Prosody Beyond Fundamental Frequency

Greg Kochanski

2006/01/29 23:02:20 UTC

This document is available at `http://kochanski.org/gpk/papers/2005/2005BeyondF0.pdf` . It will be published in *Methods in Empirical Prosody Research*, edited by S. Sudhoff, D. Lenertov'a, R. Meyer, S. Pappert, P. Augurzky I. Mleinek, N. Richter, and J. Schließer. Published in Berlin, New York by De Gruyter in the Language, Context and Cognition Series, to be published in June 2006. ISBN 3-11-018856-2.

1 Introduction

Most of this book is concerned with tactical details of experiments: relatively detailed prescriptions for techniques after the goals of the experiment are already decided. In contrast, this chapter is intended to help with the broader questions of what to measure and what experimental strategies to use or avoid.

The chapter presents a plausibility argument, based on information theory, that there is substantial amounts of interesting prosody in acoustic properties other than f_0. If true, there are obvious consequences for the design of prosody experiments.

Along with that, I argue that measures of information are applicable to prosody, and can allow new experimental tests of and constrains on phonology. These information measures are (almost) new to linguistics, even though they are commonly used elsewhere. New measures imply the possibility of new experiments. Along the way, the chapter will also touch upon some appropriate and inappropriate experimental methods.

Looking at prosody as a means of communicating information naturally raises questions that can connect theoretical linguistics, psychophysics and

articulatory modelling. This is done by following the flow of information from an abstract linguistic entries in the mind of the speaker, via articulatory motions, through its representation as part of the sound wave, through the listener's perceptual process, and back to an abstract linguistic representation. Information theory puts a global constraint on this flow, so that if we can bound the amount of information anywhere in the flow, we can make deductions about what is going on elsewhere in the flow. This chapter will focus predominantly on English, although the techniques have broader application.

2 What is Prosody?

Prosody is a general term that describes the way one says a particular sentence. I will assume, as a working definition, that prosody describes any acoustic properties that cannot be predicted by looking at the immediate lexical neighbourhood.[1]

Thus, in English, fundamental frequency (f_0) carries prosodic information but little lexical content (it only helps to separate a few pairs of words differing in stress location[2] and may help distinguish between certain consonant pairs like /b/ and /p/).

However, there is more to prosody than fundamental frequency, as one can see by considering common descriptions of speech such as "loudly", "softly", "sharply", "harshly", "clearly", "slowly", "rasped", "mumbled", and "sobbed". Not all of these descriptions can be expressed in terms of fundamental frequency; thus, beyond the widely recognised f_0 and duration, other acoustic properties contribute. Cutler et al. [1997] provides a review of the subject, touching on considerable psycholinguistic evidence.

- Loudness$_a$[3] and intensity can be important components of prosody; (Fry, 1955 Hadding-Koch, 1961; Maekawa, 1998; Kehoe et al., 1995;

[1] In tone languages like Mandarin, the tones are lexically specified and would count as part of the lexical neighbourhood. Thus, in a tone language, f_0 contains both lexical and prosodic information. Note that some people take lexical tones to be part of the language's prosody. Effectively, they are defining prosody in terms of the acoustic carrier of the information (f_0 in this case), rather than the type of the information carried (i.e. non-lexical).

[2] E.g. per**mit** vs. **per**mit.

[3] I use the term "loudness$_a$" to mean the result of an algorithm applied to the acoustical signal that approximates perceptual loudness. Loudness$_a$is a computed from the local

Silipo and Greenberg, 2000; Kochanski et al., 2005), although there is some disagreement (e.g. Fry, 1958; Turk and Sawusch, 1996).[4][5]

- The slope of the speech spectrum seems to be correlated with the prominence of syllables [Sluijter and van Heuven, 1996, Sluijter et al., 1997]. This slope is associated with the timbre of the speech.

- The degree of voicing is correlated with prominence [Ding and Campbell, 1996], and it has been noted that voiceless consonants in stressed syllables can become voiced [Shih et al., 1999, Fant et al., 2000], thereby introducing a correlation between prominence and acoustic measures of voicing.

Consequently, it seems surprising that in a sample[6] of the scientific literature, articles relating to a single acoustic measurement – f_0 – outnumber articles investigating either loudness, spectral slope, or timbre by nearly 5 to 1.

3 Communication Channels and Channel Capacities

Speech exists to communicate, at least in a broad sense, so it is interesting to compare different prosodic properties to see how much they could contribute

spectrum, and is roughly proportional to the cube root of the acoustic power; its definition can be found in Stevens [1971], as modified in Kochanski et al. [2005]. The distinction between loudness$_a$ and loudness will only be maintained where it might make a noticeable difference to the arguments. Work on intensity (the acoustic power) will also be treated as loudness$_a$, unless the distinction is crucial.

[4] I use the term "prominence" to indicate the acoustic properties of accented syllables in English that separate them from their unaccented neighbours; see Kochanski et al. [2005]. Specifically, syllables are defined to be prominent in that paper if and only if they carry an IViE accent[Grabe et al., 2001].

[5] Some references dealing with word stress are included in the list because word stress and accent tend to be expressed together. Word stress is an abstract lexical property, but evidence from which stress could be deduced is primarily observable on accented words. Likewise, the special acoustic properties associated with accents are most dramatic on syllables that are both lexically stressed *and* prosodically accented.

[6] Searches were conducted through abstracts and titles of *Phonetica*, *Journal of Phonetics*, and *Language and Speech*. Fundamental frequency-related articles (194) were counted as the sum of hits on "fundamental frequency" and "pitch"; loudness-related articles (24) summed hits on "loudness" and "intensity"; counts for timbre- (4) and spectral slope-related articles (8) were taken as twice the number of hits for the corresponding searches.

to communication. We will look at one acoustic property at a time, such as f_0 or loudness$_a$; each of these properties corresponds approximately to the concept of a communication channel.

What is a communication channel? It is any signal that can carry information. The person at one end has a strategy for encoding meaning into the signal and the person at the other has a strategy to decode the signal into something he/she considers meaningful. Often, the signal is corrupted to some degree in between. The idea comes from telecommunications research [Shannon, 1948, Shannon and Weaver, 1949]. A good review of modern information theory is Gray and Neuhoff [2000], but many of the results are in standard texts on computational linguistics such as Manning and Schütze [1999].

For an example, imagine reading a credit card number over a telephone; suppose the person on the other end did not quite hear the number and wants to confirm one digit. They might say "Was that 5555 5555 4**3**99 1017?" and select a particular digit prosodically. The speaker is then communicating some information that lets the listener pick out a particular digit; we might translate the prosodic information as "second digit in third group."

The mathematics of communication channels is very general, broadly applicable, and it is regularly applied to all modern communication systems (e.g. Kim et al., 2004).[7] Specifically, it is applicable to the normal conception of language as combinations of discrete symbols such as words, phonological features, or accents in intonational phonology.[8] If a speaker starts with such discrete symbols, information theory places limits on the listener's ability to figure out what symbols the speaker was attempting to communicate. Given a well-understood channel, these limits are exact, and experimental confirmation is embodied in every telephone and modem.

For the purpose of this chapter, the communication channel begins somewhere in the midst of the brain of the speaker, where the abstract digit string is converted to a sequence of motor commands. It includes the articulatory strategies that shape the speech, includes the telephone and the cochlea of

[7] Therefore, it is applicable to human language, as human language is one of the major payloads of telecommunication systems.

[8] Much of the math of information theory is also applicable if the information to be communicated is not entirely composed of discrete symbols. So, if information like prominence or the intensity of some emotion is best represented by a real number instead of a binary feature, the general approach remains valid although some of the estimates may change.

the listener, and terminates somewhere in the mind of the listener where the sounds are finally interpreted (in our example) as a question about one specific digit of a number.

There are four critical insights from information theory that make such a complicated and messy system tractable:

- Any communication channel has a maximum rate at which it can carry information.
- One can often describe a communication channel as a sequence of simpler channels.
- Once information is lost, it will not re-appear: a channel can be thought of as a leaky pipe. Thus, if one can prove that a certain bit of information is in the channel at any point, then it must have been transmitted into the channel and presumably specified by the speaker's phonology.
- The amount of information that a channel can carry is limited by its slowest component. So, if we can understand even one of the components, we can use that knowledge to set an upper limit on how much information the overall channel could possibly carry, even if the channel is too complex to understand completely.

One can focus on particular aspects of speech (like f_0 and loudness_a), and treat each of those as a communication channel in its own right so long as they are more-or-less independent of other aspects of speech. This approach was introduced by Miller [1956] in his classic paper; he showed how human perception could usefully be described as a collection of noisy communication channels, where the noise in the channels corresponds to the limitations in our perceptions.[9]

One reason that communication channels are a useful concept is because the information that they carry can be quantified. The relevant equation is Shannon's channel capacity [Shannon, 1948],

$$C = W \log_2(1 + S/N), \tag{1}$$

[9] To quantify this relationship between perceptual limitations and an effective noise level in a communication channel, see references on Signal Detection Theory, such as Wickens [2001], Macmillan and Creelman [1991] and references therein.

where C is the maximum possible rate at which one can communicate through a channel.[10] For this equation, the channel is assumed to be linear, to have a bandwidth W, and to have an additive noise[11] whose variance is N. The signal to be transmitted has a variance of S. The resulting channel capacity tells how many bits per syllable (or per second) could be communicated, if the language was optimized to pack information into that channel.

Equation 1 is easy to justify qualitatively. The bandwidth W corresponds to how often you can transmit a symbol down the channel. The more often you can transmit symbols, the more information you can convey. The ratio S/N measures the complexity of each symbol; S can be interpreted as the size (e.g. f_0 excursion) of the symbol, and N as the minimum size excursion that can be usefully interpreted.[12] If S/N is large, the signal stands out above the noise and you could hope to communicate a lot of detail in each symbol. Conversely, if S/N is near one (or smaller), the listener will be hard pressed to even find the signal, and any complicated detail will be lost. Putting the two parts together, it's reasonable that the total information transferred is just the product of how many symbols you can transmit times how much detail you can transmit on each symbol.

For instance, if $W = 5$/second, Equation 1 can be interpreted in terms of sending 5 symbols per second.[13] It might be sensible then to say that each syllable carries a symbol. Then, if $S/N = 16$, (a plausible value for communication via f_0), the equation asserts that it is possible to transmit four independent binary decisions (bits) within each symbol. Thus, there would be 16 distinguishable f_0 contours possible on each syllable, and to make

[10] Strictly, this is the maximum rate at which data can be sent through the channel without any errors, with the best possible encoding.

[11] Assumed to be Gaussian with a flat power spectrum.

[12] One expects that N is at least as big as the perceptual just-noticeable difference, as one presumably cannot interpret a change that one cannot detect. Also, N must be bigger than any "interference": not only must an excursion be noticeable, but it must be big enough to interpret. For example, N will likely be larger while jogging, since the physical bumping will interfere with the jogger's f_0 control. One might notice an loudness$_a$ or f_0 bump, yet not know if it was intentional or merely caused by a step off a curb. In such a case, it would not be interpretable, even if it were easily detectable.

[13] That's not the only possible interpretation, as one can think of larger symbols that each contain more information but are sent less often, or smaller, more compact symbols that carry less information but are sent more often. Equation 1 does not specify the symbol; it it specifies the maximum amount of information you can carry, no matter what set of symbols you choose, and there may be more than one equally good way to approach the limit.

use of all four bits, a combination of linguistic features must be associated with each distinguishable contour.

Equation 1 can be an over-optimistic limit to the actual amount of information transferred if the language does not use the channel optimally or if the information is intentionally transmitted redundantly.[14]

4 How Efficiently is the Channel Used?

How close are languages to optimal encodings? Are there any ways in which they are imperfect or inefficient? I'll next look at three basic strategies for efficient encoding and see how they might apply to languages. At issue is whether languages transmit one bit for each two-way phonological distinction or if some distinctions can be transmitted more efficiently.

4.1 Block Encodings for Efficiency?

Readers versed in information theory may recall that, a full bit may not be needed to encode a two-way distinction if one of the alternatives is rare. A good example might be the question/statement (Q/S) distinction; in most text and conversation, questions are relatively rare, so one might expect to be able to encode the Q/S distinction with less than one bit, on average.

Indeed, in a large corpus of N sentences, where a fraction Q of sentences are questions, one might encode the distinction in as few as

$$I = -N \cdot (Q \log_2 Q + (1 - Q) \log_2 (1 - Q)) \text{ bits.} \tag{2}$$

If we take $Q = 0.1$ as an example, the Q/S distinctions for the corpus could be encoded in $I/N = 0.47$ bits per sentence, on average. However, that does not apply to a single sentence. In order to use less than 1 bit per sentence, one would need to encode a number of things as a block: either Q/S distinctions from several sentences or multiple linguistic features (see §4.2).

[14] Why would a language transmit information redundantly? One reason might be to make the language robust against noise: languages without redundancy would be unintelligible in the presence of loud noises such as hammering. Another reason for redundancy is to allow for mismatches between the communication strategy of speakers and listeners of the language. If a language transmits some of its information twice, then it might be understandable to listeners who comprehend either representation.

Specifically, Equations 1 and 2 are derived on the assumption that the information could be encoded in large blocks. Absent that assumption, the amount of information actually transferred would be less [Shannon, 1959].

Encoding in a block is more efficient because it can be more compact to transmit the equivalent of "five statements then question" than "statement, statement, statement, statement, statement, question." However, in the real world of dialogues, neither the speaker nor the listener has the luxury of encoding data from a block of sentences. The speaker and listener are conducting a real-time conversation, and the listener must be able to decide immediately whether she heard a question or statement.[15] Thus, block encodings are not generally relevant to languages, and there is good reason to believe that when a feature needs to be transmitted, it will use up at least one bit's worth of the channel capacity.

4.2 Efficient Encoding by sharing slots?

Another possibility to encode a linguistic feature with an average of less than one bit would be to to encode together several features that co-occur in a sentence. However, this approach can work only if a suitable set of linguistic features can be identified.

Taking up the question/statement distinction again, if one wanted to show that that Q/S was encoded with less than one bit, one would need to find a linguistic feature (call it X) that shares the same acoustic representation (i.e. slot) as the Q/S distinction. Additionally, a speaker must never need to make both X and Q/S distinctions in the same utterance: one or the other must be irrelevant.

So, if X is expressed as a final rise so it can share the slot with Q/S, and it can only be transmitted in situations where the Q/S distinction is obvious from other cues. Continuation rises (Caspers, 1998; Chen, 2003 and references therein) and contradiction contours [Liberman and Sag, 1974, Pierrehumbert and Hirschberg, 1990] are plausible candidates for X. To make this work, of course, the speaker and listener must share a common strategy so that they both know when a final rise as marks a question, and when it indicates a desire to hold the floor. For instance, in a *wh*-question, the lexical

[15] This kind of conflict between encoding efficiency and block size is well known in telecommunications in the design of speech coders (where larg block sizes add unacceptable delay) and also in economics [Sims, 2003].

items make the Q/S distinction early on, so the listener does not need to use prosody to make the distinction; the slot could be used for something else.

Take prominence as another example. The simplest possible encoding is to allocate one bit per word, so that each word can be marked as prominent or not. Is there a way to transmit fewer bits, while still making sure that the listener knows? Perhaps, if some words are known in advance not to be prominent.[16]

In that case, the bits only *need* to be transmitted where prominence is possible. For instance, if a language forbids two prominences in sequence (P,P), then nothing need be transmitted for the syllable following a P, because there is no choice: the syllable is forced to be non-prominent.[17] For a language like English, where about a third[18] of the syllables are prominent, this constraint means that (on average) about 74% of the words have a prominence that is not predictable in advance, so 0.74 bits per word are required to specify prominence.[19] Likewise, if some syllables are known to be prominent (or not) based on earlier lexical or syntactic cues fewer bits would be required.

In other words, if the listener can deduce prominence from other, earlier information, then the speaker does not *need* to transmit prominence information on a certain word. However, that does not automatically give a more efficient encoding. If a word becomes an empty slot, it is available for other uses, but the encoding only becomes more efficient if the empty slot is actually put to another use. Perhaps the prominence information is transmitted anyway, just to act as confirmation of the earlier cues. In that case, we have not reduced the bit rate, although we would have a more robust language. If the earlier cues say that the syllable is non-prominent, and it is acoustically marked as prominent, then the listener can deduce that something is wrong and ask for clarification. The disadvantage of slot recycling is that the listener can make fewer cross-checks, and so would be less able to realize that

[16] Of course, it works just as well if some words are known in advance to be prominent: the essential feature is that the prominence or lack thereof be known before it is time to transmit the prosodic marker.

[17] Likewise, if it is not allowed for two adjacent words to be both prominent, words following a prominent word will be known in advance to be not prominent, so the prosody need not transmit any information.

[18] The data sets in Kochanski et al. [2005] mark 36% of the syllables as prominent.

[19] This assumes a mixture of 1- and 2-syllable words with a mean length of 1.38 syllables per word, 36% of the syllables are prominent, at most one prominent syllable per word, and that if a word has a prominent syllable, then neighboring words will not.

something was mis-heard or mis-spoken.

If the language specifies what to do with the empty slot, it would need to have a rule like this: "a loud, long syllable is prominent, except after a prominent word, where it means Y, instead." Feature Y would have to be something suitable: something that does not need to be expressed very often and something that can be delayed to the syllable after the next prominence.

The experimental signature of such slot recycling would be clear: instrumentally both words in the pair would have all the acoustic characteristics of prominence and human judges would call the first word in the pair prominent, but they would instead label the second word in each pair with property Y. No obvious candidate feature comes to mind in this particular case, but there are many combinations of languages, linguistic features and acoustic encodings that could be investigated. Experiments are needed to understand how linguistic features are encoded into speech, and to what extent the encodings depend on prior context.

While it is likely that slot recycling exists in languages, there are enough constraints in finding linguistic features that can share a slot so that it might not be possible for every feature.

4.3 Summary of Encoding Efficiency

Language does not make full use of the communication channel. Encoding of blocks of utterances (§4.1) is inconsistent with an interactive dialogue, and the lack of block encodings prevents the communication channel from being used with complete efficiency. Encoding several features together by recycling slots (§4.2) probably occurs, but it may not be applicable to all features, for lack of compatible features that can share a slot. Finding groups of linguistic features that can be encoded into the same slot without conflicts is a challenging task for general and experimental linguists.

When making numerical estimates of information (below), I will assume that each linguistic feature uses one bit of channel capacity. This is clearly an approximation, assuming an approximate balance between the number of bit slots that are recycled on one hand, and encoding inefficiencies combined with language's desired redundancy on the other.

5 How much channel capacity does a language use?

Each bit in the channel capacity corresponds to one binary decision, that is, a binary linguistic feature. Consider prominence as a concrete example. Assuming that prominence is binary, the language needs to transmit 1 bit[20] for each word that could potentially be prominent – the bit specifies whether a syllable is prominent (1) or non-prominent (0).

It has been suggested that f_0 (as one component of prosody) conveys many different varieties of information. I'll estimate how much information would be needed for each, neglecting slot recycling.

- It makes the Question/Statement distinction (e.g. Morlec et al., 2001)[21] (1 bit per sentence).
- It may mark certain words as prominent,[22] as claimed in Bolinger [1958], 't Hart et al. [1990], Ladd [1996]. (1 bit per word).
- It may mark beginnings and ends of sentences, phrases and words (e.g. Kochanski et al., 2003). (about 1 bit per word, phrase, and sentence, respectively). This may help distinguish different syntactic structures [Albritton et al., 1996].
- It may mark focus in sentences (about 1 bit per sentence), as in "Was that 5555 5555 4**3**99 1017?", above [Kochanski and Shih, 2003a]. Another example might be "I did **not** eat the dog." vs. "I did not eat the **dog**."

[20] If there were more than two levels of prominence, then the required bit rate would be higher, perhaps near 2 bits per word.

[21] Note that Grabe [2002] shows that at intonation labels do not reliably predict whether or not a question is being asked, so it is not clear to what extent f_0 actually helps the listener make the Q/S decision.

[22] Experimental support for this is mixed. Gussenhoven et al. [1997], Rietveld and Gussenhoven [1985], Terken [1991] have shown that it is possible to mark prominence by f_0 motions, but it appears that speakers under reasonably natural conditions may not normally do so: Silipo and Greenberg [1999, 2000] and Kochanski et al. [2005] saw only weak correlations of f_0 with prominence in their corpus-based analyses (also see §7.3.2). However, prominence is transmitted by some acoustic property, even if not by f_0, so this does not affect the total amount of information in prosody, just its distribution among various channels.

- It may control the flow of a dialog, keeping the floor, requesting the floor or passing it to others (about 1 bit per sentence to discriminate between wanting and not wanting the floor).

- It may express a variety of emotions. No one really knows how much information is necessary to describe emotions (see Cowie and Cornelius, 2003; Schröeder, 2001; Morlec et al., 2001 for discussions), but plausible estimates are either several levels of three coordinates (perhaps 6 bits per sentence), or a choice of six basic emotions and an intensity coordinate (a similar number of bits).

- It may mark new *vs.* old (given) information (1 bit per word).

- It may help distinguish between certain minimal pairs of words[23] (Nearly zero bits for English).

- It carries implicit information, such as the age (young vs. old is 1 bit), sex (1 bit), and dialect (several bits, perhaps including the level of education, and social class of the speaker) as noted by v. Laziczius [1935] and Firth [1937, 1950]. Because at least some of this information can be extracted from prosody (e.g. Rouas et al., 2003), it must be encoded into the available channels, whether intentionally or not. The bit rate for this information is rather uncertain, but one might estimate it as 6 bits, spread out over several phrases, for a data rate of perhaps 0.25 bits per syllable.

To put all these estimates on the same basis, they will be converted to an average number of bits per syllable, assuming 1.38 syllables per word, 6.6 syllables per intonational phrase, and 12 syllables per sentence.[24]

To transmit all this information requires a channel capacity somewhere between the largest value and the sum of all the values[25]. If all these features were independent, that is if all combinations were possible, the channel

[23] A handful of pairs in English, but many pairs in tone languages and a correspondingly larger number of bits.

[24] Syllables per word and per intonational phrase are derived from the Grabe et al. [2001] "read Cinderella" data for the EL secondary dataset, processed as per Kochanski et al. [2005]. We arbitrarily assume there to be 12 syllables (about two intonational phrases) in a typical English sentence.

[25] If one has an object that requires A bits to transmit, and another that requires B bits to transmit, then transmitting the two together requires C bits, where $\max(A, B) < C \leq A + B$ bits. Two nearly identical objects will approach the lower limit, whereas the upper

capacity would have to be greater than the sum of all the individual requirements, 3.4 bits per syllable. If the various linguistic features were correlated so that certain combinations were impossible, then the amount of information to be transmitted (and thus the required channel capacity) would be reduced.[26] However, quite strong correlations would be required to drop the information rate down substantially. Even if only 0.1% of the possible combinations of these features formed allowable (≈grammatical) sentences, the entropy would still be 2.6 bits per syllable. No such set of rules has been laid out. Thus, representing all of these linguistic and paralinguistic properties would likely require a channel capacity of 3.4 bits per syllable. Future research[27] may cast light on the interactions between these varieties of information in f_0 and may reduce the required channel capacity, but it seems unlikely that it would be reduced below 2 bits per syllable.

6 How much channel capacity is available?

There is no shortage of ways to transmit prosodic information and no reason to believe that it is all in one channel. We will consider fundamental frequency, loudness$_a$, and duration as communication channels, and see if they are individually or collectively sufficient to transmit all the information that the language requires. Other properties like spectral slope could be considered also, but their perceptual importance are less well understood at present.

6.1 Fundamental Frequency

We can estimate the information capacity of the available channels, using Equation 1 or appropriate variants. For f_0, we can estimate $W \approx 5$ Hz,[28]

limit is reached when the two objects are independent of each other so that knowledge of one does not help predict the other. The reader should consult Gray and Neuhoff [2000] for a derivation.

[26] For instance, if the question/statement distinction and dialogue control were correlated so that there were only three of the four possibilities were allowed by the language (statement & release, ask question & release, statement & hold the floor), then the information rate would drop by 0.03 bits per syllable on average.

[27] This is an opportunity for experiments. Such experiments could be valuable in understanding speech even without reference to an information-theoretic framework.

[28] The symbol "≈" means "approximately."

based on Xu and Sun [2002] and Sundberg [1979]. At a typical speech rate of 180 ms [Grabe et al., 2001, Kochanski et al., 2005] per syllable, this amounts to $W \approx 1$ cycle per syllable. S can be computed from a corpus: $S^{1/2}$ is just the fractional[29] standard deviation of f_0. Subjects reading "Cinderella" in the Oxigen corpus have a wide range of S, with the 25^{th} percentile at $S^{1/2} = 0.19$ and the 75^{th} percentile at $S^{1/2} = 0.3$.

However, N has not yet been precisely measured. While the just-noticeable frequency difference (JND or difference limen) between pure tones can be as small as 2 Hz under favourable conditions (corresponding to $N^{1/2} = 0.01$), the just-noticeable f_0 difference for speech is likely to be larger than pure tone JND because the changes in formant structure and voicing are likely to be distractors[30] [Chuang and Wang, 1978]. Chuang and Wang report $N^{1/2} = 0.015$. The JND for pitch *motions* seems to be rather larger; 't Hart [1981] and t'Hart et al. [1990] quote values of 1.5 to 3 semitones, or $N^{1/2} \approx 0.10$ (although the measurements are not strictly comparable). Such larger values are consistent with [Black and Hunt, 1996], who considered a f_0 prediction error of 9.9 Hz RMS to be unimportant ($N^{1/2} \approx 0.06$). Although the various estimates cover a substantial range, they appear inside a log function, so they make less difference to the channel capacity than one might expect.

A final factor (not shown in Eq. 1) is that speech is not always voiced, and thus no f_0 information is available in the unvoiced regions. For example, 74±7% of the speech in the the IvIE Cinderella corpus is voiced.[31] This will reduce the channel capacity by a similar factor.[32] After the reduction, the estimated amount of information that can be carried by f_0 range from

[29] S could be expressed in Hertz or semitones also; the results will not change noticeably, so long as N is expressed in the same units. Here, I use the standard deviation of (f_0 divided by that speaker's mean) as a way of normalizing away the difference between male and female speakers.

[30] There is an extensive literature on how one stimulus affects the perception of another stimulus. A useful review is Huettel and Lockhead [1999].

[31] This measurement is over marked intonational phrases, and does not include pauses between phrases. The error bar is the inter-subject variability.

[32] The actual reduction in channel capacity depends on correlations between voicing and the intended intonation. It also depends on the mean duration of unvoiced intervals, relative to the time required to change f_0. If the voicing is uncorrelated with intended intonation and if the unvoiced regions were long enough so that the f_0 just before the unvoiced interval would not be a substantial help in predicting the f_0 just after of the unvoiced interval, then if the speech is voiced v% of the time, then the channel capacity is reduced to v% of the value given in Equation 1.

$C = 1.8$ bits per syllable to 5 bits per syllable. This estimate is in general agreement with experimental results of Pollack [1952, 1953] (reviewed in Miller, 1956) who obtained a channel capacity of 2.5 bits per burst of pure tone. Quantitative modelling of Mandarin intonation also yields a similar result, where 2 bits per syllable is sufficient to accurately reproduce f_0 contours [Kochanski and Shih, 2001, §4.2].

This is an upper limit to the amount of information that can be transferred via variations in f_0; this channel capacity can be only be reached by fairly complex strategies for encoding and decoding the desired signal. Such strategies invariably involve knowledge of or adaption to the characteristics of the the communication channel. Here, the listener would need to know in detail how the speaker encodes linguistic intentions into f_0. In other words, the listener must be familiar with the speaker's dialect (and perhaps with the speaker herself) to have a chance of understanding all the the information in the channel.

Thus, there is a potential problem. There may well be too much prosodic information to encode into just f_0 so linguists may be forced to consider prosody beyond f_0. If all the information in Section 5 is indeed encoded into f_0, then the channel capacity of f_0 must be almost completely used, leading to a number of interesting implications, all of which can be experimentally tested:

- The speaker and listener must share a rather efficient and precise encoding strategy. Given that there are substantial differences between the intonation of different dialects of English [Grabe et al., 2005, Grabe, 2004], it seems likely that communication across dialects would be rather less efficient than communication within a dialect because of the lack of a shared encoding.

- If there are substantial differences in intonation between speakers, the listener must adapt to the speaker. Communication would be unreliable until the adaption completed.

- The listener must make linguistic sense from rather small f_0 changes, probably smaller than segmentally-induced f_0 shifts, possibly nearly as small as the pure-tone just-noticeable difference. This is because the only way to make the channel capacity in Equation 1 large enough to carry all the information is to make N small since S and W are fairly well determined. (Recall from §3 that N is a measurement of the

smallest signal that carries much infomation, so small N implies that small signals must be meaningful.)

If other acoustic properties contribute noticeably to prosody, then each is its own channel of information, and the overall prosodic channel capacity could be as large as the sum of the channel capacity of each acoustic property.

6.2 Loudness and or Intensity

Loudness can certainly carry information; Garner [1953] studied human discrimination of loudness levels, and found that 2.3 bits of information could be encoded in the loudness of a tone. In other words, people could reliably distinguish about 5 different levels of loudness.[33] The experiment is not directly relevant to linguistics, as the conditions were substantially different from what one might find in speech. On one hand, the ten levels were spread out over a very wide intensity range, thus making them easier to distinguish, but on the other hand, the tones were presented in isolation, requiring an absolute judgement rather than a relative comparison of the loudness of sequential syllables.

A more relevant measurement has been done by Riesz [1928] and Jesteadt et al. [1977], who looked at the discrimination of two near-by intensity levels. They showed that relatively small intensity differences ($\Delta I/I \approx 20\%$ though dependent on loudness and frequency) can be detected, and thus the ear is easily capable of making a binary louder/quieter judgement[34] on adjacent syllables.

Indeed, Kochanski et al. [2005] showed that that loudness information carries at least a modest amount of information about the prominence structure of utterances, and that it appears to carry more information about prominence than f_0 does.

[33] The equivalence of 2.3 bits and 5 levels comes from the definition of the entropy of a probability distribution. In general, if you have N equally likely levels, you need $\log_2(N)$ bits of information to select one of the levels. Here, $\log_2(5) \approx 2.3$.

[34] People can control the loudness of their speech with precision. On the reasonable assumption that the loudness is controlled by feedback from what they hear, loudness perception is probably comparably precise. Kochanski and Shih [2003b, §V.B] showed that, when measured properly, the variance in the power for identical sections of different sentences can be as small as 9%, corresponding to just a 4% standard deviation in loudness from sentence to sentence.

We can evaluate Shannon's channel capacity for loudness much like we did for f_0. In this case, we compute the normalized loudness by the techniques described in Kochanski et al. [2005], and obtain the variance of the normalized loudness[35] to be $S^{1/2} = 0.54$. The "noise" can be obtained from the above-quoted results for $\Delta I/I$, allowing for the known [Stevens, 1971] approximate cube-root scaling of loudness$_a$ with intensity to yield $N^{1/2} = 0.06$ or thereabouts. Like f_0, loudness$_a$ goes through a complete cycle per syllable, so $W \approx 1$/syllable.

However, some of this channel is unavailable for prosodic information. Vowels tend to be louder than consonants (especially stops), so some of the variation in the loudness$_a$ is driven by the phone sequence specified by the words of the utterance. We will make the fairly conservative assumption that the loudness$_a$ of each syllable nucleus can be independently controlled, but that the intervening shape of the loudness$_a$ profile is specified by the sequence of phones. In that case, it can be shown that the information capacity of the channel is reduced by a factor of two.[36]

Evaluating Equation 1 with the 50% reduction, we obtain a channel capacity of $C = 3.5$ bits per syllable, plausibly close to the estimate in Garner [1953]. The raw information capacity of the loudness channel is thus comparable to the capacity of the f_0 channel.

6.3 Duration

Similar arguments could be made for the duration of syllables. Quite small differences in phone and syllable duration are detectable [Huggins, 1968, 1972, Carlson and Granström, 1975, Quené, 2004], and thus several bits per syllable could be encoded. It is well known that many languages code vowel quantity as duration, thus duration is clearly capable of transmitting at least one bit per syllable. In languages without vowel quantity, or for vowels without a quantity distinction, that bit is also available for communicating

[35] Interestingly, the variance of the normalized loudness does not change much from person to person, compared to the variance in f_0. For instance, the subject with the least loudness variance (in Kochanski et al., 2005) has a variance that is 43% of the most variable subject. For f_0, the equivalent ratio is 12%, implying that a few of the subjects barely make any f_0 motions at all.

[36] Loosely speaking, the loudness on the nucleus carries information, but the loudness in between syllables does not. Consequently, only half of the measurements that one might take in a syllable carry useful information.

prosody.

A complication in computing the total information capacity is that the perceived loudness of sounds shorter than 200 ms (about the normal syllable duration) are related to both the intensity and the duration of the sound [Pedersen and Salomon, 1977]. In other words, one cannot use those two channels entirely independently, and if the speaker attempted to transmit different information on the duration and loudness channels, the listener would perceive a confused mixture of the two. However, the degree of correlation (confusion) is not entirely clear, as it seems that durational and loudness changes are processed separately in the brain [Giard et al., 1995].

6.4 Information and Limits

Thinking about language and prosody in terms of a comparison between the amount of information that must be carried and the capacity of various different acoustic channels looks likely to be fruitful. Some aspects of language can be considered to be the transmission and reception of well-defined bits of information. For instance, to the extent that intonational phonology is a good representation of human language, the speaker is attempting to transmit a sequence of discrete accents to the listener such as **H*L**, **L*H**, **H%**. The speaker encodes these abstract concepts in terms of time-varying values of f_0, loudness and other properties on the speech waveform. The listener then takes the waveform, extracts the pitch, and attempts to deduce what accents were transmitted.

Since linguistic content is generally considered to be composed of well-defined discrete features, the information required to transmit those features can be calculated, so long as those features are independent in the sense that all combinations are possible or if the correlations can be estimated.

On the other side of the equation, with results from psychophysical and other experiments, it is possible to compute the maximum rate at which some acoustic property like f_0 can transmit information.

It seems plausible that

1. It is not possible to transmit all the prosodic information that has been discussed in the literature by way of f_0 alone,

2. Other acoustic properties may be able to communicate as much information as does f_0.

Therefore, the field needs to consider other acoustic properties, beyond f_0, as carriers of prosody.

7 Implications for Experiments

There are three broad classes of experiments that are motivated by looking at the information content of language.

- Experiments relating to a channel capacity.
- Experiments relating to how a linguistic feature is encoded.
- Work toward establishing how much information and which information is transmitted.

7.1 Channel Capacity Experiments – Perception

These experiments attempt to put a limit on how much information can be transferred. They may be basic psychophysical experiments to help understand the limitations of human speech perception. While much psychophysics has been already done, the bulk of the work has been on tone bursts and pure tones, and much less work has been done on speech-like signals.

For instance, we know that rather small loudness variations can be detected under clean conditions [Riesz, 1928]. However, for a full speech-like signal, the threshold of detection will presumably be larger, as there are other distracting changes going on in the speech signal.[37] How much larger will it be? Experimental work is needed.

Experiments to establish a perceptual limit to a channel capacity are often synthesis-based experiments, where different variants of an utterance are prepared then presented to a subject who responds to the stimuli. Synthesis experiments can use a broad variety of responses, ranging from direct state-

[37] For example, the spectrum changes from syllable to syllable if the vowels are different. Can people reliably perceive a loudness change when the shape of the spectrum is changing at the same time? Can people reliably compare the loudnesses of two adjacent syllables, even if there is a distracting loudness dip in between (as might be caused by a /t/ or /m/)?

ments about language[38] to indirect decisions[39] to reaction time experiments. In the end, a button is pushed; the subjects classifies the stimulus as "same" or "different", "prominent" or "not". Many of these experiments share a common assumption, a weak link: they assume that subject has accurate conscious access to his or her language processing.

7.1.1 Conscious Access to the Language Mechanism

It should not be assumed that our consciously produced reports (e.g. saying "that is a question") about speech that we hear exactly reflects either our internal representations of the speech or how we would respond to the same speech in a more natural situation such as a dialogue. It is possible that we might *call* something a question when asked to classify an utterance, but if we encountered that utterance in a real dialogue, we might not *treat* it as a question. For instance, we might not answer it. The reverse is also possible.

In other words, one[40] does not normally think about the language one hears – one responds to it. Thus, any experimental situation which forces the subject to think about the language is artificial to some degree.

Consequently, the ideal experiment does not ask the subject to think about the stimuli; it observes the subject's behaviour in a situation as close to natural conversation as possible. Certain reaction time measurements match this requirement well: one could imagine asking subjects to "point to the dog", and correlating the time it takes them to respond with changes in the prosody of the stimulus.

An interesting class of experiments involves having the subject mimic stimuli that are presented. By looking at differences between the stimulus and the response, you can study details of speech perception and production. Mimicry is a normal linguistic process[41] and does not require a conscious classification of the stimulus, nor a conscious decision about what kind of response to produce.

Indeed, one particularly interesting question is to what extent we have good conscious access to our language mechanism. While it seems likely that

[38] E.g. "push 'p' if the verb seems prominent, and 'n' if not, then hit the 'enter' key to receive the next prompt."

[39] "You will be presented with two pictures; please choose the one that matches the sentence." ... "The greenhouse is full of plants." ... "The green house is full of plants."

[40] E.g. a non-linguist.

[41] Especially common before adolescence.

we have some access, its limits are unknown. It is an important question because so much of linguistics depends on introspection. Thus, if introspection were shown to be even occasionally invalid, much of linguistics would follow.

7.1.2 Forced Classification Experiments

Forced-classification experiments are interesting to the extent that the choice corresponds to a psychologically real distinction. For instance, imagine that a subject is required to say whether a syllable is prominent or non-prominent, but the brain's representation of prominence has three classes. What does the subject do? If the three internal classes form a one-dimensional sequence, then one can hope that most subjects will respond "non-prominent" to class 1, and "prominent" to class 3, but some subjects might lump classes 1 and 2 together, others might lump 2 and 3 together, and still others might waffle back and forth.

If prominence were a real-valued, rather than a discrete quantity, then the two-alternative forced-classification task causes the subjects to impose an arbitrary boundary on their internal prominence continuum.[42] In that case, the experiment would be studying this arbitrary boundary. The results then might say something about how people choose such boundaries, and perhaps how they dynamically adjust them during the experiment [Triesman and Williams, 1984, Warren, 1985], but not necessarily much about the intrinsic properties of the language.

The underlying problem is that we do not understand the relationship between discrete linguistic categories and human perception. It has been tempting to assume that linguistic categories are "wired" into our brains, through a mechanism like categorical perception [Harnad, 2003]. However, categorical perception, with the discrimination peak that is its defining feature, has been elusive in human language (see reviews in Schouten et al., 2003; Plomp, 2002, pp. 137–141; Gussenhoven, 1999). Even in situations where subjects can make sharp distinctions between classes, they are still able to discriminate within a class [Rosner, 1984, Ladd and Morton, 1997], so it is certainly not true that everything within a category is perceptually identical. The perceptual magnet effect [Kuhl, 1991, Guenther and Gjaja, 1996] is likely relevant, but also falls short of explaining how linguistic classes

[42] This can be described as a "rating scale experiment"; see references on Signal Detection Theory, such as Wickens [2001], Macmillan and Creelman [1991] and references therein.

might connect to underlying psychological reality. Basic experimental studies are needed.

Studies that do not wish to focus on the basic mechanisms of perception and linguistic categories should attempt to show that their forced-classification boundary is linguistically relevant. One possibility is to pair a two-alternative forced-classification experiment with a companion three-alternative experiment to check the sharpness of the boundary (e.g. add "maybe" to "yes" and "no"). The two experiments could be compared in many ways, but generally, the validity of the two-alternative experiment would be enhanced if the number of responses in the third alternative ("maybe") were fairly small and taken proportionally from the other two alternatives.

Some boundaries between categories are strongly embedded in our minds and resistant to manipulation, while others are ephemeral – easily set and easily moved. An ephemeral boundary could be created by an experiment that asks subjects to judge whether objects are (for instance) taller or shorter than their shoulder. Such a boundary can be learned in the course of an experiment and has little or no relevance outside the experiment.[43]

It might be expected that certain phoneme boundaries are the on the opposite extreme: highly stable boundaries that are not learned in the course of the experiment, and cannot be easily moved. The fact that it can take adults a long time to adapt to a new dialect and much effort to speak it fluently suggests that certain boundaries between classes in language, such as vowel boundaries, may be rather inflexible.

This distinction between ephemeral and stable boundaries has not been much studied, but it is a relevant question that could be asked and experimentally answered for any linguistic distinction. There are two plausible techniques for studying the stability of a boundary in a forced-classification experiment: one could measure to what extent the boundary between the two alternatives is affected by the conditions of the experiment, or to what extent it is learned in the course of the experiment. For instance, in an experiment studying a question/statement distinction, with experimental sentences embedded amidst fillers, one could change the fraction of filler sentences that are questions, and see if the boundary for the experimental sentences moves. (See Warren, 1985 and Eisner and McQueen, 2005 for a review of experiments

[43] The position of such a boundary is irrelevant, though one might hope that the width of the boundary or the process by which it is learned might say something broadly useful about perception.

studying changing boundaries.)

7.1.3 Speech Synthesis as a tool

Experiments that use synthetic speech suffer from the problem that synthetic speech, especially if it has been manipulated, does not sound very natural. Synthesized speech based on LPC, diphone, or formant synthesizers can be quite intelligible, but would rarely be confused with speech from an actual human (e.g. Donovan and Woodland, 1999).

This leads to systematic errors in experiments that use it; the experiment gives information on the perception of synthetic speech, rather than human speech. Unfortunately, we do not know how big a difference this makes, because there are very few experiments that compare natural and synthesized speech,[44] with the exception of Bailly [2003]. Schouten et al. [2003, §3] provides a discussion.

A crude estimate of the magnitude of the systematic errors can be made because the difference between natural and diphone/LPC/formant synthesized speech is much larger than differences between individual speakers (e.g. Reynolds et al., 2002), and perhaps comparable to differences between dialects. Thus, the systematic errors that synthesized speech introduces into perceptual measurements might be as large as dialect-to-dialect differences in that measurement.

7.1.4 Other things to avoid

Experimental subjects should be representative speakers of the language. Linguists, phonologists and phoneticians are expected to be very aware of language; phoneticians are trained to be able to analytically listen to speech. Consequently, they are highly unusual, and if one wants the research to be

[44] Other than acceptability judgements and intelligibility tests, that is. While acceptability and intelligibility scores have their uses for the engineering development of speech synthesizers, they are broad measures, involving an unspecified raft of perceptual and mental processes. Further they are highly subjective and dependent on experience and exposure. This combination makes them nearly useless for a scientific understanding of to what extent synthesized and natural speech might be perceived differently. (As a personal example, I recall joining the speech synthesis group at Bell Labs, and being shocked by the contrast between the positive descriptions of synthesizer performance I heard from group members and the unimpressive sounds I heard from the synthesizer. A year later, despite only modest technical advances, the synthesizer sounded much better to my adapted ears.)

broadly useful, they should not be used as subjects. Additionally, if an author of the paper is a subject or produces stimuli, the possibility is raised that bias may be introduced into the results. The author may attend to different aspects of the speech than a naive subject, or may speak differently from someone who is unaware of the goals and hypotheses involved in the experiment.

7.2 Channel Capacity Experiments – Production

Perception is not necessarily the weak link in the communication channel. Speech production could well limit the information transferred under many circumstances. For instance Shih and Kochanski [2000] and Shih et al. [2001] showed that in Mandarin[45] the inability of the larynx to adjust f_0 instantaneously can lead to dramatic intonational coarticulation, converting an underlying falling tone into a observed rise. In this case, the amount of information that can be communicated via f_0 is limited by the ability of the larynx to respond rapidly.

7.2.1 Inadvertently Biassing the Subjects

The experiment needs to be designed so that the subjects are not "told" what results to produce, even implicitly. The experimenter needs to avoid giving cues and feedback[46] that might let the subjects deduce the experimenter's desired results, and possibly influence the measurements. Subjects are typically very willing to adapt themselves to experimental requests (c.f. Milgram, 1963), so any feedback from the experimenter might well be adopted by the subjects as an instruction. The resulting experiment could become a self-fulfilling prophecy, unintentionally biassed toward the experimenter's desired results. The phenomena is known in psychology [Rosenthal and Rosnow, 1975], and presumably occurs in linguistic contexts too.

7.3 An Example

Here is a blatantly bad example of an experimental procedure to avoid, with comments in *italics*.

[45] A tone language; the Beijing dialect of Chinese.

[46] Cues can be subtle, as shown by Pfungst [1911], Miklósi et al. [1998]. Humans will likely be at least as successful as animals at reading the experimenter's intent.

- The goal: To understand of the intonation of a list of items.

  This goal contains the implicit assumption that there is a characteristic intonation produced whenever a person reads a list of items. That assumption may not be particularly accurate or useful. An alternative experiment might involve analyzing speech from naturally occuring text, and looking for common patterns in the intonation of lists.

- The task: Subjects are asked to read a few sentences in the form "I like raspberries, blackberries, gooseberries, and strawberries." The length of the list is varied, as is the order and type of item.

  Lists with more than three items are extremely rare in normal text. It is possible that a naive subject has never actually tried to read any longer list. Further, it is possible that the correct intonation for a list is dependent on the context, so that a list in different dialog acts (etc.) could have different intonation.

- The problem: The subject doesn't reliably produce the "correct" intonation. Sometimes it may be blatantly wrong (e.g. a question-like intonation), or sometimes fairly subtly wrong.

  There are two problems here: The definition of "wrong" comes from the experimenter and is therefore biassed by his expectations; Even assuming a single, well-defined intonation pattern for lists, the distinction between normal variation and errors is unclear. It's probably best to be conservative, leaving all but the most blatant mistakes cases in the data set, and setting out before-hand (based on pilot experiments) a set of simple, clear rules for rejecting a production. A good rule for prosody experiments might be to reject only if the subject speaks the wrong words or hesitates more than 200 ms.

- The cause: The experimenter decides do a full-factorial analysis so he does not want any missing data. He also realizes that the analysis assumes Gaussian-distributed data, and would be harmed if there are outliers. Consequently, he asks the subject to repeat some sentences.

  While issues of statistical analysis are outside the scope of this chapter, there are techniques ("robust statistics") for analyzing data that may contain outliers and for understanding the effect that outlying points will have on confidence intervals and hypothesis testing. Likewise, analysis procedures exist that allow for missing data. For difficult cases,

"bootstrap resampling" and "Monte-Carlo simulation" can be valuable. Consequently, there is often no need to manually reject data, and it is never best practice.

- Implicitly Teaching the Subject: The experimenter listens to the recording in real time, and if the sentence does not sound good, he tells the subject "Again, please."

  Consequently, the subject realizes that some of his productions are not good enough, and begins to modify his speech to accomodate the experimenter. However, since the experimenter is not providing an example, there is some hope that the experiment may end before the subject alters his behaviour too much.

  It is generally better practice to let the subject – instead of the experimenter – decide whether his utterance is good or bad. An experimental design that lets the subject choose will be not be biassed towards the experimenter's preferences. However, such a design might bias the experiment towards careful, formal speech.

- Accidentally Teaching the Subject: On a particularly long list, the subject has difficulty in producing a reasonable utterance: "I detest red dogs, red cars, red currants, ..., and red paint." (The underlying reason for his difficulty may be because the subject never actually says that kind of sentence in real life.) Perhaps the subject has a hard time keeping the words in the right order.

  The experimenter says "Again, please." four times in a row, and realizes that the experiment is in trouble. He says "The cars come before the currants. It's 'I detest red dogs, red cars *then* red currants."' The subject then has been inadvertently given some clues to the experimenter's preferred intonation.

  At this point, the game is lost. The appropriate action is to stop the experiment and drop the subject. It may also be worthwhile to re-think and perhaps re-design the experiment because the task may be too hard and un-natural. Spoken language (as normally spoken) is rarely hard to say; consequently, experiments involving difficult speech tasks are studying something other than normal English.

  This item also shows the value of a small informal pilot experiment. If one tries out the sentences on a small scale and finds that the sentences

are too complex to read reliably, it is much easier to re-think and re-design than if one is in the midst of a large project.

- Explicitly Coaching the Subject: The subject is now reading the words in the right order, but the intonation sounds completely unnatural. The experimenter says "OK. Now try that a little slower. It sounds like you're emphasizing 'red'. The noun is the important word: it's 'red **currants**', not '**red** currants'." From there, it is a small step to suggestions like "You're running into the bottom of your pitch range at the end. Maybe you should start a little higher?"

  This step crosses the like from bad practice to scientific misconduct.

- Data Selection: The experimenter now has recorded many versions of some productions. What to do? Obviously, he picks the one that sounds most natural (to his ears, anyway), and analyzes it.

  If data selection must be done, it is best done by a group of naive subjects. They can be asked to rate the naturalness of the utterances.

  Picking a single best or typical utterance has its risks no matter how it is done. Doing so will eliminate most variability from the data, and the variability can be as important as the typical utterance. For instance, in §7.3.2, the argument depends on comparing of the variability of two classes against the difference between typical examples.

If one follows this bad example to the end of the road, the resulting publication will have little value, even though the experimenter might not realize what has happened. Many of the problems above result from bad choices of an experimental goal and from an implicit requirement that the statistical analysis must be simple. If the earliest steps of experimental design are flawed, it becomes progressively harder to salvage an experiment as the design, experiment, and analysis proceed.

Pilot experiments can uncover unexpected problems at an early stage where they can be economically fixed. Additionally, some of these problems can be avoided by automating an experiment. An automated experiment can be more precisely described in the methods section of a publication and tends to have fewer places where the experimenter's biasses can influence the results.

7.3.1 Realism in Production Experiments

The need for naive and unbiassed subjects in production experiments is as strong as the need in perception experiments (§7.1.4). For instance, Albritton et al. [1996] found substantial differences in the prosody produced by naive speakers, as compared to trained speakers who were aware of the goal of the experiment. Likewise, Batliner et al. [2000] investigated emotional expression in different situations and built classifiers to measure how strongly the emotions appeared in prosody. They conclude that "The dilemma for our perspective is that the closer we get to real life applications, the less visible is emotion." Actors provided strong prosodic cues for emotions, but little was found for more realistic wizard-of-oz scenarios. Actors may be producing socially agreed-upon representations of emotions, which may not correspond to what people actually do when emotional.

7.3.2 Statistical Techniques in Production Experiments

It is also necessary to avoid is overly strong interpretation of statistically significant differences that might be found in an experiment.

For instance, suppose that a researcher measures f_0 near the centre of a prominent syllable, compares it to f_0 near the centre of other syllables, and finds a significant difference. What can she conclude? We can do the experiment with the IViE corpus. We find (as expected) that the average normalized[47] f_0 is significantly higher within 25 ms of the centre of a prominent syllable[48] than within 25 ms of the centre of a non-prominent syllable. Statistically, the difference is extremely significant ($z = 6.6$, $P < 10^{-6}$), but in reality, it is not particularly important.

Why? Because a listener could not use that information to reliably decide whether a syllable is prominent or not. The difference between the mean of all prominent and the mean of all non-prominent syllables corresponds to 6% of the speaker's f_0,[49] Such a difference, while probably perceptible (see §6.1), is not particularly large. Notably, it is smaller than the f_0 shifts that are found to induce prominence judgements in experiments (e.g. Gussenhoven et al., 1997; Rietveld and Gussenhoven, 1985; Terken, 1991). Thus, the f_0

[47] Normalized as per Kochanski et al. [2005], by dividing f_0 by the speaker's average f_0, then subtracting one. Thus, the speaker's mean f_0 maps to zero, and 10% above the speaker's mean f_0 maps to 0.1. The analysis here generally follows Kochanski et al. [2005].

[48] Defined as any syllable that is labelled with an IViE accent.

[49] About 7 Hz for a male or 12 Hz for a female speaker.

contrast by itself is probably not large enough to make listeners treat the syllable as prominent.

There is a good reason why listeners will not use a 6% f_0 shift to conclude that a syllable is prominent: the distributions of the f_0 of prominent and non-prominent syllables are broad enough so that they overlap (Figure 1). While the averages may be distinct, individual measurements are not.

Thus, the fundamental frequency of a single syllable does not provide much evidence or information. Imagine that a listener has just heard a syllable, and she notes that the normalized f_0 is 0.25 – well above the speaker's average. Can she reliably conclude that the syllable is prominent? No, because if she looks at the histograms near that f_0 value, she will find examples of both prominent and non-prominent syllables. If she guesses (based on that evidence) that the syllable was prominent, she would be wrong about a third of the time. For other f_0 values, the ability to correctly assign prominence to a syllable would generally be worse. Acoustics only carries information about linguistics if you can draw a border that effectively separates the prominent and non-prominent syllables.

The underlying issue is that an average over a corpus is not relevant to the listener, because the listener does not have the luxury of constructing an average. The listener has to decide whether a particular syllable is prominent or not, immediately, and simply cannot do it reliably on the basis of this f_0 measurement. This implies that statistics should be done on one element of a distibution at a time, not an average.

Thus, findings of statistically significant correlations between linguistic properties (like prominence) and averaged acoustic values does not generally prove much. It does not show that the listener makes use of that property, nor even that the listener *could* make use of that acoustic property. It merely shows that a difference exists, without necessarily showing it is big enough to transfer any information.

Statistical significance is of most interest when a null result was expected (e.g. Coleman, 2003 or Öhman, 1966 who found unexpectedly long-range coarticulatory effects). However, absent some plausible expectation of zero, a statistically significant difference is often uninteresting. Finding a small acoustic difference between two linguistic categories can be much like finding the winner in a race between a snail and a slug: no one would seriously expect them to crawl at exactly the same speed, so the mere fact of a small difference is unsurprising.

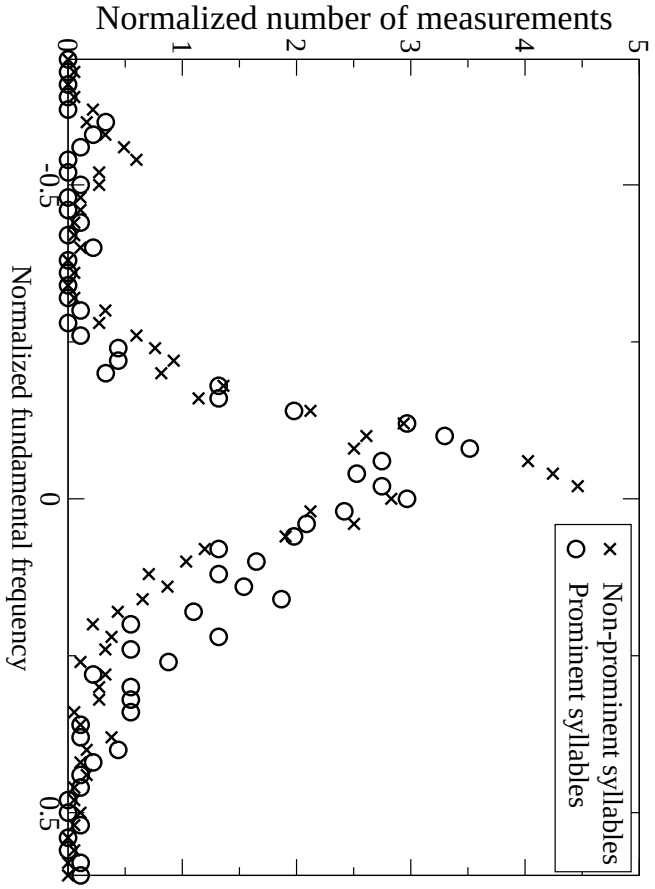

Figure 1: Histograms of central f_0 for prominent (O) and non-prominent (X) syllables in the IViE corpus.

7.4 Encoding Experiments

Encoding experiments are designed to show how a particular linguistic feature is encoded into acoustic properties. They can be either synthesis-based experiments (§7.1.3), where different variants of an utterance are prepared then presented to a subject who responds to the stimuli, or corpus-based experiments, where subjects produce speech which is then analyzed via statistical techniques after it is classified and/or labelled by a linguist.

Corpus experiments have an advantage over synthesis experiments in that the speech involved is produced by a human. However, corpus experiments suffer in that they require a human to consciously label speech (§7.1.1), and that the human labels are often forced to be in a few discrete categories.

Indeed, the very process of labelling bears an uncomfortable resemblance to a forced-classification (§7.1.2) perception experiment with the linguist playing the role of the subject (§7.1.4, § 7.2.1). While one can make a plausible case that phonologist's conscious classifications ought to mirror their own internal mental representations of speech, that is unproven, and there remains that possibility that phonology is an arbitrary classification, passed from phonologist to phonologist.[50]

Consequently, linguistic labels should be treated merely as convenient mid-points[51] between two objective measurements (e.g. text on one side and acoustic speech properties on the other). For instance, one could predict ToBI labels from a text. The labels could then be used to drive some model of intonation that would produce f_0 contours that would be compared with human-produced speech.

Papers that do not make the full link between two objective properties need to somehow make sure that a reasonably precise definition of the linguistic labels exists (preferably with examples) so that other groups can replicate the labelling (e.g. Beckman and Ayers, 1997; Grabe, 2001).[52]

[50] Experiments that can prove or disprove the reality of phonological objects (especially in the area of intonation and prosody) are clearly a very important step towards putting linguistics on a firm footing.

[51] Unless there is solid experimental evidence showing the psychological reality of the labels, or (ideally) some relationship between the labels and the way the brain stores or processes prosody.

[52] Unfortunately, even if a suitable definition of a labelling system exists, it may be difficult to convince a critical reader that labels were applied in a way that actually was consistent with the cited definition. Labelling systems like Grabe [2001] that have an extensive corpus of examples are (in principle) best, as one could check consistency with

Alternatively, the case might be made that the labels used are so obvious and universal that almost any labeller will get similar results. One could attempt to prove this by getting labels from several sources and showing that the labels are (at least mostly) in agreement. Such an argument becomes more powerful if the group of labellers is diverse. Ideally, if untrained native speakers can reliably apply the same labels to speech, it is reasonable to say that their labels are real properties of the language. An good example is the syllable in English: native speakers can count syllables with minimal instruction, and the syllable counts they produce will differ only occasionally. Thus, one can presumably trust that any labeller will produce an almost identical set of syllable counts.

The opposite extreme would be a labelling system that required extensive training by skilled linguists and still yielded substantial disagreement. In such a case, experimenters should be expected to devote a substantial amount of effort to proving that the labels that they use in an experiment are indeed consistent with the definition they cite.

Further, if labellers need to train together to achieve good agreement, the labelling system would have little value for scientific archival journals. To be a useful, permanent contribution to the scientific literature, a labelling system must be learnable from the literature. If a system depends on human-to-human transmission and consensus building, then when the originators of the system leave the scene or simply change with the years, it would become impossible to obtain a consistent set of labels. This would mean that future research would find it difficult or impossible to connect with and check older work.

7.5 How much and which information?

The question of which information is transmitted is part of the basic definition of a particular language or dialect. However, to measure the information content of a language, we need to know more than just (for example) that question/statement distinctions exist. We need to know when and how often the distinction needs to be made.

So far, in this chapter, I have applied a piecemeal approach to measuring the information content of prosody, but there is a holistic approach that

IViE by re-labelling some utterances in the corpus. One can also find examples in the corpus that show how to label many difficult or borderline cases. Unfortunately, the effort required to make use of such a labelled corpus is substantial.

could be tried to measure the information carried by prosody. The technique would be a variant of an experiment that [Shannon, 1951, Brown et al., 1992] used to measure the information content of English text.

Shannon presented text to subjects letter-by-letter, and asked them to guess the next letter, keeping track of the number of guesses they needed to get the letter right. If the subjects could (hypothetically) always get the answer on the first guess, then they did not actually need any information; the next letter was completely predictable from the previously-seen text, and it carried no information. On the other hand, if the previous text gave no information at all, and subjects would have to guess randomly from amongst 27 choices,[53] then an average subject would need about $\log_2(27) = 4.75$ guesses to get it right. In general, the average number of guesses provides a upper bound for the amount of information in each letter. In Shannon's experiment, it turned out to be an interestingly low bound, about 1 bit per letter, which implies (as one might expect) that people are very good at predicting what comes next, and that written language is highly redundant. This experiment is related to experiments by Warren and Marslen-Wilson [1987] where initial sections of words were presented to subjects for identification.

A similar experiment could be attempted for prosody. Although it may take a clever experimental design, one could imagine providing parts of spoken sentences to subjects, along with a full text, and asking them to predict the remaining intonation by speaking the entire sentence.[54] By analogy, one ought to be able to measure the information content of prosody by looking at how well people were able to predict the missing prosody. The better the prediction, the less information is carried by the missing part.

8 Conclusion

Thinking about language and prosody by comparing the amount of information that must be carried and the capacity of acoustic communication channels is likely to be fruitful. It provides motivation for a number of

[53] Twenty-six letters and space-or-punctuation, assuming no knowledge of relative letter frequencies.

[54] I thank the external reviewer for pointing out a difficulty with this experiment. Under some conditions, the speech that the subjects produce "... very often ends with a final rise. The obvious meaning of this rise is: Did I do it right?" To avoid this problem, I think you would want to make sure that there is not an obvious human authority figure to ask, and also to make it clear that their curiosity will be automatically satisfied very quickly.

experiments, and more importantly, provides a connection between several otherwise separate classes of experiments.

The main advantage of this viewpoint is that it allows connections between acoustic measurements and phonological features in the language. To be a part of the language, each phonological feature must have an encoding into some property of the sound. That encoding must be within the capabilities of human articulators, and it must be within the limits of human perception.

In principle, the maximum amount of information that can be carried in the acoustic signals from human to human can be determined. This limits the amount of phonological information that can be transmitted in each acoustic channel. This leads to a scientifically useful competition between features: absent this limit, it is very easy to propose that prosody carries any number of features, and likewise easy to show that acoustic properties are correlated to some degree with all of them. However, considerations of information can tell you that it might not be possible to transmit all of them, which would force a careful evaluation of what is actually transmitted.

I have also outlined a number of useful experiments, and pointed out some limitations of traditional prosody experiments that need to be addressed in order to make their results solid and cleanly interpretable.

References

David W. Albritton, Gail McKoon, and Roger Ratcliff. Reliability of prosodic cues for resolving syntactic ambiguity. *J. Experimental Psychology: Learning, Memory, and Cognition*, 22(3):714–735, 1996.

G. Bailly. Close shadowing natural versus synthetic speech. *International J. of Speech Technology*, 6(1):11–19, January 2003. URL `http://dx.doi.org/10.1023/A:1021091720511`.

A. Batliner, K. Fischer, R. Huber, J. Spiker, and E. North. Desperately seeking emotions: Actors, wizards and human beings. In Cowie et al, editor, *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 195–200, 2000. URL `http://nats-www.informatik.uni-hamburg.de/~fischer/isca00.ps`.

M. Beckman and G. Ayers. Guidelines for ToBI labelling. Technical report,

Linguistics Department, Ohio State University, 1997. URL http://ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf.

A. W. Black and A. J. Hunt. Generating f_0 contours from tobi labels using linear regression. In *Proceedings of ICSLP 96*, 1996. Note: this is not really a measurement of the size of segmental effects, but it is a check that assuming 10 Hz RMS segmental effects does not lead to obvious failures.

Dwight Bolinger. A theory of the pitch accent in English. *Word: Journal of the International Linguisic Association*, 7:199–210, 1958. ISSN 0043-7956. Reprinted in D. Bolinger, Forms of English: accent, morpheme, order, Harvard University Press, Cambridge, MA (1965).

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. C. Lai, and R. L Mercer. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40, March 1992.

R. Carlson and B. Granström. Perception of segmental duration. In A. Cohen and S. Noteboom, editors, *Structure and Process in Speech Perception*, pages 90–106. Springer-Verlag, Berlin, 1975.

J. Caspers. Who's next? The melodic marking of question vs. continuation in dutch. *Language and Speech*, 41(3-4):375–398, July–Dec 1998.

Aoju Chen. Language dependence in continuation intonation. In *Proceedings of the* 15th *International Congress of Phonetic Sciences*, 2003. URL http://www.let.kun.nl/gep/carlos/AojulCPhS15.pdf.

Chui-Kuang Chuang and William S-Y. Wang. Psychophysical pitch biases related to vowel quality, intensity difference, and sequential order. *J. Acoustical Society of America*, 64(4):1004–1014, 1978. URL http://dx.doi.org/10.1121/1.382083.

J. Coleman. Discovering the acoustic correlates of phonological contrasts. *J. Phonetics*, 31:351–372, 2003.

R. Cowie and R. Cornelius. Describing the emotional states expressed in speech. *Speech Communication*, 40(1–2):5–32, 2003.

A. Cutler, D. Dahan, and W. van Donselaar. Prosody in the comprehension of spoken language: a literature review. *Language and Speech*, 40 (Part 2):141–201, 1997. URL `http://psy.ucsd.edu/~dswinney/Courses/PSY218b/218b_Readings/Clutter,%2520Dahan%2520and%2520Donselaar%25201997.pdf`.

W. Ding and W. N. Campbell. Detection of sentence prominence using voice source parameters. *J. Acoustical Society America*, 100(4):2600, October 1996. URL `http://dx.doi.org/10.1121/1.417608`. Conference Abstract.

Robert E. Donovan and P. C. Woodland. A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and Language*, 13:223–241, 1999.

F. Eisner and J. M. McQueen. The specificity of perceptual learning in speech processing. *Perception and Psychophysics*, 67(2):224–238, 2005.

Gunnar Fant, Anita Kruckenberg, and Johan Liljencrants. The source-filter frame of prominence. *Phonetica*, 57:113–127, 2000. URL `http://dx.doi.org/10.1159/000028466`.

J. R. Firth. *The Tongues of Men*. Watts & Co, London, 1937.

J. R. Firth. Personality and language in society. *The Sociological Review*, xlii(2), 1950. Reprinted in J. R. Firth Papers in Linguistics 1934–1951, 177–189.

D. B. Fry. Duration and intensity as physical correlates of linguistic stress. *J. Acoustical Society of America*, 27:765–768, 1955.

D. B. Fry. Experiments in the perception of stress. *Language and Speech*, 1: 126–152, 1958.

W. R. Garner. An informational analysis of absolute judgements of loudness. *J. Experimental Psychology*, 46:373 – 380, 1953.

M. H. Giard, J. Lavikainen, K. Reinikainen, F. Perrin, O. Bertrand, and J. Pernier et al. Separate representation of stimulus frequency, intensity, and duration in auditory sensory memory: an event related potential and dipole-model analysis. *J. Cognitive Neuroscience*, 7:113–143, 1995.

E. Grabe. Intonational variation in urban dialects of english spoken in the british isles. In P. Gilles and J. Peters, editors, *Regional Variation in Intonation*, Linguistische Arbeiten., pages 9–31. Niemeyer, Tubingen, 2004.

E. Grabe, B. Post, and F. Nolan. *The IViE Corpus*. Department of Linguistics, University of Cambridge, Cambridge, UK, 2001. URL `http://www.phon.ox.ac.uk/~esther/ivyweb`. http://www.phon.ox.ac.uk/ esther/ivyweb.

E. Grabe, G. Kochanski, and J. Coleman. The intonation of native accent varieties in the british isles: potential for miscommunication? In Katarzyna Dziubalska-Kolaczyk and Joanna Przedlacka, editors, *English Pronunciation Models: A Changing Scene*, pages 311–337. Peter Lang, Bern, Switzerland, 2005. ISBN 3-03910-662-7.

Esther Grabe. IViE labelling guide. Manual, Oxford University Phonetics Laboratory, Oxford, UK, 2001. URL `http://www.phon.ox.ac.uk/~esther/ivyweb/guide.html`. Version 3; available at `http://www.phon.ox.ac.uk/~esther/ivyweb/guide.html`.

Esther Grabe. Variation adds to prosodic typology. In B. Bel and I. Marlin, editors, *Proceedings of the Speech Prosody 2002 Conference*, Aix-en-Provence, France, 2002. ISBN 2-9518233-0-4.

R. M. Gray and D. L. Neuhoff. Quantization. In S. Verdú, editor, *Information Theory: 50 Years of Discovery*, pages 281–339. IEEE Press, Piscataway, NJ, 2000. Reprinted from *IEEE Transactions on Information Theory* 4, October 1998.

F. H. Guenther and M. N. Gjaja. The perceptual magnet effect as emergent property in neural map formation. *J. Acoust. Soc. America*, 100:1111–1121, 1996.

C. Gussenhoven, B. H. Repp, A. Rietveld, H. H. Rump, and J. Terken. The perceptual prominence of fundamental frequency peaks. *J. Acoustical Society of America*, 102(5):3009–3022, 1997.

Carlos Gussenhoven. Discreteness and gradience in intonational contrasts. *Language and Speech*, 42(2-3):283–305, 1999.

K. Hadding-Koch. *Acoustico-Phonetic Studies in the Intonation of Southern Swedish*. C.W.K. Gleerup, Lund, Sweden, 1961.

Stevan Harnad. Categorical perception. In *Encyclopedia of Cognitive Science*. Nature Publishing Group/Macmillan, 2003. URL `http://eprints.ecs.soton.ac.uk/7719/01/catperc.html`.

Scott A. Huettel and Gregory R. Lockhead. Range effects of an irrelevant dimension in classification. *Perception and Psychophysics*, 61(8):1624–1645, 1999.

A. Huggins. How accurately must a speaker time his articulations? *IEEE Transactions on Audio and Electroacoustics*, 16(1):112–117, 1968.

A. W. F. Huggins. Just noticeable differences for segment duration in natural speech. *J. Acoust. Soc. America*, 51:1270–1278, 1972.

Walt Jesteadt, Craig C. Wier, and David M. Green. Frequency discrimination as a function of frequency and sensation level. *J. Acoustical Soc. America*, 61(1):169–177, 1977.

M. Kehoe, C. Stoel-Gammon, and E. H. Buder. Acoustic correlates of stress in young children's speech. *Journal of Speech and Hearing Research*, 38 (2):338–350, 1995.

D.-Y. Kim, P. A. Humblet, M. V. Eyuboglu, G. D. Forney, Jr., and S. Mehrabanzad. V.92: The last dial-up modem? *IEEE Transactions on Communications*, 52(1), January 2004.

G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. Loudness predicts prominence: Fundamental frequency lends little. *J. Acoust. Soc. of America*, 118(2):1038–1054, August 2005. URL `http://dx.doi.org/10.1121/1.1923349`.

Greg Kochanski and Chilin Shih. Automated modelling of Chinese intonation in continuous speech. In *Proceedings of Eurospeech 2001*. International Speech Communication Association, October 2001. URL `http://prosodies.org/papers/2001/automated_e2001.pdf`.

Greg Kochanski and Chilin Shih. Prosody modeling with soft templates. *Speech Communication*, 39(3-4):311–352, February 2003a. URL `http://dx.doi.org/10.1016/S0167-6393(02)00047-X`.

Greg Kochanski, Chilin Shih, and Hongyan Jing. Quantitative measurement of prosodic strength in Mandarin. *Speech Communication*, 41(4):625–645, 2003. URL `http://dx.doi.org/10.1016/S0167-6393(03)00100-6`. Draft on http://prosodies.org/papers/2003/automated_specom2003.pdf.

Greg P. Kochanski and Chilin Shih. A quasi-glottogram signal. *J. Acoustical Society of America*, 114(4):2206–2216, june 2003b. URL `http://dx.doi.org/10.1121/1.1608964`. Draft available at http://prosodies.org/papers/2003/pegg/Pseudo_EGG.pdf.

Patricia Kuhl. Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories. *Percept. Psychophys.*, 50: 93–107, 1991.

D. Robert Ladd. *Intonational Phonology*. Cambridge University Press, Cambridge, 1996. ISBN 0-521-47498-1.

D. Robert Ladd and Rachel Morton. The perception of intonational emphasis: continuous or categorical? *J. of Phonetics*, 25:313–342, 1997.

Mark Liberman and Ivan Sag. Prosodic form and discourse function. In *Proceedings of the Chicago Linguistic Society*, volume 10, pages 416 – 427, 1974.

N. A. Macmillan and C. D. Creelman. *Detection Theory: A user's guide*. Cambridge University Press, New York, 1991.

K. Maekawa. Phonetic and phonological characteristics of paralinguistic information in spoken japanese. In *Proceedings of the International Conference on Spoken Language Processing 98*, 1998. paper 997.

Christopher D. Manning and Hinrich Schütz̈le. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999. URL `http://nlp.stanford.edu/fsnlp/`.

À. Miklósi, R. Polgárdi, J. Topál, and V. Csányi. Use of experimenter-given cues in dogs. *Animal Cognition*, 1:113–121, 1998.

Stanley Milgram. Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67:371–378, 1963.

G. A. Miller. The magical number seven, plus or minus two. *Psychological Review*, 63:81–97, 1956.

Y. Morlec, G. Bailly, and V. Aubergé. Generating prosodic attitudes in french: data, model and evaluation. *Speech Communication*, 33(4):357–371, 2001.

S. E. G. Öhman. Coarticulation in VCV utterances: Spectrographic measures. *J. Acoustical Society of America*, 39:151–168, 1966.

C. B. Pedersen and G. Salomon. Temporal integration of acoustic energy. *Acta Otolaryngology*, 83:417 – 423, 1977.

O. Pfungst. *Clever Hans. The horse of Mr. von Osten.* Henry Holt, New York, 1911.

J. Pierrehumbert and J. Hirschberg. The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, editors, *Intentions in Communication*, pages 271–311. MIT Press, Cambridge, MA, 1990.

Reiner Plomp. *The Intelligent Ear: On the nature of sound perception.* Lawrence Erlbaum Associates, Mahwah, NJ, 2002. ISBN 0-8058-3867-8.

I. Pollack. The information of elementary auditory displays. *J. Acoustical Soc. America*, 24:745–749, 1952.

I. Pollack. The information of elementary auditory displays II. *J. Acoustical Soc. America*, 25:765–769, 1953.

Hugo Quené. On the just-noticeable difference for tempo in speech. *J. Acoustical Soc. America*, 115(5):2607, 2004. URL `http://lotos.library.uu.nl/publish/articles/000117/bookpart.pdf`.

M. E. Reynolds, C. Isaacs-Duval, and M. L. Haddox. A comparison of learning curves in natural and synthesized speech comprehension. *J. Speech, Language, and Hearing Research*, 45: 802–810, August 2002. URL `http://www.asha.org/NR/rdonlyres/2B0DAACA-377D-4E44-96E1-B86EA6C3B224/0/17497_1.pdf`.

R. R. Riesz. Differential intensity sensitivity of the ear for pure tones. *Physical Review*, 31(5):867–875, May 1928.

A. C. M. Rietveld and C. Gussenhoven. On the relation between pitch excursions and prominence. *J. Phonetics*, 13:299–308, 1985.

R. Rosenthal and R. L. Rosnow. *The volunteer subject*. Wiley-Interscience, New York, N.Y., 1975. 266p.

B. S. Rosner. Perception of voice-onset-time continua: A signal detection analysis. *J. Acoust. Soc. of America*, 75(4):1231–1242, 1984.

Jean-Luc Rouas, Jérôme Farinas, François Pellegrino, and Régine André-Obrecht. Modeling prosody for language identification on read and spontaneous speech. In *Proceedings, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP '03)*, volume 6, pages 40–43, 2003. ISBN 0-7803-7663-3. URL `http://dx.doi.org/10.1109/ICASSP.2003.1201602`. Volume 6 or 1.

Bert Schouten, Ellen Gerrits, and Arjan von Hessen. The end of categorical perception as we know it. *Speech Communication*, 41:71–80, 2003. URL `http://dx.doi.org/10.1016/S0167-6393(02)00094-8`.

Marc Schröeder. Emotional speech synthesis: A review. In *Proceedings of Eurospeech*, pages 561–564, 2001.

C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, July, October 1948.

C. E. Shannon. Probability of error for optimal codes in a gaussian channel. *Bell System Technical Journal*, 38:611–656, 1959.

C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, 1949.

Claude E. Shannon. Prediction and entropy of printed english. *Bell System Technical Journal*, 3:50–64, 1951.

C. Shih and G. P. Kochanski. Chinese tone modeling with Stem-ML. In *Proceedings of the International Conference on Spoken Language Processing*, volume 2, pages 67–70, 2000. URL `http://prosodies.org/papers/2000/tonemodel_2000.pdf`.

C. Shih, B. Möbius, and Bhucana Narasimhan. Contextual effects on consonantal voicing profiles: A cross-linguistic study. In *Proceedings of the* 14^{th} *International Congress of Phonetic Sciences (ICPhS)*, volume 2, pages 989–992, August 1999.

Chilin Shih, G. P. Kochanski, E. Fosler-Lussier, Melody Chan, and Jia-Hong Yuan. Implications of prosody modeling for prosody recognition. In M. Bacchiani, J. Hirschberg, D. Litman, and M. Ostendorf, editors, *Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pages 133–138. International Speech Communication Association, October 2001. URL `http://prosodies.org/papers/2001/ASRprosody2001.pdf`.

Rosaria Silipo and Steven Greenberg. Prosodic stress revisited: Reassessing the role of fundamental frequency. In *Proceedings of the NIST Speech Transcription Workshop*, May 2000.

Rosario Silipo and Steven Greenberg. Automatic transcription of prosodic stress for spontaneous English discourse. In *Proceedings of the XIVth International Congress of Phonetic Sciences (ICPhS99)*, pages 2351–2354, August 1999.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, April 2003.

Aagath M. C. Sluijter, Vincent J. van Heuven, and Jos J. A. Pacilly. Spectral balance as a cue in the perception of linguistic stress. *J. Acoustical Society of America*, 101(1):503–513, January 1997.

Agaath M. C. Sluijter and Vincent J. van Heuven. Spectral balance as an acoustic correlate of linguistic stress. *J. Acoustical Society of America*, 100 (4 part 1):2471–2485, October 1996.

S. S. Stevens. Perceived level of noise by Mark VII and decibels. *J. Acoustical Society of America*, 51(2 (part 2)):575–602, 1971.

J. Sundberg. Maximum speed of pitch changes in singers and untrained subjects. *J. Phonetics*, 7:71–79, 1979.

J. 't Hart. Differential sensitivity to pitch distance, particularly in speech. *J. Acoust. Soc. America*, 69(3):811–821, 1981. URL `http://dx.doi.org/10.1121/1.38559`.

J. 't Hart, R. Collier, and A. Cohen. *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. Cambridge University Press, Cambridge, 1990.

Jacques Terken. Fundamental frequency and percieved prominence of accented syllables. *J. Acoustical Society of America*, 89(4):1768–1776, April 1991.

J. t'Hart, R. Collier, and A. Cohen. *A perceptual study of intonation: An Experimental-phonetic approach to speech melody*. Cambridge University Press, Cambridge, UK, 1990.

M. Triesman and T. C. Williams. A theory of criterion setting with an application to sequential dependencies. *Psychological Review*, 91(1):68–111, 1984.

Alice E. Turk and James R. Sawusch. The processing of duration and intensity cues to prominence. *J. Acoustical Society of America*, 99(6):3782–3790, 1996. URL `doi:10.1121/1.414995`.

J. v. Laziczius. Probleme der phonologie. In *Proceedings of the Second International Congress of Phonetic Sciences*, 1935.

Paul Warren and William Marslen-Wilson. Continuous uptake of acoustic cues in spoken word recognition. *Perception and Psychophysics*, 41(3): 262–275, 1987.

Richard M. Warren. Criterion shift rule and perceptual homeostasis. *Psychological Review*, 92(4):574–584, 1985.

Thomas D. Wickens. *Elementary Signal Detection Theory*. Oxford University Press, Oxford, UK, 2001. ISBN 0-19-509250-3.

Yi Xu and Xuejing Sun. Maximum speed of pitch change and how it may relate to speech. *J. Acoust. Soc. Am.*, 111(3), March 2002. URL `http://dx.doi.org/10.1121/1.1445789`.