Clear Speech

Technologies that Enable the Expression and Reception of Language

Synthesis Lectures on Assistive, Rehabilitative, and Health-Preserving Technologies

Editor
Ronald M. Baecker, ***University of Toronto***

Advances in medicine allow us to live longer, despite the assaults on our bodies from war, environmental damage, and natural disasters. The result is that many of us survive for years or decades with increasing difficulties in tasks such as seeing, hearing, moving, planning, remembering, and communicating.
This series provides current state-of-the-art overviews of key topics in the burgeoning field of assistive technologies. We take a broad view of this field, giving attention not only to prosthetics that compensate for impaired capabilities, but to methods for rehabilitating or restoring function, as well as protective interventions that enable individuals to be healthy for longer periods of time throughout the lifespan. Our emphasis is in the role of information and communications technologies in prosthetics, rehabilitation, and disease prevention.

Clear Speech: Technologies that Enable the Expression and Reception of Language
Frank Rudzicz
2016

Designed Technologies for Healthy Aging
Claudia B. Rebola
2015

Fieldwork for Healthcare: Guidance for Investigating Human Factors in Computing Systems
Dominic Furniss, Rebecca Randell, Aisling Ann O'Kane, Svetlena Taneva, Helena Mentis, and Ann Blandford
2014

Fieldwork for Healthcare: Case Studies Investigating Human Factors in Computing Systems
Dominic Furniss, Aisling Ann O'Kane, Rebecca Randell, Svetlena Taneva, Helena Mentis, and Ann Blandford
2014

Interactive Technologies for Autism
Julie A. Kientz, Matthew S. Goodwin, Gillian R. Hayes, and Gregory D. Abowd
2013

Patient-Centered Design of Cognitive Assistive Technology for Traumatic Brain Injury Telerehabilitation
Elliot Cole
2013

Zero Effort Technologies: Considerations, Challenges, and Use in Health, Wellness, and Rehabilitation
Alex Mihailidis, Jennifer Boger, Jesse Hoey, and Tizneem Jiancaro
2011

Design and the Digital Divide: Insights from 40 Years in Computer Support for Older and Disabled People
Alan F. Newell
2011

Copyright © 2016 by Morgan & Claypool

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Clear Speech: Technologies that Enable the Expression and Reception of Language

Frank Rudzicz

www.morganclaypool.com

ISBN: 9781627058261 paperback
ISBN: 9781627058278 ebook

DOI 10.2200/S00672ED1V01Y201509ARH008

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON ASSISTIVE, REHABILITATIVE, AND HEALTH-PRESERVING TECHNOLOGIES

Lecture #8
Series Editor: Ronald M. Baecker, *University of Toronto*
Series ISSN
Print 2162-7258 Electronic 2162-7266

Clear Speech

Technologies that Enable the Expression and Reception of Language

Frank Rudzicz
Toronto Rehabilitation Institute and
Department of Computer Science, University of Toronto

SYNTHESIS LECTURES ON ASSISTIVE, REHABILITATIVE, AND HEALTH-PRESERVING TECHNOLOGIES #8

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Approximately 10% of North Americans have some communication disorder. These can be physical as in cerebral palsy and Parkinson's disease, cognitive as in Alzheimer's disease and dementia generally, or both physical and cognitive as in stroke. In fact, deteriorations in language are often the early hallmarks of broader diseases associated with older age, which is especially relevant since aging populations across many nations will result in a drastic increase in the prevalence of these types of disorders. A significant change to how healthcare is administered, brought on by these aging populations, will increase the workload of speech-language pathologists, therapists, and caregivers who are often already overloaded.

Fortunately, modern speech technology, such as automatic speech recognition, has matured to the point where it can now have a profound positive impact on the lives of millions of people living with various types of disorders. This book serves as a common ground for two communities: clinical linguists (e.g., speech-language pathologists) and technologists (e.g., computer scientists). This book examines the neurological and physical causes of several speech disorders and their clinical effects, and demonstrates how modern technology can be used in practice to manage those effects and improve one's quality of life. This book is intended for a broad audience, from undergraduates to more senior researchers, as well as to users of these technologies and their therapists.

KEYWORDS

computational linguistics, speech-language pathology, assistive technologies, rehabilitation science, machine learning

Dedicated to those overcoming barriers in communication

Contents

Preface

When I was a grad student, the purpose of my research was to improve the accuracy of speech recognition software for people with speech disorders. I started by working with cerebral palsy (CP), which remains the most common cause of hard-to-understand speech today. Most people with CP were not very well understood by speech recognition at the time—less than 1% of their words could be correctly recognized whereas a speaker without a speech disorder might be comfortably understood 85% or 90% of the time. It wasn't that their words didn't make sense—people with CP can normally understand and produce *language* just fine—it was that their voices are quite different from those of the general population, which can profoundly confuse speech software. It was my job to un-confuse the software.

Not being understood *almost all of the time* can be annoying in itself—and speech recognition certainly did a dismal job for people with CP. It was therefore perhaps somewhat frustrating that speech was often the *most effective* means of communication these individuals had. Although CP limits the control of the muscles of speaking (e.g., the tongue), CP *also* affects other muscles (e.g., those controlling the fingers). This means that while speech in CP can be approximately three times slower than typical speech, typing can be over a *hundred* times slower.

So if a computer can't understand what you say *and* it takes too long or is too difficult to type by hand, then merely participating in our modern society becomes a tremendous challenge. According to the U.S. Census bureau, less than 10% of people with severe disabilities are employed, partially due to difficulty in communication, which has considerable consequences for social and health well-being.

Something must be done.

So how could I make my own small dent toward cracking this huge problem? Since the *sounds* of speech in cerebral palsy were so difficult for computers to understand, I reasoned that it might help to "teach" the computer *why* those sounds were difficult—to teach it about differences in the *physical* origins of speech. How do you teach a computer? These days, we use MACHINE LEARNING where you basically program the computer to find patterns and relationships in data by itself, typically given lots of carefully curated examples that you provide. In my case, I needed to provide examples of speech sounds and their corresponding vocal tract movement, and for that I needed participants to come into the lab to have their voices and facial movements recorded during speech.

Many of the participants were in their early twenties and came in with their parents or other caregivers. One young man with CP was particularly talkative, and his father was equally eager to insert himself into the conversation, usually to repeat or to clarify what his son said. They were both very outgoing, and we had about as non-serious a chat as you can imagine in a research

setting, in the basement of a satellite building of the University of Toronto. At one point, the young man revealed that one of his main motivations for volunteering (and for getting his dad to take time off of work to drive him into the lab), was "girls." I told him that was not part of our research protocol. This young man's father then chimed in to say that it wasn't so much "girls" as it was a *particular* girl, and that she and his son were "courting,"[1] but communication between them remained difficult. The young man had tried a number of devices and programs to help him be understood, but he found each of them to be insufficient—he didn't feel like he could properly express himself. The alternative to talking through a computer was to talk through the filter of your father, which can also be non-ideal in courtship. He wanted to help us improve the technology.

Can advanced speech technology improve your love life? More data is required. However, what was clear to me from that exchange was that so much of who we are, collectively and as individuals, depends on our ability to communicate. Language is not just about communicating facts or making plans—to a large extent it defines how others perceive us and how we perceive ourselves in the world. Being able to define yourself in your own words—to speak for yourself—is liberating.

I hope that this book can bring together people who really should be talking together, especially technologists, therapists and clinicians, and people affected by speech disorders. Technologists need to know what challenges exist in the real world and how clinicians are currently meeting those challenges. Therapists need to know how artificial intelligence that can help to diagnose, monitor, and overcome issues of communication. Perhaps most importantly, people affected by speech and language disorders need to know that there is light at the end of the tunnel, and that technology is helping to provide that light.

> [Language has a] unique role in capturing the breadth of human thought and endeavour...We look back at the thoughts of our predecessors, and find we can see only as far as language lets us see. We look forward in time, and find we can plan only through language. We look outward in space, and send symbols of communication along with our spacecraft, to explain who we are, in case there is anyone there who wants to know. [Crystal, 1998]

Frank Rudzicz
February 2016

[1] Is that what kids do these days?

Figure Credits

Figure 3.1 Used with permission from Microsoft © 2016.

Figure 5.2 Based on: Paul Boersma. Functional Phonology: Formalizing the interactions between articulatory and perceptual drives. Ph.D. thesis, Universiteit van Amsterdam, September 1998.

Figure 5.3 Based on: Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken, and Janice Staehely. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49(9):743–759, September 2007.

Figure 6.2 Keith L. Moore and Arthur F. Dalley. *Clinically Oriented Anatomy*, Fifth Edition. Lippincott, Williams and Wilkins, 2005. Copyright © 2005 Lippincott, Williams & Wilkins. Used with permission.

Figure 6.4 Based on: Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.* Prentice Hall, Second edition, 2009.

Figure 8.1 Based on: Fangxin Chen and Aleksandar Kostov. Optimization of dysarthric speech recognition. In *Proceedings of the 19th Annual International Conference of the IEEE*, volume 4, pages 1436–1439. Engineering in Medicine and Biology Society, November 1997.

Figure 8.2 Prasad D. Polur and Gerald E. Miller. Investigation of an HMM/ANN hybrid structure in pattern recognition application using cepstral analysis of dysarthric (distorted) speech signals. *Medical Engineering and Physics*, 28 (8):741–748, October 2006. Copyright © 2006 Elsevier. Used with permission.

Figure 9.1 Used by permission of the Tetra Society of North America.

Figure 9.3a Courtesy of National Public Website on Assistive Technology.

Figure 9.3b Courtesy of Poule, `https://en.wikipedia.org/wiki/Augmentative_and_alternative_communication#/media/File:Minimo.jpg`.

Figure 9.4 Courtesy of © 2016 Copyright PrAACtical AAC.

Figure 9.5 Courtesy of David J.C. MacKay.

CHAPTER 1

Introduction

About one in every ten people in the world, from newborns to the oldest among us, has some communication disorder affecting speech. These disorders can manifest themselves physically (as in reduced control of the muscles of speech in cerebral palsy and Parkinson's disease), cognitively (as in difficulty understanding words in autism and dyslexia), or both physically and cognitively (following, for example, cardiovascular stroke), according to the U.S. National Institutes of Health. These figures are increasing with the age of the population and the incidence of stroke. The Canadian Association of Speech-Language Pathologists and Audiologists estimates that one in ten people are affected to some degree by language impairments and that this proportion will rise significantly over the next decade with the prevalence of cardiovascular stroke expected to rise as populations in various countries become older. In fact, speech and language disorders are present in nearly 85% of those who have experienced stroke and are one of the first symptoms of Alzheimer's disease. This prevalence is especially worrying since aging populations across many nations will result in a drastic increase in speech disorders brought on by age. This will place a tremendous burden on speech-language pathologists, therapists, and caregivers who are often *already* overworked or, in many cases, devoted to language disorders that occur *earlier* in life, such as in cerebral palsy or in developmental delays. This has been referred to as an impending healthcare crisis. At the very least, it will require massive changes in how healthcare is delivered, globally.

Fortunately, modern technology has matured to the point where it can now have a profound positive impact on the lives of millions of people with speech and language disorders. This book serves as common ground for several communities, especially clinical linguists (including speech-language pathologists), and technologists (including computer scientists and engineers). Hopefully, sharing common ground will help to accelerate collaboration in this area. The book is, however, written for a broad audience, from advanced undergraduates and more senior researchers to users of assistive technologies, their families, and their therapists.

Before we continue, we should make a few terminological clarifications. To be properly pedantic, we should distinguish between SPOKEN LANGUAGE (referring to word-, grammar-, and meaning-level aspects of language in spoken utterances), WRITTEN LANGUAGE (referring to those aspects in writing), and SPEECH (referring to acoustics and articulatory aspects of speech acts). While it is important to be cognizant of the differences between these terms, we will occasionally use the term "speech" as a superset of its physical and acoustic properties *and* the linguistic aspects of speech acts. We should also be clear as to the scope of this book. We will not cover all topics in speech and language, naturally. Our focus is on technologies that assist in speech communication.

This will include technologies that can interpret difficult speech and those that can synthesize easy-to-understand speech. Along these lines, we will also discuss the entry of written text to drive those systems.

Part I provides some mathematical and terminological background to help interpret the rest of the book. Not all of Part I will be applicable to you, but if you are a technologist missing a background in linguistics, or a linguist without experience in modern machine learning, these chapters will at least help you to communicate in the same language (so to speak) as the other people in the room. Part II covers the NEUROLOGY AND ANATOMY of speech and language; to a large extent, this covers information about how the brain processes and produces language, and what can go wrong in the vocal tract and hearing mechanism. While the focus will be on speech, we will also take a brief foray into cognitive disorders affecting language comprehension.

Part III covers TECHNOLOGIES THAT ENABLE, especially those that speak for people with impairments of speech production (including eye-typing and word prediction), and those that help interpret for those with impairments of speech reception (including cochlear implants and other hearing aids). This will be a whirlwind tour of this area of research and will in many ways only scratch the surface. You are therefore invited to follow the various citations and references that will be provided throughout these sections.

PART I

Background

CHAPTER 2

Math & Stats for Language Technology

Like so much of human behavior, language opens itself up to formal analysis by mathematics and statistical processes. This includes everything from describing the motions of the physical articulators to mimicking auditory processes in artificial neurons. In order to make meaningful progress in our field of research, we often require a thorough familiarity with various statistical tools. This chapter surveys mathematical approaches that are relevant to certain subareas within speech and language processing in a fairly introductory manner, using examples relating to our domain. By no means is this survey exhaustive—for almost every tool we discuss, you will find more intricate varieties that can be more suitable to your task, so you are encouraged to dig deeper by following the included references and by doing your own research.

We begin by discussing basic probability theory, which is a central component in modern computer systems of language, which use statistics to make interpretations. Probability theory is also a central component of INFORMATION THEORY, which concerns the uncertainties present in the transmission of information between abstract producers and receivers of messages. Resolving and modeling these uncertainties using statistical probabilies can be an important component of speech technologies.

2.1 PROBABILITY THEORY

PROBABILITY THEORY deals with representing the likelihood of events. The canonical examples involve games—how likely are you to roll a 4 with a fair 6-sided die? How likely are you to pull the Queen of Hearts from a deck of playing cards? In these examples, we're using probability theory for one of its chief purposes—to assign a likelihood to a particular EVENT. In the fair die example, the event is that a particular side, $A = 4$ will turn up, but there are several *possible* events (in this case, six). We often call the list of all possible events the SAMPLE SPACE or DOMAIN, represented as Ω, where the number of elements in that set are denoted by double bars. For example, if Ω is sides on a die, $\|\Omega\| = 6$.

We can often think of language in similar terms. Imagine that our domain Ω is the set of all English words, of which there are approximately $\|\Omega\| = 250{,}000$.[1] You may have stumbled on it or stepped over it, but there was an AMBIGUITY in the previous sentence, which is an important

[1]According to the Oxford English Dictionary, which generally includes only common, non-slang words. Technical terms, including medical/clinical language, greatly expand this set.

concept in computational linguistics.[2] The ambiguity is around the meaning of the word "*word*". Sometimes, the word "*word*" is synonymous with TERM, which is like an entry in a dictionary of which there are, as estimated, about 250,000 in English. Other times, the word "*word*" means a sequence of characters separated by spaces—an instance of a term—of which there have been countless trillions scribbled down over the centuries. An *instance* of a word is called a TOKEN. For example,

Counting terms Since$_1$ the$_2$ terms$_3$ "the", "in$_4$", "this$_5$", "terms", and$_6$ "sentence$_7$" are$_8$ repeated$_9$ in this sentence, the number$_{10}$ of$_{11}$ *terms* in this sentence is$_{12}$ thirteen$_{13}$.

Counting tokens This$_1$ sentence$_2$, by$_3$ contrast$_4$, has$_5$ seven$_6$ tokens$_7$, no$_8$ wait$_9$, make$_{10}$ that$_{11}$ thirteen$_{12}$ tokens$_{13}$.

We'll return to ambiguity in Chapter 3, so this digression is merely to emphasize that when dealing with probabilities, it's important to carefully define your domain first.

Given a domain, we now have the task of defining how likely a given event is. In the case of a fair 6-sided die, the probability of any one side turning up should be the same for every side, i.e.,

$$P(A = 1) = P(A = 2) = \ldots = P(A = 6) = P(A) = 1/6.$$

To a large extent, research in natural language by computers treats language itself as no more than an $\|\Omega\|$-sided die, which we metaphorically roll each time we write a word on a page (or, more likely these days, a keyboard). Imagine you have a very limited vocabulary of 25 terms, $\Omega = \{the, of, in, house, cat, hat, \ldots, pulchritudinous\}$,[3] where each word is equally likely. This is what we call a "flat" or UNIFORM DISTRIBUTION, so called because the resulting probability distribution of the words, shown in Figure 2.1, is uniformly flat.

Now, naturally not all words in a human language are equally likely. You are far more likely to write the word "the" than you are to write the word "pulchritudinous". Here, the metaphor of language-as-dice must be amended, but it can still work. *Loaded* dice are dice that are weighted assymetrically so that certain sides are more likely to come out on top than others. If you are handed a loaded die and tasked with discovering its BIAS, that is, the probabilities of each side, a good approach would be to roll the die a very large number of times, count the occurrances of each side, and divide each of those counts by the total number of rolls to get the probabilities of each side. The same can be done with language. Using the loaded die metaphor, we can imagine that every word token we see on a page (or on a screen, or uttered by a mouth, etc.) is the result of a "roll of the language die". If an imaginary book has a million word tokens ($N = 1 \times 10^6$) and 40,000 of those are the term "the" (i.e., we *Count*(*the*) forty thousand times), then the PRIOR

[2] Computational linguistics is a field of research that processes natural language using algorithms and statistics, and is discussed in Chapter 3.

[3] **pulchritudinous** *adj.* Beautiful. **Origin** 1910–15 Americanism.

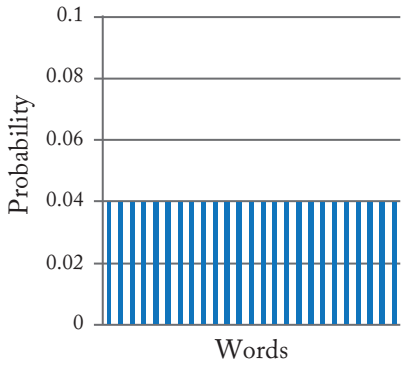


Figure 2.1: A uniform distribution where each bar represents the probability of each of 25 equally likely words, i.e., each word in this limited vocabulary has a 4% probability of being said.

PROBABILITY of a word W being "the" is

$$P(W = the) = Count\,(the)/N = \frac{40,000}{1 \times 10^6} \approx 0.04.$$

As in all probabilistic models, we are bound by the rules of probability. The two chief rules any probabilistic model must obey are:

- The probability of any event A must be between 0.0 and 1.0 inclusive, i.e., $0.0 \leq P(A) \leq 1.0$. If $P(A) = 0$, A is impossible; if $P(A) = 1$, no event *other* than A is possible.
- The sum of probabilities over all events (possible or otherwise) must sum to 1.0, i.e., $\sum_i P(A_i) = 1$.

Since every word token in our imaginary book must count toward exactly one of the terms used in that book, clearly $\sum_W P(W) = 1$.

One can estimate the probability of each word in a collection of text by simply counting the occurrence of each term and dividing by the total number of tokens in that collection. This constitutes a LANGUAGE MODEL, which is a centrally important concept in computational linguistics. At its most basic, it is a representation of the likelihood of words, which permits all manner of applications to be possible, from making more informed guesses in interpreting hard-to-understand speech to predicting the next word someone is trying to type. Representing the probabilities of words helps the computer to make smarter interpretations of language.

Like all models in science, a language model is a set of PARAMETERS that **describes** data that we've seen already and can **predict** future or unseen data. In Ptolemy's geocentric model of the solar system, the parameters were the widths of mystical transparent spheres on which the sun and planets revolved, convolutedly, around the Earth. In our language model, the parameters

are merely the probabilies of each word. Ptolemy's geocentric model could also be used to make predictions, such as the location of Mars in the sky on a particular future evening. Similarly, our simple language model can be used to make predictions of "future words", as we'll discuss in Section 3.1. For example, you could find the probabilities of all of the words in George R.R. Martin's series of novels, *A Song of Ice & Fire*, to predict each subsequent word in the (as yet unpublished) rest of the series. Whether that prediction is *accurate* is another matter altogether, to which we will similarly return in a future section.[4]

2.1.1 MULTIPLE EVENTS

Just as Ptolemy's Earth-centered model of our solar system turned out to be inadequate, so too will our simplistic model of language. Merely finding the probability of a word will encapsulate (almost) nothing with regard to grammar, meaning, or context. It's not possible to learn a language model from the first $N-1$ chapters of a murder mystery in this way, for example, to accurately predict the culprit in the N^{th} chapter.

An account of modern computational theories of syntax, semantics, and pragmatics is beyond the scope of this book. However, for now it suffices to say that, even at our somewhat naïve level of language modeling, we can learn models that are slightly more complex by using more than just one RANDOM VARIABLE. There are many ways we can do this—we are really only limited by our imagination.[5] For example, for any word W, we could model its joint probability with the speaker being in a particular emotional state, E; $P(W = dang, E = angry)$ could be the JOINT PROBABILITY of uttering the word "dang" while simultaneously being angry. We're also not bound to using only two random variables, so long as we obey the laws of probability. Specifically, to generalize our earlier rules:

- The probability of any n joint events $x_1 \ldots x_n$ must be between 0.0 and 1.0 inclusive, i.e., $0.0 \leq P(x_1, \ldots, x_n) \leq 1.0$.
- The sum of probabilities over all events (possible or otherwise) must sum to 1.0, i.e., $\sum_{x_1} \sum_{x_2} \ldots \sum_{x_n} P(x_1, \ldots, x_n) = 1$.

Often, we use this notation not to refer to n events happening *simultaneously* but instead as a sequence of n single events occuring in strict succession. If x_1 is the first word in a sequence, x_2 is the second. Also, with multiple events, we can also introduce new rules. Perhaps the most familiar of these is the CHAIN RULE which is:

$$P(A, B) = P(B \mid A)P(A) \tag{2.1}$$

or, more generally,

$$P(x_1, x_2, ..., x_n) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1, x_2), \ldots, P(x_n \mid x_1, x_2, \ldots, x_{n-1}). \tag{2.2}$$

[4] Can you predict which one?
[5] And data!

When combined with the notion of a joint probability as a probability of a sequence, we can talk about things like:

$$\begin{aligned} P(w_1, w_2, w_3, w_4) &= P(a, long, time, ago) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)P(w_4 \mid w_1, w_2, w_3) \\ &= P(a)P(long \mid a)P(time \mid a,\ long)P(ago \mid a,\ long,\ time) \end{aligned} \tag{2.3}$$

where we describe the probability of reading the sequence *a long time ago* as the probability of seeing "a", times the probability of seeing "long" given that we just read the word "a", times the probability of seeing "time" given that we just read *a long*, and so on. Importantly, you are not bound to read the words left-to-right. Any permutation of reading order is permitted—the probability of a three-word sequence can be the prior probability of the second word times the probability of the first given the second, times the probability of the third given the first two. This allows us to use another new rule of joint probabilities, namely Bayes' rule which states

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{2.4}$$

This relationship[6] is visualized in 2.2 and is not merely provided here for your amusement—Bayes' rule is fundamental to many kinds of modern machine learning algorithms.

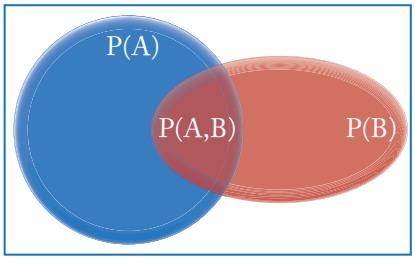


Figure 2.2: Bayes' rule represented in a standard Venn diagram of probabilities over random variables A and B.

2.2 INFORMATION THEORY

At some level, human communication is really about transmitting information, whether explicit words or a spectrum of intentions and emotions. Therefore, it can be useful to draw on mathematical theories collectively called information theory to quantify the *amount* and type of information in a signal. Information theory dates to the end of the second World War and was initially a means to determine how to build in error-correction and redundancy given imperfect communication channels that could corrupt or distort your message [Shannon, 1949].

[6] You can work it out for yourself, using the chain rule, knowing that $P(A, B) = P(A)P(B|A) = P(B)P(A|B)$.

Let's go back to the metaphor of language as a die whose sides are words. Imagine we have an individual who can only say two words: *yes* and *no*—they are about to speak, and you're unsure as to which word they are about to utter. You have a certain *amount* of UNCERTAINTY—a lack of information. Imagine that this entity now utters the word *no*. Your uncertainty is gone and you've received information. How much? If *yes* and *no* are equally likely, then $P(no) = 0.5$ and therefore

$$\begin{aligned} I(E = no) &= \log_2 \frac{1}{P(E = no)} \\ &= \log_2 \frac{1}{\frac{1}{2}} = 1 \text{ bit.} \end{aligned} \tag{2.5}$$

This might be intuitive—it takes a "bit", in computer terms, to encode a binary value. Not so surprising, after all. What if all this entity did was roll 6-sided dice instead? If the die is fair, each side is equally likely. How much information would we receive if the die came up with a 5 (whose probability is $P(E = 5) = 1/6$)?

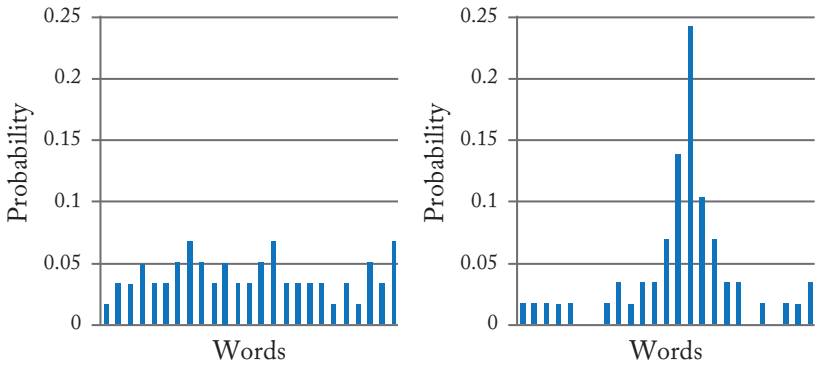
$$\begin{aligned} I(E = 5) &= \log_2 \frac{1}{P(E = 5)} \\ &= \log_2 \frac{1}{\frac{1}{6}} \approx 2.59 \text{ bits.} \end{aligned} \tag{2.6}$$

We receive more information when we observe the roll of a 6-sided die because there are more options—each possible outcome is less likely so we're more "surprised" when we observe a particular value in a sequence. Note that the base of the logarithm, 2 here, is not connected to the number of possible values. We could have chosen any base (e.g., base-10 or the natural base, e)—this value merely determines the units of information we're dealing with. Base-2 gives "bits" of information.

As we've previously discussed, the words in natural languages are not all equally likely. If you're wondering what the next word I might utter might be, you'll be far less surprised if the word is *the* than if it was *octogenarian*.[7] Fortunately, the same formulation above still holds. Even with a vocabulary of 150,000 words, if $P(E = the) = 0.05$, then $I(E = the) = \frac{1}{P(E=the)} = 20$ bits.

In general, the amount of information provided by a single symbol (e.g., word) is not as informative as the *average* amount of information provided by symbols observed from a system or entity over time, which leads us to the concept of ENTROPY. Before we discuss it formally in Section 2.2.1, let's find some entropy. Beyond representing an average amount of information gained, per symbol, for a system, entropy can in many ways characterize a distribution. Figure 2.3 shows two probability distributions over a set of words—one relatively flat and the other relatively "peaked". In many aspects of computational modeling of language, we prefer (in some sense) low entropy distributions because we can be more confident in our predictions—we have less uncertainty about what we will observe next in the sequence, on average.

[7] **octogenarian** *n.* someone who is between 80 and 89 years of age.



(a) High entropy (b) Low entropy

Figure 2.3: Distributions with high (flatter) and low ("peakier") entropy.

Entropy is therefore equivalently:

1. The average amount of information provided by symbols in a vocabulary,
2. The average amount of uncertainty you have before observing a symbol from a vocabulary,
3. The average amount of "surprise" you receive when observing a symbol, and
4. The number of bits needed to communicate that alphabet.

2.2.1 ENTROPY

In many aspects of atypical speech, we will have considerably more statistical uncertainty than in typical speech. As discussed in Chapter 8, this is often the case in motor disorders such as cerebral palsy, where the control of the articulators might be less precise. We may wish to measure and compare the degree of statistical uncertainty in both acoustic and articulatory data for speakers with and without these disorders, as well as the *a posteriori* uncertainty of one type of data given the other. This quantification will inform us as to the relative merits of incorporating knowledge of articulatory behavior into speech technology systems for individuals with these disorders.

Entropy, $H(X)$, is a measure of the degree of uncertainty in a random variable X. When X is discrete, this value is computed with

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

where b is the logarithm base, x_i is a value of X, of which there are n possible, and $p(x_i)$ is its probability. When our observations are continuous, as they are in many acoustic and articulatory

data, we must use *differential entropy* defined by

$$H(X) = -\int_X f(X) \log f(X) dX,$$

where $f(X)$ is the probability density function of X. For a number of distributions $f(X)$, the differential entropy has known forms [Lazo and Rathie, 1978]. For example, if $f(X)$ is a multivariate normal,

$$\begin{aligned} f_X(x_1, \ldots, x_N) &= \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{(2\pi)^{N/2}\,|\Sigma|^{1/2}} \\ H(X) &= \tfrac{1}{2} \ln\left((2\pi e)^N\,|\Sigma|\right), \end{aligned} \tag{2.7}$$

where μ and Σ are the mean and covariances of the data. However, since we observe that both acoustic and articulatory data follow non-Gaussian distributions, we choose to represent these spaces by mixtures of Gaussians. Huber et al. [2008] have developed an accurate algorithm for estimating differential entropy of Gaussian mixtures based on iteratively merging Gaussians and the approximate upper bound of the entropy,

$$\tilde{H}(X) = \sum_{i=1}^{L} \omega_i \left(-\log \omega_i + \tfrac{1}{2} \log\left((2\pi e)^N\,|\Sigma_i|\right)\right),$$

where ω_i is the weight of the $i^{th}(1 \leq i \leq L)$ Gaussian and Σ_i is that Gaussian's covariance matrix. Note that while differential entropies *can* be negative and not invariant under change of variables, other properties of entropy are retained [Huber et al., 2008], such as the chain rule for conditional entropy

$$H(Y \mid X) = H(Y, X) - H(X),$$

which describes the uncertainty in Y given knowledge of X, and the chain rule for mutual information

$$I(Y; X) = H(X) + H(Y) - H(X, Y),$$

which describes the mutual dependence between X and Y. Here, we quantize entropy with the *nat*, which is the natural logarithmic unit, e (≈ 1.44 bits).

These representations are very general, if also a bit technical, and can be useful in a wide variety of contexts. Representing language with this and similar information-theoretical models allows for a number of uses, including explaining how messages can be distorted as they are passed through a medium, such as speech over a telephone wire or Skype. This use, called the NOISY-CHANNEL MODEL, will be discussed in Section 8.1.7, in which we suggest that certain speech disorders can be explained by probabilistic distortions to the control signals that are passed from the brain to the articulators of speech (e.g., the tongue).

CHAPTER 3

(Computational) Linguistics

The field of COMPUTATIONAL LINGUISTICS (CL),[1] in broad terms, concerns getting computers to process human language. This pursuit has taken many forms, each of which has involved different challenges. The goal of creating a CONVERSATIONAL AGENT, in modern times, dates back at least to 1950 when Alan Turing proposed that the best way to determine if a machine *actually thinks* is to have a conversation with it [Turing, 1950] (through textual tele-type, the equivalent of modern text messaging, but the same principle applies).

Computational linguistics, by definition, must naturally be at the center of any system that processes language computationally, which includes tools to help individuals with linguistic disorders. The following subsections (very briefly) introduce core aspects of computational linguistics that are especially relevant in this application area. The first involves predicting words using statistics; while this is often the *goal* of many applications, this simple idea forms the foundation for many other aspects of language processing, including translating texts between languages and speech recognition. The second core component concerns linguistic *features*, which are numerical or quantitative measurements of specific aspects of a piece of text or speech. Accurate measurement of *relevant* features is often essential to the function of the third component, which is machine learning—a topic which extends beyond computational linguistics.

3.1 WORD PREDICTION

WORD PREDICTION is now a technology that almost anyone with a mobile phone has used—it is the automatic presentation of a list of possible continuations given what you've already typed, so you can save a bit of typing. Quite often (though not always!) modern phones will present the correct next word to you given only a few keystrokes. While the ubiquity of this technology may make it appear generic, it has tremendous implications for various populations with communication disorders. This assistive technology has reduced the number of keystrokes required of an individual by as much as $\sim$ 69% in adaptive-lexicon systems [Matiasek et al., 2002, Swiffin et al., 1987], thereby increasing communication speed and allowing improved individual expression [Alm et al., 1992]. Prediction is especially valuable to those for whom fatigue or frustration often accompany attempts at communication [Garay-Vitoria and Abascal, 2006].

[1] The term "computational linguistics" is often synonymous with NATURAL LANGUAGE PROCESSING or HUMAN LANGUAGE TECHNOLOGY, although there are partisans of each camp for whom the distinction, which generally involves the adherence to classical linguistics theory in CL, is paramount.

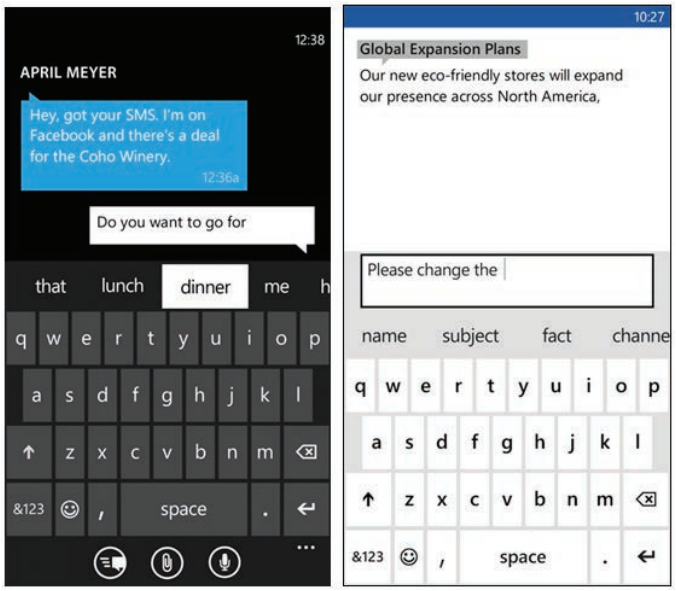


Figure 3.1: Word prediction on a modern mobile touch-screen device. Used with permission from Microsoft © 2016.

Word prediction can be accomplished easily by simply knowing the probabilities of sequences of word types. A sequence of N words, $w_1 \ldots w_N$, is called an N-GRAM and the probability of that N-gram is $P(w_1 \ldots w_N)$; often, we only have enough confidence in our statistics to look at small values of N, so we use the statistics of unigrams (1-grams, like *in*), bigrams (2-grams, like *in the*), or trigrams (3-grams, like *in the blue*). We can also transform N-gram probabilities to conditional probabilities.

If a user of an assistive communication device (or, indeed, any modern smartphone) has just typed *see you*, then we can use the trigram model to consider all possible words that come next. More formally, if $P(\textit{see you}\ \mathbf{w}^*) \geq P(\textit{see you}\ \mathbf{w_x})$ for any word $\mathbf{w_x}$, then $P(\mathbf{w}^* \mid \textit{see you}) \geq P(\mathbf{w_x} \mid \textit{see you})$ and $\mathbf{w}^*$ is therefore the best next possible word, and we can suggest it to the user.

The probabilities of N-grams can be determined very easily. All that is required is a large sample of text data similar in nature to the kind of text we expect people to type. For example, if we expect people to type "financial news", we might learn our probabilities from the *Wall Street Journal*. If we expect people to type "medical diagnoses", we might learn our probabilities from medical textbooks. In either case, the data we gather is called a *corpus*,[2] sometimes denoted as C. Our probabilities can then simply be obtained by counting the occurrences of a particular N-gram in that corpus, divided by the total number of N-grams in that corpus. For example, if $N = 1$ and we want to know the prior probability of the word *octopus*, we would just count the number

[2]**corpora** *n.pl.* the plural of *corpus*.

of times *octopus* occurs in our corpus and divide it by the total number of 1-grams (i.e., words); if *octopus* occurs once in a corpus with 100 word tokens in total, we estimate $P(octopus) \approx 1/100 = 0.01 = 1\%$.

In general, this extremely simple approach works incredibly well and across contexts much more varied than word prediction. Knowing the probabilities of N-grams allows us to estimate the probabilities of much longer sequences. For example, if we have a bigram model, we can estimate the probability of longer sequences like *the cat in the hat* by multiplying together all the component bigram probabilities. That is:

$$P(the\ cat\ in\ the\ hat) \approx P(the\ cat) \cdot P(cat\ in) \cdot P(in\ the) \cdot P(the\ hat).$$

What this allows us to do is evaluate whether one sentence is more likely than another. In speech recognition, this means that if the system is having difficulty deciding between two competing hypotheses for what was said, we can rely on which hypothesis is simply more "likely" in a language. While this approach does have broad applicability, there are complications. Not least among these is the problem that everyday people utter or encounter sentences or even component phrases and N-grams that have never been uttered before. For example, you've probably never read or heard the sequence *Just Google the Instagram app*, despite its innocuousness. If you have never encountered the trigram *Google the Instagram* in your life (which constitutes a *very* large corpus of language), then our approach would estimate that $P(Google\ the\ Instagram) = 0$ and therefore that our nice little sample phrase is *impossible*. Fortunately, there are many algorithmic solutions. A gentle introduction to these can be found in the textbook by Jurafsky and Martin [2009].

The current word w_i can also be anticipated given an n-gram context augmented by part-of-speech tags t_j. For example, Fazly and Hirst [2003] describe an algorithm that ranks possible completions based on the estimate

$$\begin{aligned} P(w_i|w_{i-1}, t_{i-1}, t_{i-2}) &\approx \sum_{t_i \in T(w_i)} P(w_i|w_{i-1}, t_i) P(t_i|t_{i-1}, t_{i-2}) \\ &\approx \sum_{t_i \in T(w_i)} \frac{P(w_i|w_{i-1}) P(t_i|w_i)}{P(t_i)} P(t_i|t_{i-1}, t_{i-2}) \\ &= P(w_i|w_{i-1}) \sum_{t_i \in T(w_i)} \frac{P(t_i|t_{i-1}, t_{i-2}) P(t_i|w_i)}{P(t_i)} \end{aligned} \tag{3.1}$$

where $T(w_i)$ is the set of all possible PoS tags associated with word w_i. Combining PoS with lexical context in this way reduces the percentage of keystrokes needed to produce text by $\sim 6\%$ over purely *a priori* statistical methods [Fazly and Hirst, 2003]. Other extensions to text prediction to further refine the list of hypothesized completions include the use of grammatical syntax and semantics [Erdogan et al., 2005, Li and Hirst, 2005], as well as trained neural networks [Garay-Vitoria and Abascal, 2006].

Empirically observed improvements in the rate of typed communication with prediction might not overcome improvements gained through the use of speech (see above), but applying the same approach to predicting spoken communication may reduce the amount of effort required for both the dysarthric speaker and their audience. If speech input is coupled with a visual display for output, for example, that display could be updated "on-the-fly" with the results of predicted queries before those queries are completed.