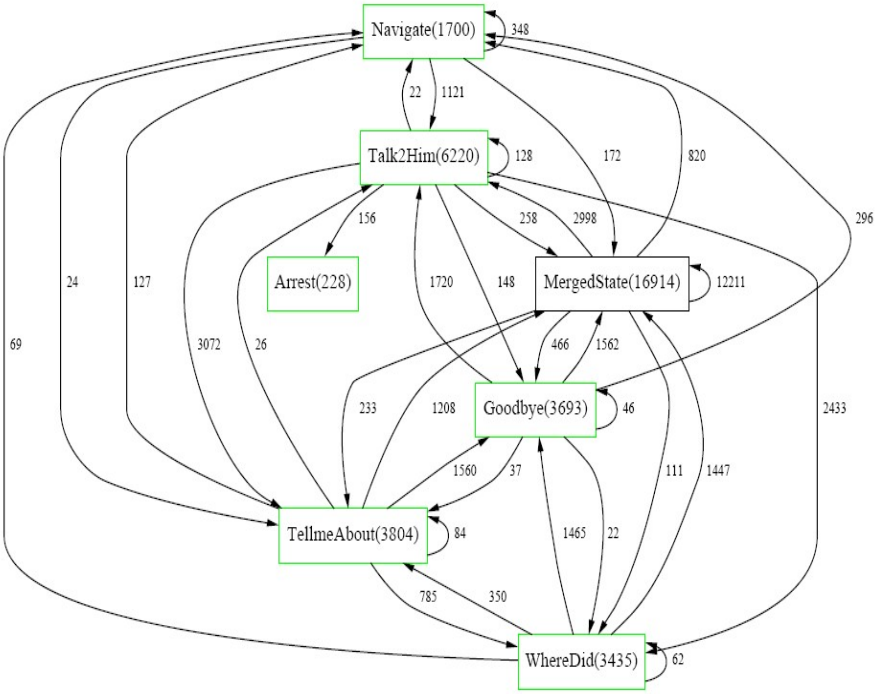


Figure 4.2: Dialog state and state transition diagram (with counts) for all children players for the navigation/query subdialog ('MergedState' denotes combination of all dialog states not shown in plot).

The total number of times a state is visited (in parenthesis) and the total number of state transitions (arrow labels) are shown for all games played by children players (total of 290 games).

Useful information about problem-solving and dialog strategies of children can be drown by such graphs. For example, consider the state marked 'Tellme-About' in Fig 4.2. It can be seen that only about 20% of the time (785/3804) the child requested a second piece of clue; instead, the child preferred to utilize the first piece of

information obtained about 72% (2768/3804) of the time this state was visited. In other words, most children preferred to concentrate on a single task per turn.

4.5 Summary

In this chapter, we presented the basic aspects of the WoZ experiment contacted and the main characteristics of our database, i.e. the database acquired by the experiment. We are now ready to proceed with the presentation of the calculated metrics and the method we used in order to acquire our results.

Chapter 5

Linguistic Analysis of the Corpus

5.1 Introduction

Our linguistic analysis of spontaneous children speech was based on the CHIMP database, as described in the previous chapter. Data from the 'game' sessions was excluded as well as the data for children aged 6-7. While performing our analysis we also spotted and excluded certain outliers (children with distinctive regional pronunciation)

5.2 Metrics Calculated

The list of acoustic and linguistic correlates measured and method of measurement follows:

- *Duration metrics:* Phone and sentence durations, rate of speech information, and between-word silence duration were computed from automatically estimated phone-level corpus segmentation. Average durations were computed per speaker and plotted per age and gender. Intra- and inter-speaker duration variability were also computed.

    1. *Phone duration*: computed per phone, age and gender group.
    2. *Sentence duration*: computed per age and gender group (average of all sentences in the corpus).
    3. *Sentence duration per dialogue state*: computed per dialogue state, age and gender group.
    4. *Between-word silence duration*: computed per age and gender group.
    5. *Speaking rate*: computed as the average number of phones/sec, per age and gender group.
    6. *Speaking rate excluding silence*: computed as the average number of phones/sec (excluding between word silence segments), per age and gender group.

- *Fluency metrics:* False-starts, mispronunciations, hesitations and filled pauses were manually labeled on the spontaneous speech corpus.

1. *False-starts and mispronunciations*: per age and gender group.
2. *Hesitations and filled-pauses*: per age and gender group.

- *Lexical and Syntactic metrics:* Sentence length, vocabulary size and lexical-variation were estimated on the manually transcribed corpus.

1. *Sentence Length (in words)*: per age and gender group.
2. *Vocabulary size (unique words)*: per age and gender group.
3. *Vocabulary size (total words)*: per age and gender group.
4. *Linguistic variability (perplexity)*: computed as bigram language model perplexity per age and gender group.
5. *Linguistic variability (perplexity) within dialogue state*: computed as bigram language model perplexity per dialogue state, age and gender group.
6. *Intra- vs inter- speaker linguistic perplexity*: computed as the average ratio of the perplexity of the bigram language model of the one speaker's utterances vs all speaker's utterances (in an age and gender group).
7. *Linguistic variability turn to turn (within a dialogue state)* : computed as the Levenshtein distance between two adjacent utterances in the same dialogue state, age and gender group.

All the metrics presented here were calculated both for all ages and gender and for the 3 age-gender groups. Thew were plotted per age and gender group using MATLAB and statistical significance of the results was tested using 2-way ANOVA analysis.

5.3 Duration Metrics

5.3.1 Preparing Corpus

In order to calculate the duration metrics we needed a transcribed corpus identical to the acoustic data (the acoustic model was trained in clean data). We removed all utterances that contained special labels or dashes (indication of false start) (this left us with 26655 out of the 35988 original utterances, 74%) and the empty utterances (leaving us with 25719 utterances). We didn't change words that contain an apostrophe ('), like don't, since they are included in our lexicon. We also removed outlier users and ended up with 25502 utterances of clean data.

In order to check whether, in our cleaned corpus, there are words that do not appear in our lexicon, we created a list of all the words of our clean corpus and another one with all the words in the lexicon, we compared these two lists and we created a 'missing words' list. This first missing words list was quite long, including both uppercase and lowercase characters. We lowercased the corpus and the lexicon and compared the lists again. We also added manually a few words (about 5) from a different lexicon, one used in $AURORA_4$ database. Finally the missing words were narrowed down to 512 (most of them very rare). We then removed all the utterances that contain at least one of the 'missing words', and we ended up with 24395 utterances out of the original 25502 clean utterances.

5.3.2 Forced Segmentation and HTK

In order to compute the desired metrics we performed automatically estimated phone-level corpus segmentation using HTK Toolkit. We used a context-independent phone-level acoustic model [1] (three-state per phoneme with eight Gaussians per state, Hamming window 25 ms and a frame update of 10 ms, 16KHz sampling frequency) and hand-labeled word-level transcriptions. The model was trained on the whole spontaneous corpus.

We used the HVITE tool to compute forced alignments. Fig. 5.2 illustrates the way HVITE performs this.

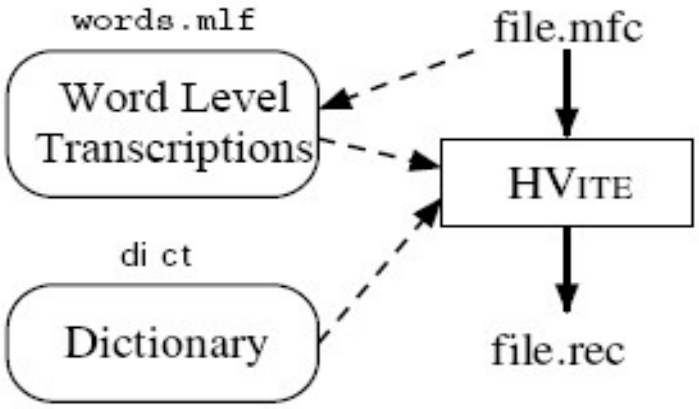


Figure 5.1: Forced Alignment with HTK.

After the above mentioned preparation of the corpus, we ended up with three files: the file *words.mlf*, that contains the word-level transcriptions of the utterances, the dictionary *dict*, which contains phone level transcriptions of the words and a file containing a list of the utterances and the path to the audio files. These three files are the input for HVITE. The decoder uses the list of files to locate the corresponding file.mfc (for each audio file), list them in a .scp file, finds the best matching path through the network and constructs a lattice which includes model alignment information. Finally the lattice is converted to a transcription. An output.mlf file with both model and word level transcriptions of all the utterances is returned. For example, a fragment of our output, for the utterance 'read that note' was:

'/august01.cristina.sub148.1.1_0107.rec'*
0 2900000 r -2770.488525 read
2900000 3800000 iy -831.825256
3800000 5100000 d -1246.907349
5100000 5100000 sp -0.215807
5100000 5400000 dh -315.434052 that

[1] The acoustic model used was trained and tested by Michail Maragakis as part of his MSc[35].

5400000 6000000 ae -664.818726
6000000 6300000 t -333.542114
6300000 6300000 sp -0.215807
6300000 6800000 n -566.184448 note
6800000 9400000 ow -2504.242188
9400000 11600000 t -1999.376099
11600000 11700000 sp -72.969818

This shows the start and end time of each word and the total log probability. For example this sentence's duration including the silence fragment at the end is 117 frames, i.e. 1170 msec or 1.17 sec (10msec/frame). We were now able to calculate the desired duration metrics.

The segmentation information was obtained using the HTK speech recognition software; the hand-labeled word-level transcriptions of the corpus and a context-independent phone-level acoustic model (trained on the whole spontaneous corpus) were used for this purpose.

5.3.3 Phone Durations

We computed the duration of each one of the ten monophthongal vowels ('aa', 'ae', 'ah', 'ao', 'eh', 'er', 'ih', 'iy', 'uh', 'uw') averaged per age and gender group. We also computed the average duration of all these vowels.

5.3.4 Sentence Duration

We computed the duration of all the sentences in the corpus with and without the silence segments at the end of each sentence and averaged this numbers per age and gender group. We also computed the averaged per age-gender group sentence duration (with and without the silence segments at the end of each sentence) for five dialog states: "Goodbye", "Talk2Him", "TellmeAbout", "Travel", "WhereDid". These are the dialogue states with the largest number of utterances.

Furthermore we computed the averaged sentence duration for specific phrases (with and without the silence segments at the end of each sentence). In order to choose which phrases to process, we divided the corpus in files, one for each phrase and counted the appearances of each phrase, choosing the most common ones. However it seems that we do not have a lot of data and in some cases, we have age and gender groups with no data. A larger corpus is necessary for calculating sentence duration of specific phrases.

5.3.5 Between Word Silence duration

We computed the averaged over age and gender duration of between word silence fragments. However in our corpus the non-zero between-word silence fragments are only 2.4% of all between-word silence fragments (1389/57796).

5.3.6 Speaking Rate excluding Including Silence Fragments

We finally computed the average number of phones /sec, excluding between word silence fragments per age and gender group.

5.4 Fluency Metrics

The manually transcribed corpus contained special labels, indicating disfluencies of the speech, hesitations, filled pauses etc. The most common of these labels are shown in Tbl. 5.1

Label	Definition
[.misp]	mispronunciation
[.brth]	breath noise
[.nspn]	noise
[.hst]	hesitation
[um]	filled pause
[uh]	filled pause
[.lps]	lips sound
-	false start

Table 5.1: Special corpus labels and their meaning.

After listening to the corresponding audio files we decided to consider as **False Starts and Mispronunciations** all the utterances that contained a dash '-' (for example, 'susp-' instead of 'suspect') as well as the utterances containing the label [.misp]. **Hesitations and Filled Pauses** on the other hand were the utterances that contained [.brth], [.nspn], [.hst], [um], [uh], [.lps]

In order to calculate Fluency Metrics, we focus only on the 132 successful sessions, i.e. the ones ending in apprehending the suspect ('Arrest' dialogue state), in order to reduce the effect of hesitations due to poor game playing. We computed the average number of disfluencies and hesitations as percentage of total number of words per utterance. We furthermore computed the percentage of utterances that contained disfluencies and hesitations per age and gender group.

5.5 Linguistic Complexity and Variability

Corpus Preparation
Our corpus included hesitation, filled pauses and disfluency labels, such as [.hst], [.misp], etc. It also included fragments of words containing a dash and indicating false start, e.g susp- (instead of suspect). All the above mentioned were removed before calculating perplexities.

5.5.1 Language Models and CMU Toolkit

In order to calculate Linguistic Complexity and Variability, bigram language models were built using the CMU Statistical Language Modeling Toolkit. In all three language models, we applied Witten-Bell discounting and used back-off weights to compute the

probability of unseen bigrams.

Fig. 5.2 summarizes the usage of CMU for building each language model.

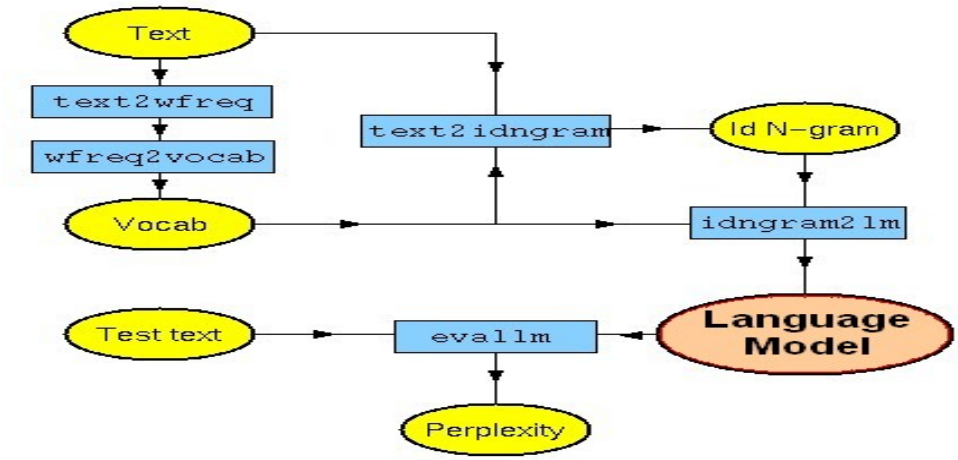


Figure 5.2: CMU Toolkit Usage.

Given a large corpus of text in a file a.text, but no specified vocabulary, CMU

- Computes the word unigram counts
  cat a.text | text2wfreq > a.wfreq
- Convert the word unigram counts into a vocabulary
  cat a.wfreq | wfreq2vocab > a.vocab
- Generate a binary id 3-gram of the training text, based on this vocabulary
  cat a.text | text2idngram -vocab a.vocab > a.idngram
- Convert the idngram into a binary format language model
  idngram2lm -idngram a.idngram -vocab a.vocab -binary a.binlm
- Compute the perplexity of the language model, with respect to some test text b.text
  evallm -binary a.binlm

  An example output of this process would be:

  Reading in language model from file a.binlm
  Done.
  evallm : perplexity -text b.text
  Computing perplexity of the language model with respect to the text b.text

Perplexity = 128.15, *Entropy = 7.00 bits*
Computation based on 8842804 words.
Number of 3-grams hit = 6806674 (76.97%)
Number of 2-grams hit = 1766798 (19.98%)
Number of 1-grams hit = 269332 (3.05%)
1218322 OOVs (12.11%) and 576763 context cues were removed from the calculation.
evallm : quit

CMU built language models were used for calculating Language Model Perplexities and Speaker Linguistic Variabity, while for Levenshtein Distance we used pure PERL programming.

5.5.2 Language Model Perplexities

Our corpus was split into subsets that contained the utterances of each age and gender and the combination of the two. Three different language models were built, with different training-testing subsets of our data:

1. **Train Global - Test Category**: This language model was trained to the whole corpus and tested on each of the subset files. Average perplexities were calculated for each age and gender group.

2. **Train Category - Test Category**: Training and testing of this language model was done on the same data set, i.e. the files mentioned beforehead.

3. **Partial Train - Partial Testing** : 2/3 of each of the above mentioned files was used for training this language model and 1/3 for perplexity calculation. We also repeated this procedure in Round Robin (Leave-one-out cross validation) fashion.

The same procedure was followed for calculating linguistic variability (perplexity) within dialogue state. We split our corpus into subsets that contained the utterances of each age and gender and the combination of the two for each dialogue state,for example 8.Goodbye, 9.Goodbye, etc,10.WhereDid, M.Talk2Him, F.Talk2Him, etc as well as M.10.WhereDid, F.10.WhereDid. The file M.10.WhereDid for example contains the "WhererDid" dialogue state utterances of all the 10 year old males. These files were used for the testing of the language models.The same was done for each age-gender group, 8-9, 10-11, 12-14 We computed perplexities with all 3 models for five dialogue states: "Goodbye", "Talk2Him", "TellmeAbout", "Travel", "WhereDid". However 'Travel" is a rather 'special' dialogue state since it contains a lot of state and cities names. Thus, the measures acquired for this particular dialogue state can not be considered representative [2].

Corresponding to the above mentioned example on the usage of CMU, results acquired for type 2 language model and using as test file the utterances of 14 year old children can be seen below:

[2]Semantic Analysis could be of great use here

evallm : perplexity -text 14
Computing perplexity of the language model with respect to the text 14
Perplexity = 7.58, Entropy = 2.92 bits
Computation based on 6040 words.
Number of 2-grams hit = 5684 (94.11%)
Number of 1-grams hit = 356 (5.89%)
0 OOVs (0.00%) and 0 context cues were removed from the calculation.

5.5.3 Speaker Linguistic Variability

Two language models were developed for calculated Inter- and Intra- vs Inter- Speaker Linguistic Variability. The first one was trained over all user utterances while the second was trained over the utterances of each age group. The testing corpus was the utterances of each speaker within the group. This way Speaker Linguistic Variability was calculated as the average ratio of the bigram language model perplexity of one speaker's utterances vs all speakers utterances (in an age and gender group).The perplexities calculated differed in the absolute value, something totally expected (linguistic variability of the first model is higher), but the age-gender trends ware identical.

5.5.4 Linguistic variability turn to turn

Linguistic Variability turn to turn within a dialogue state was calculated as the Levenshtein Distance between two adjacent utterances in the same dialogue state. *Levenshtein distance* is the minimum number of required transitions (insertions,deletions and substitutions) in order to transform one utterance to another, given that insertions and deletions are given a 0.7 weight while substitutions are given an 1 weight. The Levenshtein distance was calculated by creating an array of transitions costs and selecting the minimum element[51].

We calculated the average Levenshtein distance for all sessions, only successful sessions and only successful ones, within dialogue states where the user makes the same type of requests, e.g. "WhereDid".(This metric was also used in [42]).

5.6 Lexical Metrics

Corpus Preparation
All the disfluencies, hesitations and mispronunciations that were excluded in calculating perplexities were also excluded while computing the lexical metrics. We also focused only on the successful sessions.

5.6.1 Vocabularies

Three type of vocabularies were calculated:

1. Vocabulary Size Total : the total number of words used in each session normalized to the duration of this session (number of utterances)

2. Unique Vocabulary Size : the number of unique words in a session as a percentage of the total number of words in this session

3. Stemmed Unique Vocabulary Size : the same as Unique Vocabulary Size, only that we first reduced inflected words to their stem, base or root form. For stemming, we used the Porter stemming algorithm, implemented in PERL.[41]

5.6.2 Words Per Utterance

The words per utterance were calculated as: The total number of words in each session divided by the total number of utterances in the session and averaged per age and gender group.

5.7 Statistical Significance and ANOVA

In statistics, a null hypothesis (H_0) is a hypothesis set up to be nullified or refuted in order to support an alternative hypothesis. When used, the null hypothesis is presumed true until statistical evidence, in the form of a hypothesis test, indicates otherwise ὁ that is, when the researcher has a certain degree of confidence, usually 95% to 99%, that the data does not support the null hypothesis. It is possible for an experiment to fail to reject the null hypothesis.It is also possible that both the null hypothesis and the alternate hypothesis are rejected if there are more than those two possibilities.

In scientific and medical applications, the null hypothesis plays a major role in testing the significance of differences in treatment and control groups. The assumption at the outset of the experiment is that no difference exists between the two groups (for the variable being compared): this is the null hypothesis in this instance.

The significance level of a test is defined as the probability of making a decision to reject the null hypothesis when the null hypothesis is actually true. The decision is often made using the *p-value*: if the p-value is less than the significance level, then the null hypothesis is rejected. The smaller the p-value, the more significant the result is said to be.

ANOVA is short for analysis of variance. Analysis of variance is a collection of statistical models, and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables.

In practice, there are several types of ANOVA depending on the number of treatments and the way they are applied to the subjects in the experiment:

- One-way ANOVA is used to test for differences among two or more independent groups.

- One-way ANOVA for repeated measures is used when the subjects are subjected to repeated measures; this means that the same subjects are used for each treatment. Note that this method can be subject to carryover effects.

- Factorial ANOVA is used when the experimenter wants to study the effects of two or more treatment variables. The most commonly used type of factorial ANOVA is the $2x2$ (read: two by two) design, where there are two independent variables

and each variable has two levels or distinct values. Factorial ANOVA can also be multi-level such as $3x3$, etc. or higher order such as $2x2x2$, etc. but analyses with higher numbers of factors are rarely done by hand because the calculations are lengthy and the results are hard to interpret. However, since the introduction of data analytic software, the utilization of higher order designs and analyses has become quite common.

- When one wishes to test two or more independent groups subjecting the subjects to repeated measures, one may perform a factorial mixed-design ANOVA, in which one factor is a between subjects variable and the other is within subjects variable. This is a type of mixed effect model.
- Multivariate analysis of variance (MANOVA) is used when there is more than one dependent variable.

We can use N-way ANOVA to determine if the means in a set of data differ when grouped by multiple factors. If they do differ, you can determine which factors or combinations of factors are associated with the difference.N-way ANOVA is a generalization of two-way ANOVA. For example, the two-way ANOVA model for perplexity can be written

$$y_{ijk} = \mu + \alpha_{.j} + \beta_{i.} + \gamma_{ij} + \varepsilon_{ijk} \tag{5.7.1}$$

In this notation

- y_{ijk} is a matrix of perplexity observations (with row index i, column index j, and repetition index k). is a constant matrix of the overall mean gas mileage.
- μ is a constant matrix of the overall mean gas mileage.
- $\alpha_{.j}$ is a matrix whose columns are the deviations of each perplexity measure (from the mean perplexity μ) that are attributable to the age.
- $\beta_{i.}$ is a matrix whose rows are the deviations of each perplexity measure (from the mean perplexity μ) that are attributable to the gender.
- γ_{ij} is a matrix of interactions
- ε_{ijk} is a matrix of random disturbances.

The MATLAB anovan function performs N-way ANOVA. Unlike the anova1 and anova2 functions, anovan does not expect data in a tabular form. Instead, it expects a vector of response measurements and a separate vector (or text array) containing the values corresponding to each factor. This input data format is more convenient than matrices when there are more than two factors or when the number of measurements per factor combination is not constant.

In MATLAB 2-way ANOVA requires the data to be balanced, which in this case means there must be the same number of samples for each combination of age and gender. Since this is not the case, we decided to perform n-way ANOVA, with n=2 and p=0.05 in order to find out whether data from several groups have a common mean.

We also used 'full' model in order to compute the p-values for null hypotheses on the 2 main effects and interactions at all levels.[3]

5.8 Summary

In this chapter we discussed the process by which we performed our analysis of the corpus, the tools used (PERL,CMU,HTK,MATLAB) and how we calculated the desired results for each category, namely: duration statistics, fluency statistics, lexical metrics and language perplexity metrics. The results acquired are presented next.

[3] Although not performed here, a 3-way ANOVA analysis, with success/not success of each session of the game as third factor, might be interesting.

Chapter 6

Acquired Results and Evaluation

6.1 Introduction

In this chapter, we present the most important results for each category and we discuss their significance and their interpretation. A lot more results were acquired, but the ones presented here are the most significant and the ones that, in a way, summarize the general age and gender trends that were spotted during this analysis. An interesting observation when looking at the results altogether is that there are no results contradicting each other. This gives us a certain degree of confidence on the conclusions we were led to. Each figure presented here is accompanied with the corresponding 2-way ANOVA analysis with respect to age and gender group. All statistics presented here are computed for three age groups: 8-9, 10-11 and 12-14, with the exception of vowel duration where data are presented (also) for all ages.

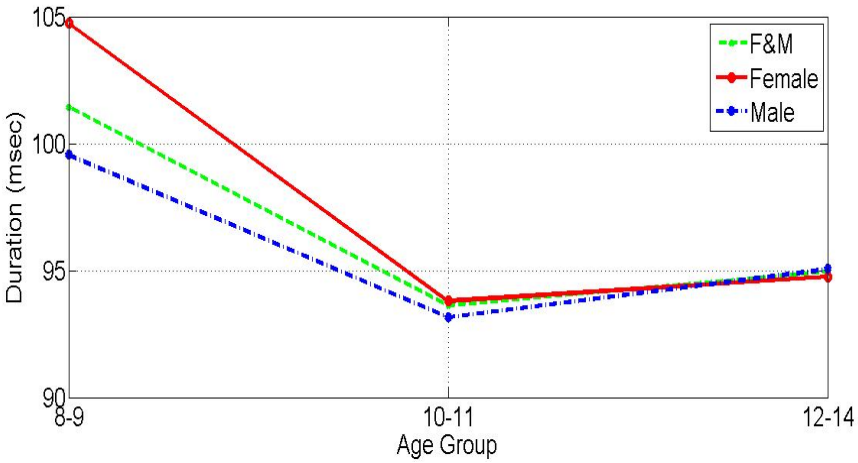
6.2 Duration Metrics

In Fig. 6.1(a), the average vowel phone duration is shown for all age and gender groups. Both the age and the gender trend here are statistical significant, as proved by the corresponding 2-way ANOVA analysis.

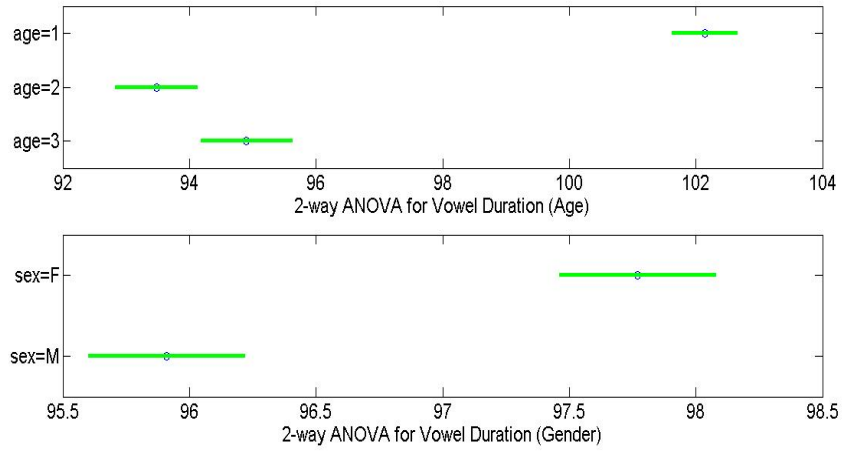
As expected, average vowel duration decreases with age. Specifically, there is a significant reduction in average vowel duration between the younger (8-9) and middle (10-11) age group, and then the duration levels off (and increases somewhat) for the older age group (12-14). The trend is similar for both genders.

The average vowel phone duration for all ages is shown in Fig. 6.2(a) . We should note here however that the amount of data for ages 6 and 7 is very limited; the data points are only shown for completeness.The same conclusions as before can be drawn here, since we can see that the average durations decreases for ages 8 to 11 and then increases for ages 12-14.

In Fig. 6.2(c), results from previous studies on phone duration for 'Read' speech are shown[33]. Although the two trends seem different at first sight, and the phone duration differs significantly for the same age, we should keep in mind that the two plots correspond to different kind of speech (spontaneous vs. read). This is an interesting observation that would be discussed further on.

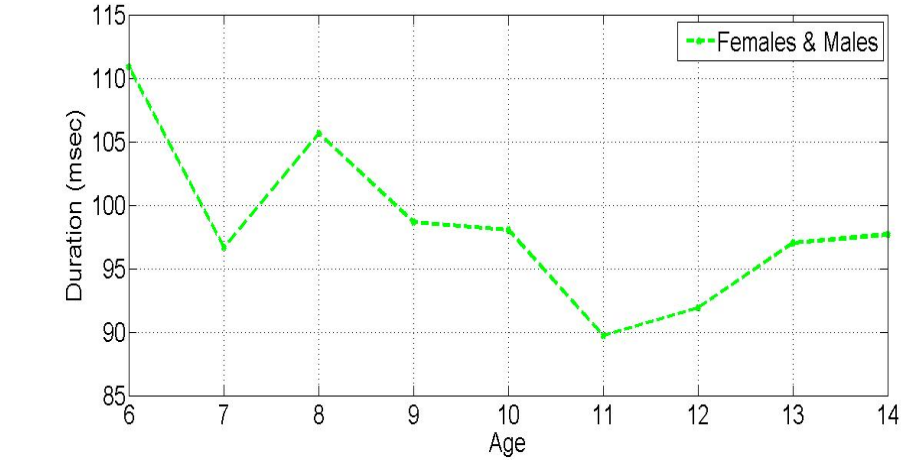


(a)

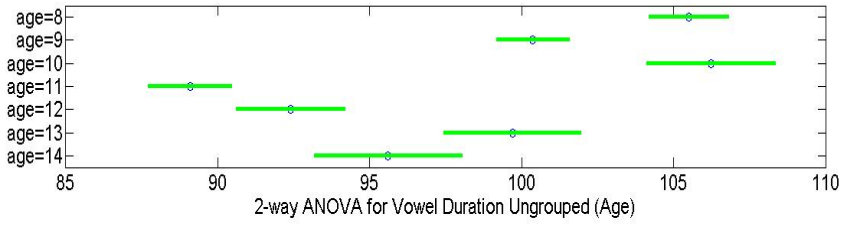


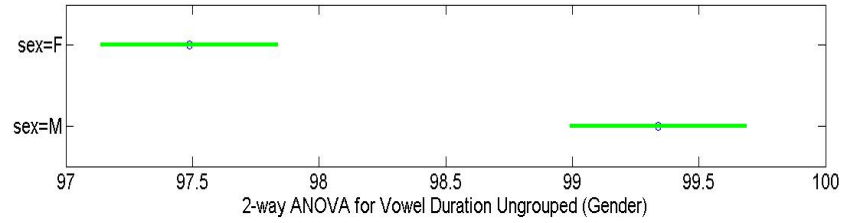
(b)

Figure 6.1: (a) Vowels Duration as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender).

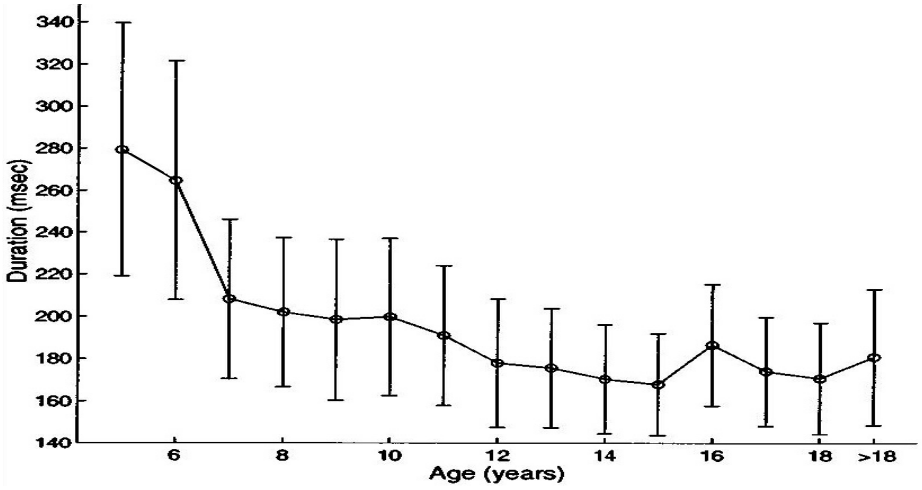


(a)





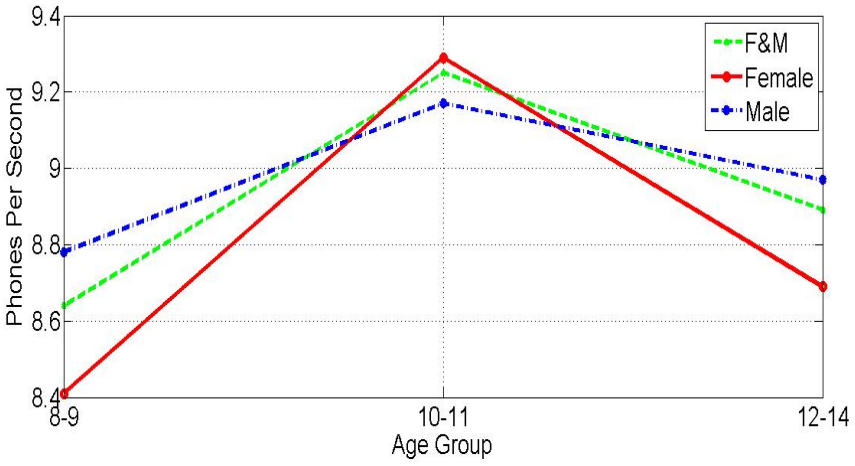
(b)



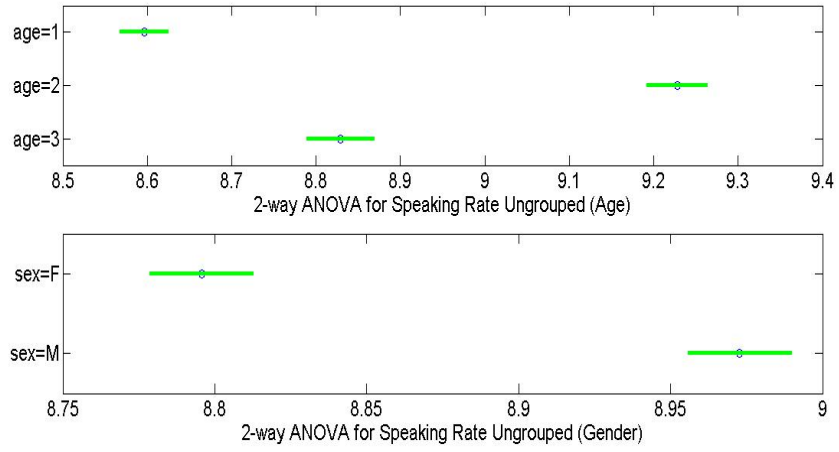
(c)

Figure 6.2: (a) Vowels Duration for all ages as a function of age and gender, and (b)2-way ANOVA results (age-group and gender).

In Fig. 6.3(a), the speaking rate is shown in terms of phones per second (including between word silence fragments) for all age and gender groups. Speaking rate when excluding inter-word silences, although not presented here, increases (as expected) but the trend remains the same. The middle age group (10-11 years) is speaking significantly faster than the younger and older groups. The differences in speaking rate among the age groups is up to 10% for female speakers. It is interesting to note that there is a significant reduction in speaking rate between the age groups 10-11 and 12-14. ANOVA analysis in Fig. 6.3(b) shows that this results are statistically significant both for age and gender.



(a)

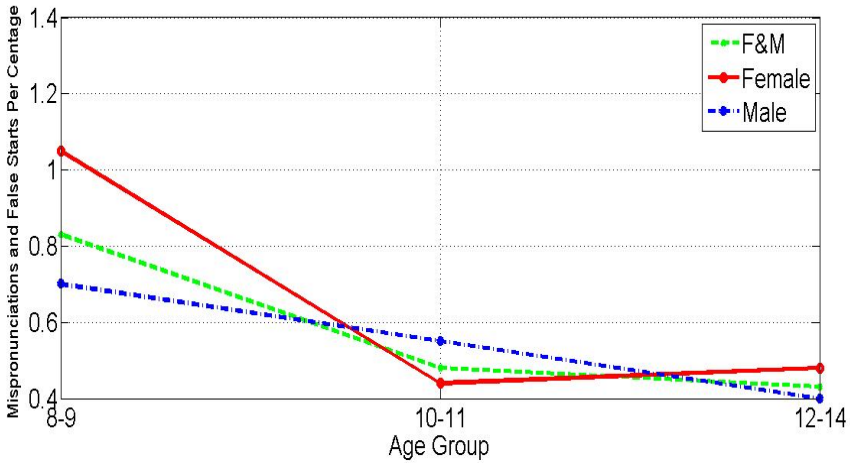


(b)

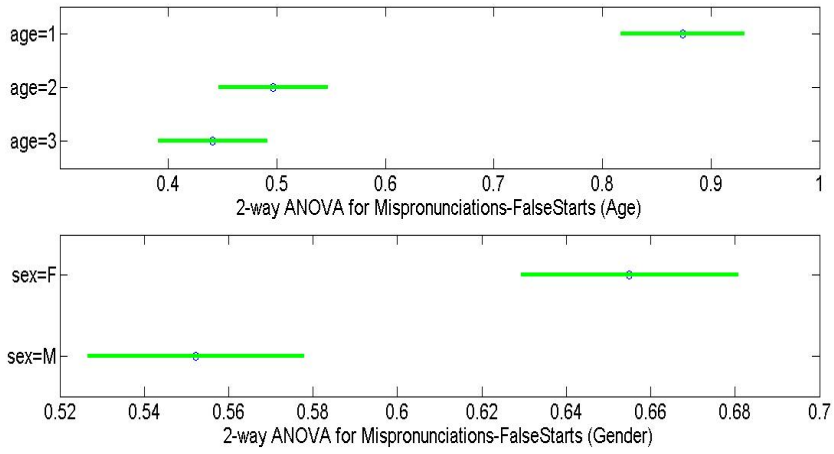
Figure 6.3: (a) Speaking Rate as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender).

6.3 Fluency Metrics

In Fig. 6.4(a), the average number of false starts and mispronounciations are shown as the percent of the total spoken words. Disfluencies decrease as a function of age. Specifically, there is a significant relative reduction of 30% between the 8-9 and 10-11 age groups. The reduction is even bigger for female speakers. Disfluencies decrease from the 10-11 to the 12-14 age groups, especially for male speakers; however, the reduction is not statistically significant.



(a)

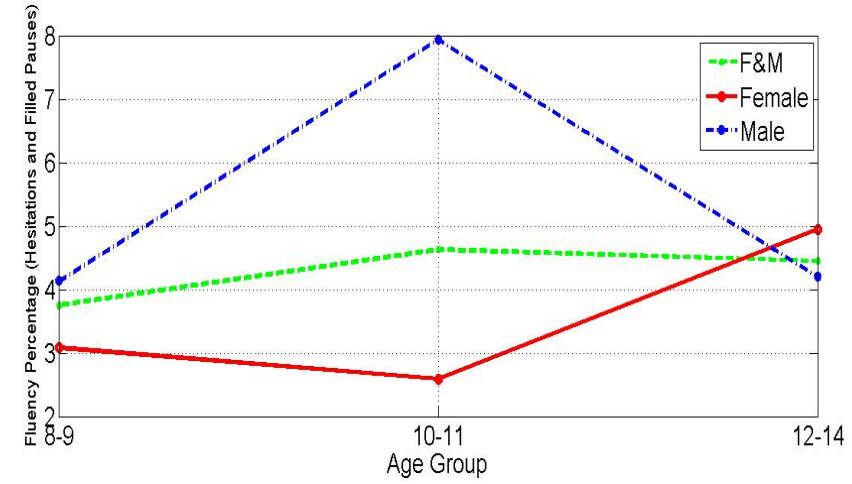


(b)

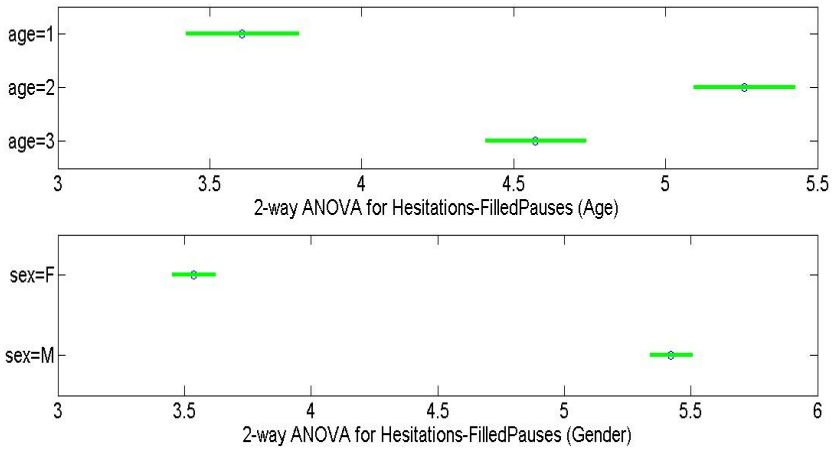
Figure 6.4: (a) False starts and Mispronunciations Per Word Per Utterance as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender).

In Fig. 6.5(a), the average number of hesitations and filled pauses are shown as a percent of the total spoken words. Both the age and the gender trends here are statistically significant as shown in Fig. 6.5(b). Hesitations increase somewhat with age. This is consistent with the observations in [42]. In addition, boys tend to hesitate much more than girls in the 10-11 age group(at least twice as much). When looking at

the breakdown of hesitations vs. filled pauses (not shown here) hesitations in the form of breathing noises are significantly higher for younger children, while filled pauses are much more common for older children (see also [42]). Breathing noises occurred 60% more often for younger children. Surprisingly, this trend was reversed for filled pauses which occurred almost twice as often for the 12-14 age group.



(a)



(b)

Figure 6.5: (a) Hesitations and Filled Pauses Per Words Per Utterance as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender).

6.4 Lexical and Syntactic Metrics

In Fig. 6.6(a), the average number of words per utterance is shown as a function of age group and gender. ANOVA results show that the trends spotted here are clearly statistically significant. Thus we can confirm a clear gender trend here: girls tend to be more verbose than boys after the age of 9 and especially for the older age group. The average number of words per utterance for boys consistently decreases with age, while

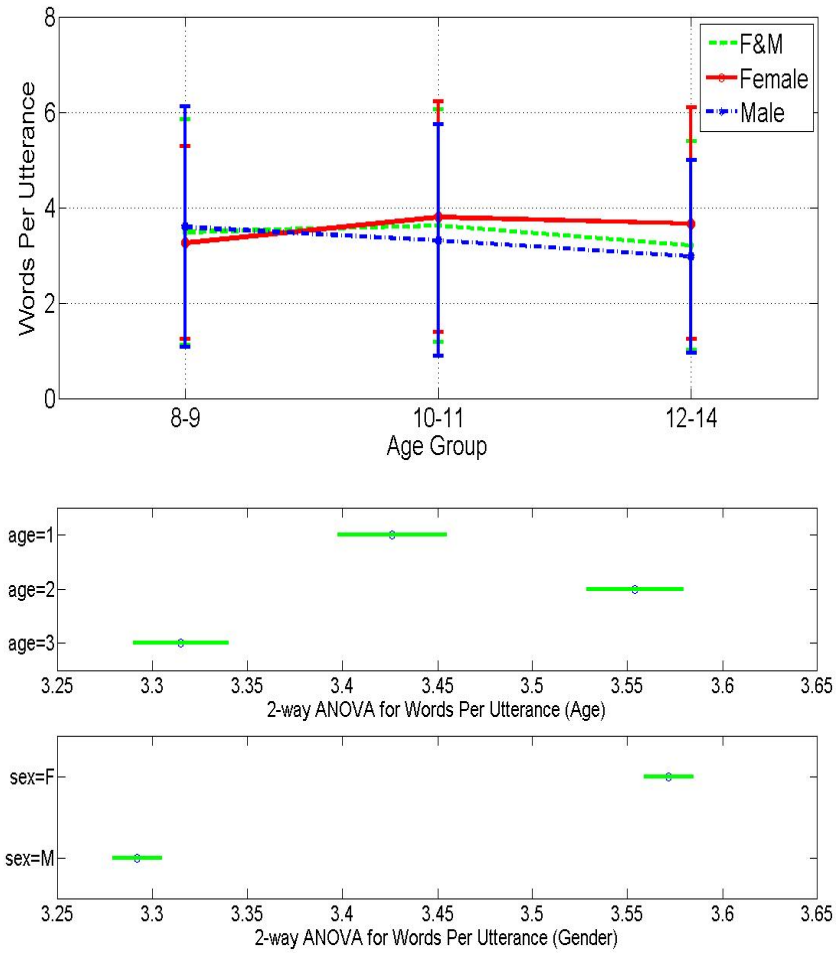


Figure 6.6: (a) Words Per Utterance as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender).

for girls verbosity increases between the age groups 8-9 and 10-11, and then levels off. The relative difference in verbosity between age groups is between 10 and 20%.

In Fig. 6.7(a), the average vocabulary size per session is shown as a function of age group and gender. Although not shown here, the vocabulary trends and statistics where very similar if stemmed words were used instead of word forms. The average vocabulary size tends to increase with age but as ANOVA results in Fig. 6.7(b) show, the age trend is not significant. There is however a statistically significant and maybe unexpected and surprising gender trend: boys have a richer vocabulary than girls for the 8-9 and 10-11 age groups.

Three different measures of linguistic complexity and variability are shown as a function of age group and gender.

Specifically in Fig. 6.8(a), the language model perplexity is shown for language

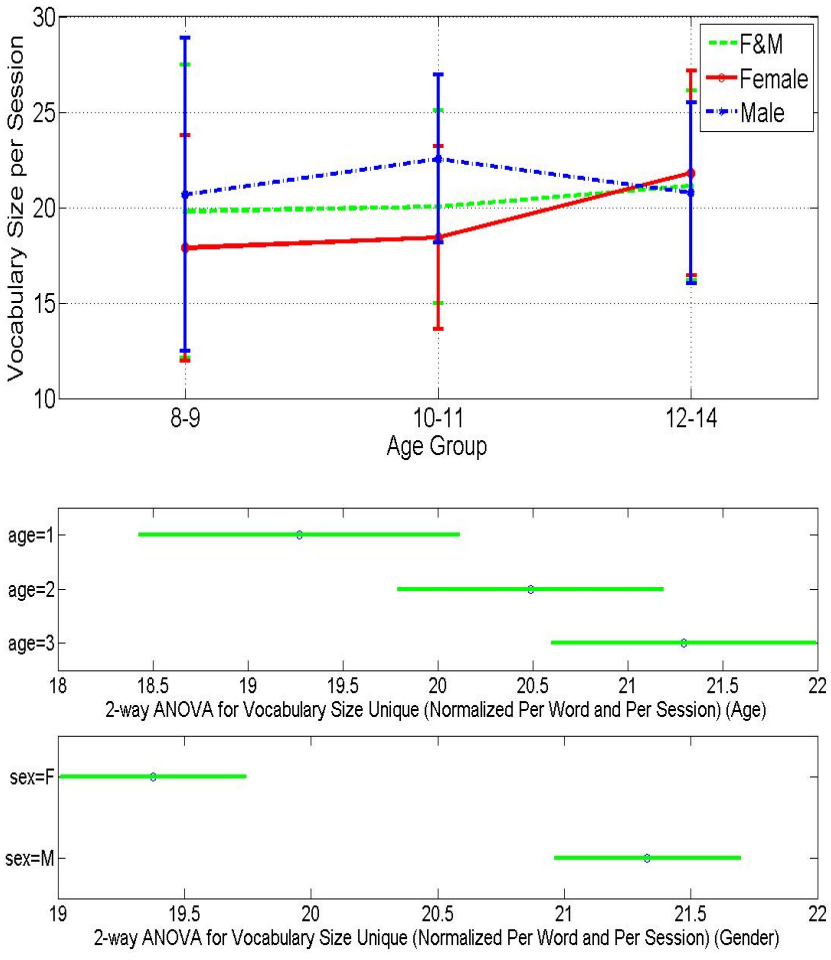
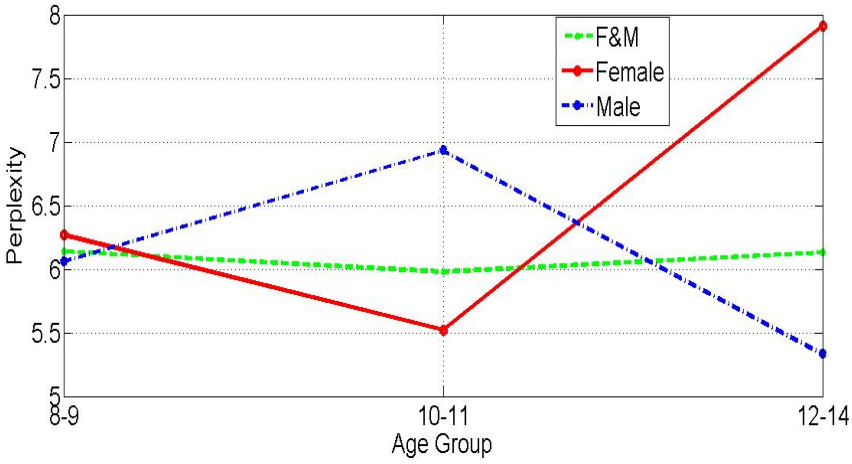


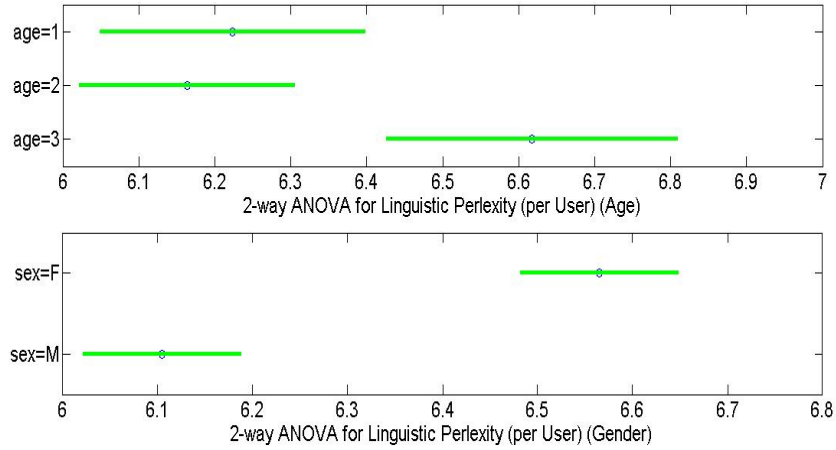
Figure 6.7: (a) Vocabulary Size Unique Per Words Per Session as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender).

models trained on partial data using round-robin (type III language model) for each age and gender group. As analyzed in the previous chapter different language models were used for calculating perplexities.The results acquired differed in the absolute values of perplexities (as the models were trained on different data sets), but the age and gender trends were the same.

In Fig. 6.9(a), inter- and intra-speaker linguistic perplexity is shown for type I models. Error bars depict the inter- (within groups) speaker variability while the plotted curve depicts the intra- (between groups) speaker variability. These results correspond to the LM trained over group utterances. In Fig. 6.10(a), the average Levenshtein distance between two adjacent utterances of the same speaker from the "WhereDid" dialogue state is shown as a function of age and gender. This plot corresponds to perplexities acquired over all sessions not only successful ones. Levenshtein distance for



(a)

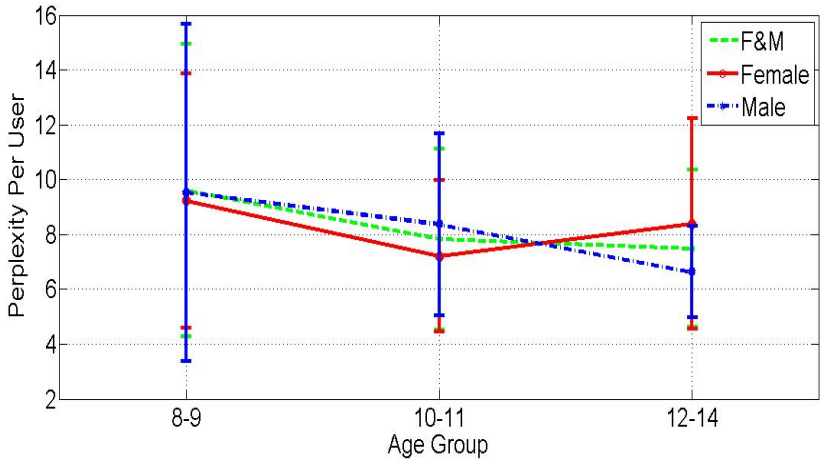


(b)

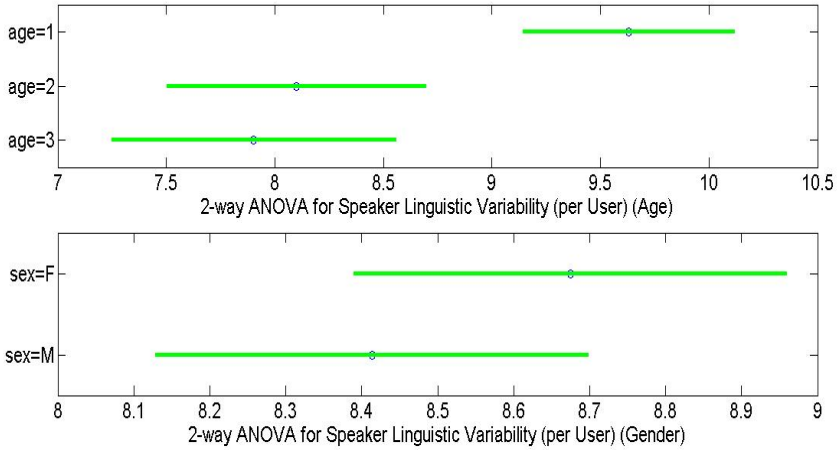
Figure 6.8: (a) Language model perplexity (type II) per user as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender).

only successful sessions was also calculated, but the results acquired were not statistically significant.

As ANOVA results show, not all these trends are statistically significant, however all three linguistic complexity and variability measures follow very similar trends. In Fig. 6.8, statistically significant is the difference in the perplexity for age groups 10-11 and 12-14, in Fig. 6.9 for age groups 8-9 and 10-11 and in Fig. 6.10 for gender groups. Therefore we can conclude that there is reduction of perplexity as a function of age for boys, reduction of perplexity between the 8-9 and 10-11 groups for girls and then a significant increase for the 12-14 age group. The reduction in perplexity between the 8-9 and the 10-11 age groups is larger for girls than for boys, although, this result is not always statistical significant. Note also that the increase in perplexity between the girls aged 10-11 and 12-14 holds both within- and across-speakers.



(c)

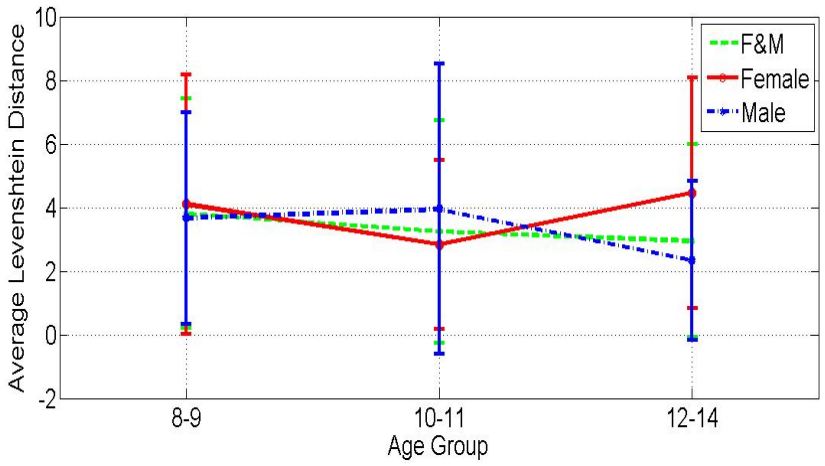


(d)

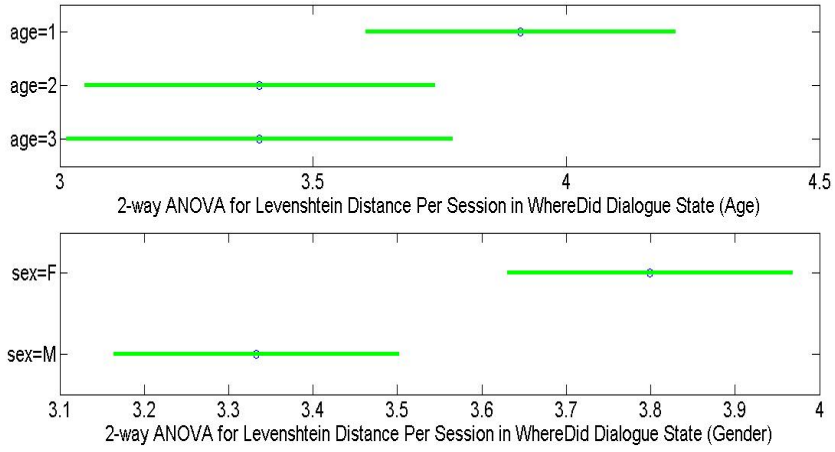
Figure 6.9: (a) Inter- (error-bars) and intra-speaker (plotted curve) language model perplexity (type I) as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender).

6.5 Discussion

Linguistic choices of children are not just a function of age and gender. Various other factors interfere and affect them. Children respond differently to these factors in different ages. Thus to better interpret the results of this study it is important to broaden our view. The game itself for example holds an important role. The task to be accomplished affects the linguistic choices of the children. In general, it is hard to select a task that is both engaging and challenging for a wide range of age groups. In our case, the selected game is rated for children 8 years and older. Results from the exit interview (and the rate of successful games played) indicate that the game is challenging for the 8-9 age group, a good fit for the 10-11 age group, while the game does not provide much of a challenge for the 12-14 age group[39].



(e)



(f)

Figure 6.10: (a) Average Levenshtein distance between two adjacent utterances of the same speaker from the "WhereDid" dialogue state as a function of age and gender, and (b) 2-way ANOVA results (age-group and gender).

A significant difference was noticed between spontaneous and read speech as far as durations and speaking rates are concerned. Comparing the results reported here with the ones reported in [33], we notice that vowels duration are significantly lower in spontaneous speech and speaking rate is higher.

The age trends for spontaneous speech and read speech however are very similar. Between the 8-9 and 10-11 age group there is a 10% relative reduction in average vowel duration. Then from 10-11 to 12-14 the durations seem to level off or even increase somewhat. Children seem to reach adult-level skill at articulation speed around the age of 11 year, and girls seem to be somewhat more adept than boys in the 8-9 age group.

Adult-level skills for read speech are reached around 13-14 years, as claimed in [33]. In this study adult level skills for duration metrics, e.g. speaking rate seem to be reached approximately 2 years later (11-12 in this study).

This can be explained by taking into account the *cognitive load* applied to read speech by the process of reading itself. Extra load on the working memory is imposed while reading, since the mind has to temporary store and manipulate the information. (Read speech can be divided into three stages, read the text, understand it and then speak.) Reading skills evolve with age and gender and affect speaking rate and durations. Thus the durations observed in read speech could be biased by the reading speech of the child and the maturation of the reading skills with age. The higher absolute values for read speech could be explained by the additional cognitive load that reading incurs. However more experiment in larger corpora are needed in order to verify these claims.

The absolute values of mispronounciations and filled pauses are higher in human-machine than in human-human interaction, as reported in the literature [5]. Disfluencies decrease with age and children reach adult-skill level at around 12-13 years of age (somewhat earlier for boys than girls). The age trend is reversed for hesitations. The high number of hesitations for boys aged 10-11 compared to girls of the same age group is hard to interpret and could be due to social reasons. Note that the 10-11 age group is fully engaged by the game and find the game most fun. Further research is needed to interpret this result.

In general, the ability of the children to use language efficiently to achieve a task improves with age for all three age groups. Children use less words per utterance to convey the same message, and, in general, use linguistically simpler constructs as they become more adept with using language over the years. Specifically, note that linguistic variability is reduced with age. Also older children keep repeating linguistic patterns that have been successful at achieving the task at hand (see Levenshtein distance). It is interesting to note that for girls in the 12-14 age-group the linguistic variability increases as does the average sentence length. In fact, sentence length increases also for girls aged 10-11 compared to the 8-9 age group. In general, we conclude that *girls show more linguistic exploration* than boys in the 12-14 age group. This trend seems to emerge around 11 years of age. It is unclear if this trend also correlates with the fact that the game is "easy" for older children, i.e., for girls aged 12-14 game is no longer challenging and thus the opportunity emerges to explore more complex and interesting linguistic patterns. One might conclude that girls ages 12 and older consider language as part of the game not just a tool to successfully complete the game.

6.6 Summary

In this chapter we presented the main results acquired and the ANOVA analysis for testing their statistical significance. We also discussed their meaning and their interpretation in combination with various other aspects of children's speech and children's interaction with computers.

Chapter 7

Conclusions and Future Work

The analysis of acoustic, lexical and linguistic characteristics of spontaneous children's speech has shown significant age and gender trends. Average vowel duration was shown to be significantly lower for spontaneous speech compared with read speech. The age trend (reduction in duration, increase in speaking rate) was similar for read and spontaneous speech, but adult-level values were reached 1-2 years earlier for read speech. The additional cognitive effort that reading imposes on speech duration was claimed to be affecting these statistics, however further research on this field is required. It would be interesting, for example, to test how speaking rate evolves in time, as the cognitive load is reduced and the affect that cognitive load has on durations as a function of time.

Disfluencies decreased with age and leveled off for the 12-14 age group, while hesitations increased with age and where especially pronounced for boys in the 10-11 age group. Older children used simpler linguistic constructs and shorter utterances to complete the task. An important finding was that girls showed significantly more linguistic exploration than boys, as was evident, by the increase in average sentence length and linguistic perplexity for the 12-14 age group.It is clear that girls, as soon as they get adept to the game, consider language as part of the game rather than just a tool to complete the game, as boys do. There is a clear difference here between exploiting the language (in order to complete a task) and exploring the language. Another interesting and rather surprising finding was that boys seem to have a richer vocabulary than girls for the 8-9 and 10-11 age groups. This finding should be further investigated and could also be related to the game itself.

Future research directions include factor analysis for the acoustic and linguistic measures. The emotional state of the child playing the game also affects the characteristics of the speech and has to be investigated. Child frustration, politeness or even neutral attitudes has to be taken into account. Moreover the exit interviews of the children that played the game will be investigated in order to study the effect user satisfaction has on the characteristics of the speech.It is important to know not only what was said but also how something was communicated so that the spoken interaction between system and user will be more natural.

Furthermore a semantic and pragmatic analysis will be performed in order to better understand the way children use the language to communicate. Computing statistics such as for the number of semantic attributes filled per utterance are also of great

interest.

Apart from the results presented in this thesis, our research led to a rather large amount of calculated metrics and not all of them were investigated thoroughly and in detail. Examining these results more closely can possibly lead to additional observations regarding linguistic characteristics of children speech.

Children 's speech analysis and especially spontaneous children speech is a scientific area still developing and open. Since interactive dialogue systems for children are widely used nowadays and their use increases continuously, research in this field is both extremely interesting and important in order to improve both the acoustic and linguistic models of children speech as well as the interface of such applications. Moreover adaptive interfaces that change according to age and experience of the child, interfaces with 'intelligence' and personality, as well as interfaces specifically designed for children with disabilities based to their special linguistic characteristics are scientific directions of enormous interest.

At this point, there is only one thing to say: To be continued...

Bibliography

[1] A. Papoulis and U.S. Pillai, "*Probability, Random Variables and Stohastic Processes.*" McGraw-Hill, 2002.

[2] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan, "Politeness And Frustration Language In Child-Machine Interactions," in *Proc. European Conf. on Speech Communications and Technology*, (Aalborg, Denmark), 2001.

[3] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, and D. Giuliani, "The PF-STAR Children's Speech Corpus," in *Proc. of Interspeech*, (Lisbon, Portugal), 2005.

[4] A. Batliner, C. Hacker, S. Steidl, E. Noth, S. Arcy, M. Russel, and M. Wong, "You Stupid Tin Box - Children Interacting With The AIBO Robot: A Cross-Linguistic Emotional Speech Corpus," in *Proc. of the 4th Intern. Conf. of Language Resources and Evaluation*, (Lisbon, Portugal), 2004.

[5] L. Bell, J. Boye, J. Gustafson, M. Heldner, A. Lindstron, and M. Wiren, "The Swedish NICE Corpus - Spoken Dialogues Between Children And Embodied Characters In A Computer Game Scenario," in *Proceedings of Interspeech*, (Lisbon, Portugal), October 2005.

[6] L. Bell and J. Gustafson, "Children's convergence in referring expressions to graphical objects in a speech-enabled computer game," in *Proc. of Interspeech*, (Antwerp, Belgium), 2007.

[7] C.D. Manning and H. Schutze, "*Foundations of Statistical Natural Language Processing.*" The MIT Press, 2000.

[8] C.E. Shannon, "*A Mathematical Theory of Communication.*" The Bell System Technical Journal, Vol. 27, 1948.

[9] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Clow, and I. Smith, "The Efficiency Of Multimodal Interaction : A Case Study," in *Proc. ICSLP 98*, (Sydney, Australia), 1998.

[10] D. Jurafsky and J.H Martin, "*Speech and Language Processing.*" Prentice Hall, 2000.

[11] S. Eguchi and I. Hirsh, "Development of speech sounds in children," *Acta Oto-Laryngologica*, vol. 257, pp. 1–51, 1969.

[12] M. Eskernazi, "Kids: A database of children's speech," *Journal of the Acoustical Society of America*, vol. 100, 1996.

[13] A. Esposito, "Children's organization of discourse structure through pausing means," *Lecture Notes In Computer Science*, no. 3817, pp. 108–115, 2005.

[14] A. Esposito, M. Marinaro, and G. Palombo, "Children speech pauses as markers of different discourse structures and utterance information content," in *From Sound To Sense*, (MIT), June 2004.

[15] F. Jelinek, "*Methods for Speech Recognition.*" MIT Press, 1997.

[16] F. Jelinek and L.R. Mercer, "*Interpolated Estimation of Markov Source Parameters from Sparse Data.*" In: Proc. Workshop on Pattern Recognition in Practice, 1980.

[17] V. Farantouri, A. Potamianos, and S. Narayanan, "Linguistic Analysis of Spontaneous Children Speech," in *Child, Computer and Interaction Workshop, ICMI*, (Chania, Crete), 2008.

[18] F.P. Brown and A.S. Della Pietra and J.V. Della Pietra and C.J. Lai and L.R. Mercer, "*An Estimate of an Upper Bound for the Entropy of English.*" Computational Linguistics. Vol.18, 1992.

[19] F.S. Chen, "*Probabilistic Models for Natural Language.*" Ph.D. thesis, Harvard University, 1996.

[20] F.S. Chen and J. Goodman, "*An Empirical Study of Smoothing Techniques for Language Modeling.*" In: Proc. Thirty-Fourth Annual Meeting of the Association for Computational Linguistics, 1996.

[21] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, pp. 847–860, October 2007.

[22] G.J. Lidstone, "*Note on the General Case of the Bayes-Laplace Formula for Inductive or a posteriori Probabilities.*" Transactions of the Faculty of Actuaries, 1920.

[23] U. G. Goldstein, *An Articulatory Model For The Vocal Tracts Of Growing Children.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, June 1980.

[24] H. Jeffreys, "*Theory of Probability.*" Clarendon Press, 1948.

[25] H. Ney and U. Essen and R. Kneser, "*On Structuring Probabilistic Dependencies in Stochastic Language Modeling.*" Computer Speech and Language, 1994.

[26] A. Hagen, B. Pellom, and R. Cole, "Children's Speech Recognition With Application To Interactive Book And Tutors," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003.

[27] H.I. Witten and C.T. Bell, "*The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression.*" IEEE Transactions on Information Theory, 1991.

[28] I.J. Good, "*The Population Frequencies of Species and the Estimation of Population Parameters.*" Biometrika, 1953.

[29] E. M. Iosif, "Unsupervised induction of semantic classes using semantic similarity metrics," Master's thesis, Technical University of Crete, July 2007.

[30] R. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *Journal of Speech and Hearing Research*, vol. 19, pp. 421–447, 1976.

[31] R. Kent and L. Forner, "Speech segment durations in sentence recitations by children and adults," *Journal of Phonetics*, vol. 8, pp. 157–168, 1980.

[32] H. J. Ladegaard and D. Bleses, "Gender differences in young children's speech: the acquisition of sociolinguistic competence," *International Journal of Applied Linguistics*, vol. 13, pp. 222–233, October 2003.

[33] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children speech: Developmental changes of temporal and spectral parameters," *JASA*, vol. 105, pp. 1455–1468, March 1999.

[34] L. B. Leonard, M. E. Fey, and M. Newhoff, "Phonological considerations in children's early imitative and spontaneous speech," *Journal of Psycholinguistic Research*, vol. 10, pp. 123–133, March 1981.

[35] M. Maragakis, "Region-based vocal tract length normalization for automatic speech recognition," Master's thesis, Technical University of Crete, 2008.

[36] J. Mostow, A. Hauptmann, and S. Roth, "Demonstration of a reading coach that listens," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, 1995.

[37] M.S. Katz, "*Estimation of Probabilties from Sparse Data for the Language Model Component of a Speech Recognizer.*" IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.35, Iss.3, 1987.

[38] M.T. Cover and J.A. Thomas, "*Elements of Information Theory.*" John Wiley, 1991.

[39] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Trans. on Speech and Audio Processing*, vol. 10, pp. 65–78, February 2002.

[40] M. Perakakis and A. Potamianos, "A study in efficiency and modality usage in multimodal form filling systems," *IEEE Transactions on Audio, Speech and Language Processing*, 2008.

[41] M. Porter, "An algorithm for suffix stripping," *Electronic Library And Information Systems*, vol. 14, no. 3, pp. 130–137, 1980.

[42] A. Potamianos and S. Narayanan, "Spoken dialog systems for children," in *Proc. Internat. Conf. on Acoust., Speech, and Signal Process (ICASSP)*, (Seattle, Washington), pp. 197–201, May 1998.

[43] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, November 2003.

[44] A. Potamianos and S. Narayanan, "A review of the acoustic and linguistic properties of children's speech," in *Proc. Intern. Workshop on Multimedia Signal Processing (MMSP)*, (Chania, Greece), October 2007.

[45] P.R. Clarkson and R. Rosenfeld, "*Statistical Language Modeling Using the CMU-Cambridge Toolkit.*" In: Proc. 5th European Conference on Speech Communication and Technology, 1997.

[46] R. Kneser and H. Ney, "*Improved Backing-off for N-gram Language Modeling.*" In: Proc. International Conference on Acoustics, Speech and Signal Processing, 1995.

[47] R.L. Bahl and F. Jelinek and L.R. Mercer, "*A Maximum Likelihood Approach to Continuous Speech Recognition.*" IEEE Transactions on Pattern Analysis and Machine Intelligence, 1983.

[48] M. Russell, B. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker, "Applications of automatic speech recognition to speech and language development in young children," in *Internat. Conf. Speech Language Processing*, (Philadelphia, PA), October 1996.

[49] R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward multimodal human computer interface," *Proc. IEEE*, vol. 86, pp. 853–869, 1998.

[50] E. Strommen and F. Frome, "Talking back to big bird: Preschool users and a simple speech recognition system," *Educational Technology Research and Development*, vol. 41, pp. 5–16, 1993.

[51] B. Thomson-McInnes, "Levenshtein distance, users.cs.umn.edu/ bthomson/tools/programs/."

[52] M. Tomasello, "The item-based nature of children's early syntactic development," *Trends in Cognitive Sciences*, vol. 4, no. 4, pp. 156–163, 2000.

[53] W.A. Gale and K.W. Church, "*Estimation Procedures for Language Context: Poor Estimates are Worse Than None.*" In: Proc. 9th Symposium in Computational Statistics, 1990.

[54] W.A. Gale and K.W. Church, "*What's Wrong with Adding One?.*" Corpus-Based Research into Language, Rodolpi, 1994.

[55] W.E. Johnson, "*Probability: Deductive and Inductive Problems.*" Mind, 1932.

[56] B. Xiao, C. Girand, and S. Oviatt, "Multimodal integration patterns in children." Proc. of the 7th Intern. Conf. on Spoken Language Proc., 2002.

[57] S. Yildirim, C.Lee, S. Lee, A. Potamianos, and S. Narayanan, "Detecting politeness and frustration state of a child in a conversational computer game," in *Proc. Intl. Conf. on Speech Communication and Technology (INTERSPEECH)*, (Lisbon, Portugal), September 2005.

[58] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Olasson, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2).* Cambridge University Engineering Department, 2002.