View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Hochschulschriftenserver - Universität Frankfurt am Main

Holger Quast

Automatic Recognition of Nonverbal Speech

An Approach to Model the Perception of Para- and Extralinguistic Vocal Communication with Neural Networks

Drittes Physikalisches Institut,
Georg August Universität Göttingen

and
Machine Perception Lab, Institute for Neural Computation
University of California, San Diego

To my parents,
Helmut and Heide Quast,
And to the memory of my uncle,
Horst Quast

Contents

1 Introduction

A couple of years ago, working on some lab reports while the TV was running in the background, I was listening to a political debate with half an ear. The discussion was clearly dominated by one person. Not paying attention to what was said, I assumed this person was the most competent one in the field. When I *did* listen later on, I was surprised to find this speaker was neither the most knowledgeable one nor the superior rhetorician (verbally), but managed to lead the discussion solely by the way he talked, his *non*verbal vocal demeanor. As Oscar Wilde aptly put it, *in matters of grave importance, style, not sincerity, is the vital thing*.

I started thinking about this phenomenon, and immediately a number of useful software gadgets came to my mind that could originate from the understanding of *prosody*, i.e. the intonation and rhythm of speech; programs that could possibly assist in speech therapy, be useful in human-computer interaction, affective computing, and speaker training.

At the University of Göttingen, I joined Manfred Schroeder's and Hans Werner Strube's Speech and Neural Networks group where Professor Schroeder kindly and expertly acted as my Master's thesis advisor for the topic I had on my mind: pattern recognition on nonverbal speech. In San Diego, Terry Sejnowski and Javier Movellan invited me to UCSD's Institute for Neural Computation and provided me with excellent counsel on pattern recognition. It is in Javier's Machine Perception Lab where I carried out the biggest part of this research.

The work presented here consists for the most part – with small additions that didn't make the January deadline – of my Physics Master's thesis which I submitted at Göttingen at the beginning of this year. Style or sincerity, limbic system or cortex, nonverbal or verbal speech, I hope the reader finds this text both affecting and informative.

Göttingen,
Spring 2001

Holger Quast

holcus@physik3.gwdg.de
http://mplab.ucsd.edu/~holcus

1.1 Motivation

Vocal communication incorporates a *verbal* and a *nonverbal* communication channel: whereas the verbal part of speech is represented by words, the nonverbal channel is carried by the *prosody* – i.e. the stress and intonation patterns of the utterance – and holds information about the speaker's physical state, emotions, the attitude towards the object of the conversation, etc.

Long before mankind learned to communicate through words, our progenitors vocalized utterances for a variety of purposes[1] such as to express affect, contact family members, threaten an enemy, or, as Charles Darwin, the first prominent modern researcher of the expression of emotions in man and animals (Darwin 1872), hypothesized, "to charm or excite the female". In today's speech, nonverbal information is just as important for effective communication. How a message is received depends to a large extend not only on *what* was said, but also on *how* it was said.

Human speech recognition works in a multimodal manner and on multiple interacting levels. We combine a deep knowledge about grammar, the meaning of previously said words, the speaker, the culture, and so forth with the heard sounds, the prosody, and gestures to parse a speech stream and extract meaning. If one took away a listener's knowledge about meaning, grammar, and use of prosody of a language – say by having a person spot words from a dictionary in a language the listener doesn't understand – experience shows the ability to transcribe speech to text falls short of a native speaker's. Current computer speech recognition (or, to be more precise, word recognition) systems for the most part suffer from the same impediment, they only exploit the verbal channel. As a consequence, both their performance and functionality as natural human–computer interfaces are strongly limited. If a computer wanted to interact with a user, it seems reasonable it would make use of the many modalities that human communication offers, such as the *linguistic* (verbal) data but also information about the speaker's state, if he was serious or joking, emotionally agitated or calm, etc.[2] A good speech recognition system could use the affective message of an utterance to assign probabilities to words in cases where the right selection based on acoustic clues is difficult. For instance, if a software is to decide between the words "fun" or "gun" and the nonverbal content signals a hectic atmosphere, the latter candidate appears to be the rational choice.

The use of prosody becomes even more crucial in *meaning* recognition. In irony, for example, the words usually carry the opposite message of their literal content. If a human–computer speech dialogue system assesses the machine's behavior considering only the verbal information of a user's contemptuous "Well done computer, you just deleted my most important file!", this clearly does not capture the intended message. The common use of :-) smiling and :-(frowning *emoticons* in emails is another example showing that the linguistic channel is often not enough to convey the whole meaning of a message.

At the other end of vocal dialogue systems lies speech synthesis. Even high-quality artificial speech generated from real voice recordings sounds unnatural after a while because it lacks variability and the ability to adjust to a situation on an affective level. If the emotional content of a conversation is understood by the machine, the speech output can adapt to that situation. Janet Cahn created such an *Affect Editor* that is able to produce synthesized speech with recognizable and, at times, natural affect for the six basic emotions angry, disgusted, glad, sad, scared, surprised (Cahn 1989). Murray and Arnott (1993, 1996) describe a number of correlates between affective content and acoustic parameters that modified the prosody of

[1] cf. Tembrock 1975, Scherer 1985

[2] Reports of patients that have suffered frontal lobe brain damage and, as a result of this, are emotionally impaired, suggest that emotional capabilities of computers are not merely an interface issue, but that machines need emotions or similar concepts to perform intelligent tasks like decision making or scheduling (Damasio 1994, see also Picard 1997).

the DECtalk speech synthesizer as needed. Even a lower-level, less elaborate system could add to the naturalness of computer speech production using a phenomenon called *internal simulation* (Fiukowski 1984; Eckert, Laver 1994): during the course of a conversation, we partially adopt certain aspects of the other person's prosody, for instance average pitch, speech rate, loudness and so forth. The measurement of these parameters is fairly robust, and a speech synthesizer can then be tuned accordingly.

The quantity perceived loudness, which is derived from the absolute loudness model presented in section 3.4, is a good indicator for agitation and therefore can find use in stress monitoring devices, e.g. for air traffic radio communication or driver speech dialogue systems (Fernandez and Picard 2000). There is substantial interest in stress monitoring applications as the number of publications shows, but so far only parameters like *microtremors*, i.e. inaudible modulation in the 4–12 Hz range (as measured by the Teager energy operator $T(s_n) = s^2(n) - s(n-1)\,s(n+1)$; Teager & Teager 1990; Cairns and Hansen 1994), floor f_0 (Tolkmitt and Scherer 1986) and maximum f_0 (Protopapas and Lieberman 1996) are mentioned as moderately accurate correlates. It is also desirable to know the effects of stress on speech in order to be able to design speech recognition software that is resistant to these changes (Steeneken, Hansen 1999).

People who were born deaf often speak with an unnatural prosody since they lack acoustic feedback. With programs that can measure nonverbal content, e.g. how happy, angry, or confident a speaker sounds, speech therapy aids – as a tireless and inexpensive supplement to human experts treatment – for these patients can be built that lead them towards a more natural prosody and can train them to express their emotions understandably. The same programs would also be useful to assist in the therapy of patients with Autism or Asperger Syndrome who have problems understanding and coding emotions.

Once the prosodic features that lead to an impression are understood, utterances can be composed that lead to a very pure and intense perception of only this impression. These speech samples can then be used in cognitive neuroscience to study what events go on in the brain during the understanding and processing of affect.[1]

Similar tools can be used to help language students learn both pronunciation and prosody of a foreign language. A lot of people feel uncomfortable speaking in a foreign language because they fear their accents might make them target for ridicule. With a software tutor, students can learn in the privacy of their own home with a most patient teacher.

Whereas effective nonverbal vocal communication is important in most conversations, these skills become a crucial prerequisite for a large number of professions: lawyers need to be perceived as convincing, teachers as interesting, politicians as competent, managers need to convey leadership ability, and so forth to be successful. With the proper prosody evaluation tools it seems possible to practice the desired impression and optimize the speaker's *sending accuracy* so that a message can be received as intended.

[1] Neuroscientists often use very strong magnetic fields to stimulate areas of interest in the brain, and then investigate what influence corresponds to which part of the cortex. As Sir Francis Crick once suggested to me, listening to affective speech to elicit emotions during a functional magnetic resonance scan sounds like a by far more natural stimulus.

1.2 Goals

The overall intent of this project is to devise and test a framework for machine nonverbal speech perception that could be used in applications such as the ones mentioned above. Specifically,

1. to create a versatile database that contains a number of natural sounding speech recordings that can be used to generate a variety of different impressions in a listener,
2. to extract digital signal processing (dsp) parameters that possibly carry the nonverbal information in the speech samples and lead to a memory efficient data representation,
3. to evaluate the recordings with respect to their nonverbal content, and
4. to perform pattern recognition on the psycholinguistic and the signal processing data to see if the nonverbal content of speech is indeed carried by the extracted dsp-parameters and can be learned by a pattern recognition scheme.

1.3 Overview of this work

In order to train a pattern recognition system to recognize affective speech, it must be provided with examples. The next part, Chapter 2, describes how a database of speech recordings of German actors and nonactors is assembled. These recordings are then evaluated in the seven categories pleasant – happy – confident – strong – agitated – leadership – angry by Californian listeners. The results are normalized and refined to yield a measure of the affective content for each recording and category. A second value, a confidence factor, is derived that describes how high listeners' agreement was in each category, for each recording. The results are qualitatively compared to scores from German judges who received the evaluation program and submitted their scores over the internet.

The third chapter deals with the *acoustic* parameters that represent a speech recording. These parameters fall into the categories fundamental frequency, intensity, and spectral composition. To extract these, a pitch tracker is developed, a new psychoacoustically motivated speech loudness model is introduced, and other signal processing tools are build that lead to a total of 18 acoustic parameters representing each recording.

To see if the affective, psycholinguistic value of a speech recording can be represented as a function of its signal processing parameters, neural networks are used as pattern recognition engines to map one set of features onto the other one. This process is elaborated in Chapter 4. The network is trained by showing it example pairs of acoustic and affective data, and, if it has learned successfully, the network is able to generalize, i.e. see new examples, and be able to assess its affective content. Since listeners' agreement strongly fluctuated in the evaluation of the speech recording, the neural networks are programmed to consider the quality of a data point when learning from it by means of the confidence factor computed in the second chapter. A new psychoacoustically motivated data representation technique called lombada, based on the loudness model developed in Chapter 3, is introduced. With the lombada technique, the prosodic information can be stored at a fraction of the space occupied by the original recording.

Chapter 5 shows how the affect recognizer developed here was successfully applied to build a nonverbal speech dialogue interface as can be used for a pet robot.

Chapter 6 Concludes this work, discusses the findings and gives an outlook to possible future investigations in this field.

2 Psycholinguistic Evaluation

This chapter describes the speech data, that is used later to train the nonverbal-speech recognizer, on a psycholinguistic level. At first, a number of speech recordings is collected both from actors and nonactors. These samples are evaluated by American and German listeners who report their impressions for each recording in 7 categories. The scores are then normalized to find a useful description of affect for each datapoint.

2.1 The Speech Database

The set of data used in this project contains 145 recordings of the eight-sentence German monologue noted below; 117 of professional actors producing the monologue picturing themselves in different given situations, 28 of nonactors who spoke in their natural registers. The 13 actors of the *Deutsches Theater Göttingen* (6 women, 7 men) ranged in age from 27 to 66 with an average of 38 years. The nonactors (2 women, 12 men) were 21-72 years old, the average age was 41. The speakers came from a broad range of backgrounds and the recordings took place in different environments (workplace, home, a friend's home etc.) Three actors had slight accents, namely Bavarian, Austrian, and Swiss. Some people had very faint local dialects, but the majority spoke High German. Age, sex, and profession of the speakers were noted.

The actors were not asked to produce a specific expression, but to imagine themselves in different scenarios and then speak the monologue in a suitable manner. This way, the desired expression categories were not presented in an artificially detached way but together with the complete affective content that would naturally occur. This also takes the interpretation of the nonverbal content away from the actors and leaves it to the listeners who later evaluate the recordings. Since the evaluators judge the recordings in categories related to leadership ability, confidence, agitation etc., the situations given to the actors are supposed to contain this affect in varying degrees and included talking with family or friends, speaking to employees of one's company, or addressing a parliament as the head of the government. If they came up with other expressions or situations, they were encouraged to portray those as well in additional recordings.

The text tries to combine a variety of different sentence structures without compromising coherence as a whole. It contains two exclamations, one of which is an interjection (8), the other one a request (5); one question (7), and regular sentences, two of which (4,6) have the same number of syllables and intonation structure to allow for training on one sentence and test generalization ability on the other.

> *(1) In der Vergangenheit ist schon einiges an guter Vorarbeit geleistet worden. (2) Die Ziele, die wir jetzt verfolgen, sind die gleichen und müssen auch auf die gleiche Weise behandelt werden. (3) Unsere Aufgabe ist nun, noch einmal die Zeiteinteilung durchzusehen. (4) Sie überprüfen dann das Weitere. (5) Bitte notieren Sie die Punkte, die Sie heraussuchen, und tragen Sie uns diese vor! (6) Wir erledigen alles Andere. (7) Glauben Sie, daß Sie das schaffen? (8) Gut!*

The length of one recording averages about 30 seconds.

The speech samples were recorded with an active microphone worn on the speakers' heads to keep the mouth-to-microphone distance constant. The signal was augmented by a custom-made preamplifier, recorded on DAT with a sampling rate of 48 kHz and stored as 16-bit mono linear pcm 48-kHz .raw data file.

2.2 Nonverbal Information in the Speech Signal

2.2.1 The Linguistic, Paralinguistic, and Extralinguistic Channels

The information communicated in spoken language can be categorized as *linguistic*, *paralinguistic*, or *extralinguistic* (Eckert, Laver 1994), see Figure 2.1. Whereas the verbal content, the actual meaning of the words, is thought of as linguistic information, the extralinguistic channel contains information about the speaker's basic state, e.g. a big person with a large vocal tract will usually have a lower voice than a child. Some extralinguistic parameters are also determined by the culture of the speaker. Compare for instance Swedish – where the pitch vividly goes up and down – to American English, where pitch changes by far less throughout a sentence.

The paralinguistic channel carries information about momentary deviations from the usual (extralinguistic) baseline, such as whispering in a situation that calls for silence, or expression of emotions.

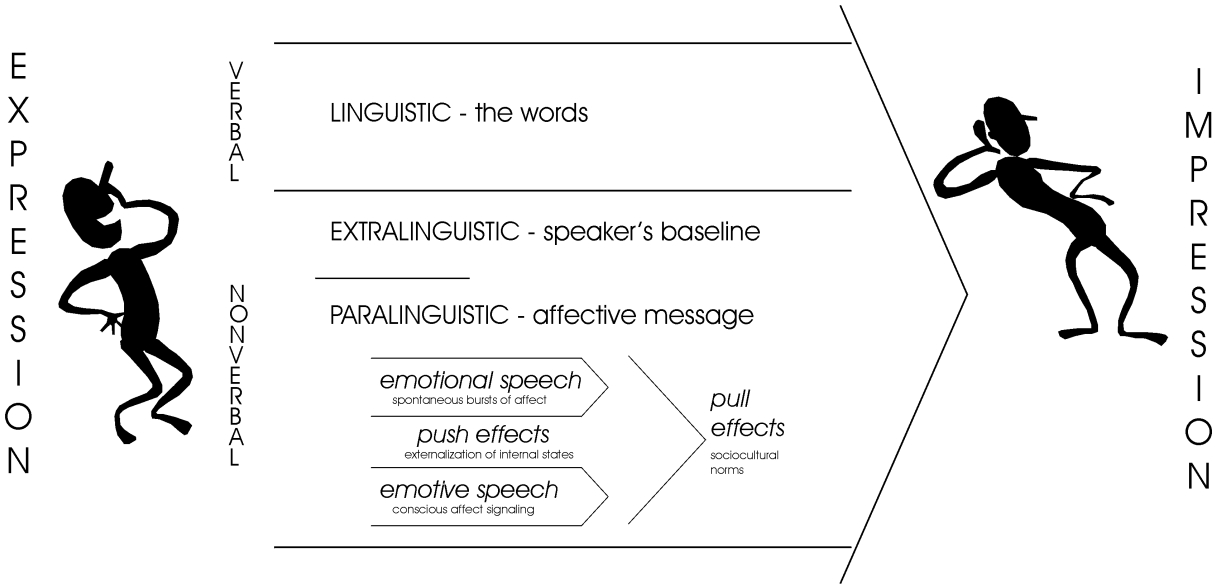


Figure 2.1 Psycholinguistic speech communication: The sender articulates an expression that generates an impression in the receiver

Different techniques have been used to hide the linguistic information from the listener in experiments to isolate the nonverbal content. Margaret Friend and Jeffrey Farrar (1994) compared content masking with *low-pass filtering*, *random splicing*, and *reiterant speech*. In low-pass filtering, all spectral content above 400 Hz is suppressed. Random splicing, as used by Scherer (1971) describes a method where an utterance is divided into 200–400 ms segments that are then recombined in random order. Reiterant stimuli are produced by replacing the syllables of an utterance with nonsense syllables which generate a similar f_0 contour. Each procedure preserves some forms of acoustic information while disrupting or degrading others. Obviously, random splicing and low-pass filtering cannot be expected to yield speech that sounds natural. As Friend and Farrar predicted, all three content-masking procedures generated bias in observers' affective ratings. Scherer (1984) also used speech recordings played backwards to hide linguistic information, but since a lot of information is contained in the pitch and loudness contours, especially the last voiced sound, important prosodic information is lost, if not improperly presented. For the 1996 Banse and Scherer paper, actors were asked to produce meaningless, fantasy utterances – "Hat sundig pron you

venzy." and "Fee gott laish jonkill gosterr." – with phonemes from several Indo-European languages. This, like reiterant speech, preserves both spectral and temporal information of the prosody, but since the speakers have to produce fictive sentences, they might have a lower sending (affect encoding) accuracy than when speaking in their first language. Lea Leinonen and Tapio Hiltunen used an actual word from their language ("saara") and achieved recognition rates from 50-99% in a 10-category forced-choice experiment (1997). Since the same word is used at all times, the linguistic information is moved into the background, and the listeners can concentrate on affective content. They chose to use a name because, as the authors point out, speakers often express their emotional state when uttering the listener's name.

In this work, the recordings remain untouched, and all prosodic information remains intact. The verbal content is hidden by using German sentences, recorded from native speakers of German, for Californian listeners. Also, the same sentence is used for all listeners' evaluations. To allow investigation whether it makes a difference if the recordings are presented in a foreign or one's native language, the evaluation experiments are repeated with German listeners.

2.2.2 Expression vs. Impression

As in linguistic communication, nonverbal vocal information is also transmitted from a *sender* as *expression* to a *receiver* who obtains an *impression* (Scherer 1978, see also Scherer 1982), which implies that the message sent at one end is not necessarily the one understood at the other end. Take for instance the Swedish speaker that generates the impression of a happy extrovert person in a non-Scandinavian listener because of her fundamental frequency's strong modulation, which is a normal (extralinguistic) expression characteristic of her language. The prosodic elements of Russian sometimes make its native speakers sound offended and angry to listeners from other cultures – possibly enough to scare a second grader out of Russiatown, see Figure 2.2. In this case, a listener who is not familiar with the foreign prosody cannot distinguish between the extralinguistic (personal and language-specific) baseline – and the paralinguistic (affective) content.[1]

Figure 2.2 Para- and Extralinguistic message mixed up: Lisa Simpson who is asking for the way to the *Springsonian* Museum, panics as a Russian, relaxedly playing chess outside a hot dog parlor, answers. Although he addresses her in his normal register, she misinterprets the vivid intonation as a sign of anger (Groening 1998).

[1] In general, however, nonverbal vocal communication of emotions seems to be effective across language boundaries. Robert W. Frick summarizes eight studies that either suggested cross-cultural recognition to be as good as within-cultural recognition (5 studies), or slightly impaired, but still operative (Frick 1985).

It is interesting to note that tonal or tone languages, i.e. languages that use intonation to carry linguistic information, such as Thai, Taiwanese or Mandarin, have reduced ability to signal affect prosodically, and therefore are often forced to use *segmental markers* for affective signaling (Ross, Edmondson, Seibert 1986).

The origin of the expression is yet another level away from the addressee. *Push effects*, described as externalizations of internal states (Scherer 1988), are portrayed through culture-specific standards (*pull effects*). The utterance may have an *emotional* – e.g. displaying true inner emotion – or an *emotive* cause, consciously bringing affective information across (Marty 1908). Socially less proficient people can have a low *sending accuracy* (Picard 1997), meaning they are not efficient in communicating their emotions.

When attempting to automatically recognize patterns in nonverbal speech, it thus seems advantageous to stay as close as possible to perception. That is, model the impression a listener has, rather than a speaker's expression such as emotions, or internal state. The impression can be easily quantified in evaluation experiments, and one only has to worry about *receiving accuracy* (Picard 1997), i.e., how well a listener can decode the spoken message and understand emotional content.

2.3 Evaluation of the Recordings

2.3.1 Choice of Categories

In the vast majority of studies on affective vocalization, forced-choice experiments were used to classify speech recordings. Usually, actors are asked to portray given emotions, a number of strong examples is pre-selected, and then listeners match recordings and emotional labels. The recognition accuracy varies strongly among different nonverbal expressions. Pittam and Scherer (1993) compare a study of their own with findings from van Bezooyen and note a decoding accuracy of 28% to 72% (Scherer et al. 1991) and 49–74% (van Bezooyen 1994) for the 5 emotions fear, disgust, joy, sadness, anger (after correcting for chance hits). Banse and Scherer (1996) report an even wider range of 15–78% recognition accuracy for 14 emotions. Frick lists results as high as 90% in his summary on the role of prosodic features in communicating emotions (1985).

For this work, the recordings are evaluated by listeners according to a *semantic differential* approach (Osgood, Snider 1969). The idea behind the semantic differential scheme is to rate data in categories belonging to four groups:

Evaluation (valence)	– description of personal appeal
Activity	– description of the item/process
Understandability	– a meta-category group, e.g. to describe a sample's naturalness
Potency	– an intensity category group

The categories then span a vector space that contains the speech samples. In this work, for example, the psycholinguistic dimensions are pleasant, happy, confident, strong, agitated, leadership, angry; and each recording is described by each listener through an evaluation vector such as 2,1,1,2,-1,1,0. This offers the advantage of a finer-grained scoring procedure than the binary forced-choice dimensions. Clearly, speech is not necessarily happy or unhappy, but can be assigned different degrees of happiness. Moreover, with the semantic differential, the recordings are assessed in every category, datapoints do not have to fall in one bin (be either pleasant or happy or confident, but not all of these), i.e. lie on the category axes of the semantic differential as is the case in the forced-choice evaluation. The data points can populate the whole space, i.e. be represented by linear combination of the category basis vectors. Only in rare cases will pure emotions occur in natural speech.[1]

[1] This, of course, depends on the definition of "pure" or "basic" emotions. Klaus Scherer's *component process theory* (Scherer 1984) conceptualizes that "there are as many different emotions as there are differential outcomes of emotion-antecedent situation appraisal" (Banse, Scherer 1996).

Most of the time multiple affective qualities contribute to the vocal content of an utterance.[1]

Since this work was originally motivated by the ability of public speakers such as politicians who are often able to generate a competent impression even if a conversation's subject matter is unfamiliar terrain for them, the nonverbal categories in this evaluation form a semantic space around the impression of leadership ability. Namely, the listeners rate the recordings in the dimensions unpleasant – pleasant, unhappy – happy, unconfident – confident, (physically) weak – strong, calm – agitated, leadership ability, not angry – angry. The translated German labels are unangenehm – angenehm, nicht glücklich – glücklich, nicht selbstbewußt, selbstbewußt, (physisch) schwach – stark, ruhig – erregt, keine Führungsqualität – Führungsqualität, nicht ärgerlich – ärgerlich. A group of native German speakers initially also evaluated the naturalness of the recordings to make sure no samples that for instance sounded as if the speakers were uncomfortable with the recording situation or that contained unnaturally strong actor pathos would enter the training process of the pattern recognition system.

Tradeoffs in the selection of the categories were to collect as much useful data as possible to provide for the means of analyses possibly beyond the scope of this thesis work and pose different scenarios for the pattern recognition scheme, and on the other hand to avoid weariness and ennui of the human subjects who had to rate 150 recordings.

For the most part, the dimensions are bipolar, i.e. opposite ends of the rating scale represent opposite impressions. It seems to be a matter of definition if one should consider opposite affective states to exist for the (basic) emotions happy and angry, or whether they are unipolar. In the latter case, a speaker who is talking at his extralinguistic baseline would be considered to be both not happy and not angry. However, some listeners might perceive a furious, angry speaker as less happy than just baseline zero. From this perspective, a number of contrasting states now form the negative "not happy" scale, for instance basic emotions with negative valence such as angry, sad, disgusted, or other complex attributes like annoyed or bored. These aspects are also dealt with in the evaluation process and the data postprocessing.

The categories are expected to show variety in how subjective the listener's judgement is, the most personal dimension being 'pleasant'. Here the database for instance allows an abundance of analyses to see if and how appeal is a matter of the speaker's and listener's age, sex, and personal taste, what acoustic parameters are perceived as pleasant, if a happy voice is perceived as more pleasant than an angry voice etc.

With this selection of seven classes, the listeners get to judge both low-level nonverbal impressions such as agitation, and high-level ones like leadership ability, and it can be investigated if the complexity of the percept is reflected in the pattern recognition process.

Even without considering acoustic cues, a close look at the interdependence between the psycholinguistic categories can reveal if some categories might be correlated or show causal dependencies between one or multiple other impressions. For instance, it seems plausible that angry recordings are often also judged as unhappy, physical strength might sometimes correlate with confidence. Perhaps a high-level category such as 'leadership ability' can be fully represented and learned by a pattern recognizer as a function of the other impressions, without knowledge of signal processing parameters.

[1] Rosalind Picard (1997) gives an overview of different models for the representation of emotions. Different theories subsist as to whether pure emotions occur in sequence after each other and thus, averaged over time, form more complex emotional states, or whether a number of basic emotions can occur at the same time and the inner state is a mixture of multiple coinciding emotions. Although most theorists support the mixture model, experimental evidence suggests that for each representation scenarios exist in which it seems valid. Many sets with varying numbers of 'basic emotions' were defined; a prominent one is described by Paul Ekman who lists the 6 emotions fear, anger, sadness, happiness, disgust, and surprise (see Ekman 1992).

2.3.2 The Evaluation Process

20 native English speakers (10 men, 10 women) each evaluated the complete 150-speech-recording database in the categories shown in Fig. 2.3. The evaluators' ages ranged from 18 through 54 years, the average age was 27.
The scoring process and the goal of the research were explained to them, and they signed the UCSD's Human-Subject-Committee consent forms.

Before the actual evaluation started, the subjects did a test run of 10–20 recordings to familiarize themselves with the process and also get a feel for the intensity range of the impressions. The evaluators needed 1–2 hours (there was no time constraint, the speed was determined by the user) to complete all ratings. Short breaks were taken every 30 minutes – or earlier if desired.

Each person received $15 for a completed evaluation. As an incentive to perform well and thoughtfully, the user whose score showed the minimum summed absolute deviation from the mode of all subjects' answers in each category and recording was rewarded with an additional $20.

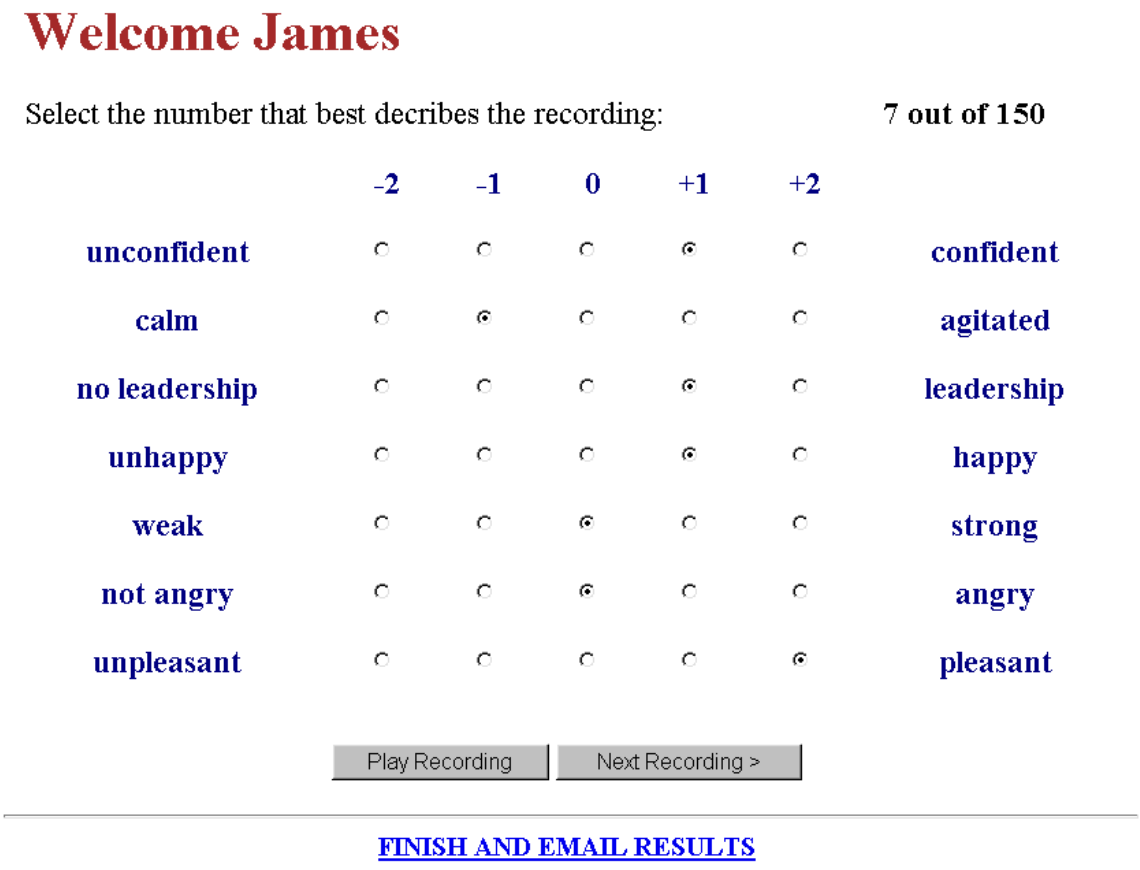


Figure 2.3 The evaluation screen

The bipolar scales contain 5 possible choices from –2 to +2, for the negative and positive extremes, respectively, and neutral at 0.[1] Each psycholinguistic category was explained in detail. For the 'pleasant' dimension, the listeners were asked to record their personal liking

[1] Initially, finer grids of 7 and 11 were tried to lower discretization noise, but the challenge to quantify the affect more precisely required an exorbitantly higher amount of time and concentration which resulted in worse accuracy and consistency than with the coarser 5-point scale.

for the current voice. It was clarified that the voice didn't need to be young, beautiful, or happy. The 'happy' category was explained as a bipolar category, ranging from –2 through +2, so was 'confident.' 'weak – strong' here describes the impression the listener has of the speaker's physique. To avoid confusion with confidence/mental strength, two scenarios were outlined to stress that these impressions do not necessarily coincide. In the first situation, a physically strong athlete who has just won a major championship is nervously giving his first public speech in front of a full stadium (and therefore scores low on 'confidence.'). The second imagery has an old, physically weak sage calmly but confidently teach his disciples.

The notion whether 'strong' also means 'big' is left to the interpretation of the listeners. For the 'calm – agitated' rating it was pointed out that no assumptions about the valence of the sample are made in this category (one of the early test evaluators confused agitated with hectic). Examples of both positive and negative valence were explained to the evaluators for 'calm' (cozy, comfortable vs. depressed) and agitated (happily excited – hectic). 'Leadership' describes how well the listeners believe a speaker would do in a position where that person is in charge of directing and leading people, for instance as chairman of the board of a company or a high-ranking politician. The only unipolar scale, 'angry,' is explained to reach from 0 for not-angry voices to +2 for extremely furious ones. For all categories, the listeners were asked to judge the momentary state of the voice, not what they believed to be the character of the person. The order in which the categories are presented to the listeners by the user interface shown in Figure 2.3 is randomized between evaluators to eliminate a possible bias.

The 145 recordings of the monologue's second sentence, see Sect. 2.1 make up the biggest part of the 150 speech samples. When reading or interpreting the monologue, both actors and non-actors in a few instances sounded slightly unnatural at the first moments of the recording. By the beginning of the second sentence, they were in a normal flow. Sentence two also has the most complex sentence structure and the largest number of syllables, and therefore allows to collect the greatest number of acoustic parameters with the *lombada* feature extraction, see Section 4.4.2. Two instances of sentence 2 (one male, one female speaker) were repeated to check if an evaluator's rating is consistent throughout the whole process. This also permits to analyze whether the perception of a voice is influenced by previous samples, if for instance a normal recording after a series of very strong ones is judged slightly weaker than normal. The remaining three database entries are two examples of the question (sentence 7) and sentence 6 of the monologue. This way, it can be tested if the score that sentence two of one recording receives is the same as for another part of the monologue. If so, the labels can be adopted by all the sentences of the recording. Also, it can be seen if the pattern recognition system that is trained on one sentence spoken by 145 people is able to correctly assess the nonverbal content of another, previously unseen, utterance. The order in which the 150 speech samples are presented to the listeners is randomized and changed for every user.

The evaluators used headphones to listen to the recordings. All 20 sessions took place at the same terminal at UCSD's Machine Perception Lab. The evaluators interacted with the computer through the user interface shown in Figure 2.3. They listened to a speech sample first, then rated it by clicking on the appropriate radio buttons in all categories. The sample could be replayed and the selected buttons changed as often as desired until the user advanced to the next recording. Once the evaluation of all samples was completed or when the user decided to finish the process, the evaluator was asked to fill out a (voluntary) short form which asked for the person's age, training in German and other languages, comments, and an email address as contact information if the user wanted to be eligible for the extra best-evaluation bonus. The scores and user information were then sent per email to the computer on which further data processing was done.

The evaluation program is written in Java script and runs on any contemporary web browser. It can either be run directly on the web or downloaded and then run offline. This offers the great advantage that the same database can be evaluated by anybody who has a

computer with sound capabilities, and allows to compare the results of listeners from all over the world; either of people who do the evaluation at home, or of subjects who work in cooperating labs. Converting the audio files to MP3 format preserved the high audio quality at a fraction of the file size through perceptual and entropy coding. The download method proved to work well in collecting scores from 20 German listeners (also 10 men, 10 women; ages 22–58, average age 29) who performed the same evaluation process at their own computers. The results of the German evaluations are included below.

The Java script evaluation software with the speech database is currently available at http://mplab.ucsd.edu/~holcus/Speech.html.

2.3.3 Postprocessing of the Evaluation Scores

To account for different evaluation behaviors, the scores of each person are normalized with regard to mean and average absolute deviation in every category.

It has been shown that moods affect evaluative judgment (Niedenthal, Kitayama 1994). (*Moods* – as opposed to emotions – describe longer-term affective states that persist from hours to days, possibly longer. The phenomena denoted here as *emotions* last a few minutes.) Moods bias the interpretation of the likelihood of events, i.e. a person with a positive predisposition will consider positive outcomes as more probable than negative ones, and vice versa (Mayer, Salovey 1993). They also influence judgment in the sense that a person's mood is reflected in the quality of the evaluation (Clore 1992). A happy person, for instance, will in general perceive her environment in a more positive way than someone who is angry. To eliminate a possible mood bias, each evaluator's scores are shifted so that the average responses in every category are zero.

Another elementary normalization issue that needs to be addressed are the differing scoring *amplitudes* among the evaluators. Depending on character, mood, or experience, some people will tend to deviate further from the zero evaluation baseline than others. Three variants were compared: leaving the scores untouched, setting the standard deviation to 0.5 in each category, and adjusting the average absolute deviation to 0.5.

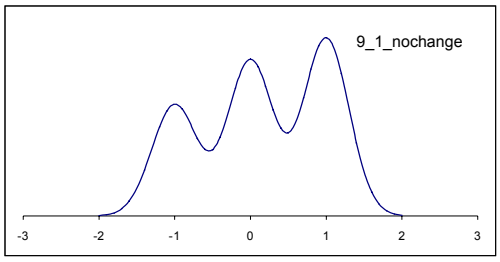


Figure 2.4a One of the 1050 evaluation histograms of the original 20 responses

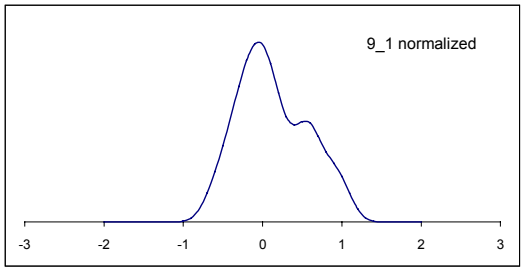


Figure 2.4b The same data with normalized mean and absolute deviation

2.4 Evaluation Data

As seems plausible considering the nature of the problem, normalizing the average absolute deviation maximized the sum of datapoints that fall in a 2-standard-error interval for each category and recording, and also visibly smoothed the histograms of the 20 answers for each category and recording. An example of the normalization effects is displayed in Fig. 2.4. The left histogram shows the three distinct modes where the evaluators cast their scores, the adapted graph on the right exhibits a more uniform shape with only one major peak.

The absolute deviations (before normalization) averaged over all users, recordings and categories are a good indicator of how much (varying) affective information is contained in the individual categories. They rank as follows:

Psycholinguistic category	Average absolute deviation
Agitated	1.04 ±0.05
Confident	0.99 ±0.04
Leadership	0.93 ±0.04
Strong	0.86 ±0.05
Angry	0.85 ±0.08
Pleasant	0.74 ±0.06
Happy	0.66 ±0.04

Table 2.1 Average absolute deviation as an indicator of affective information in the psycholinguistic categories

Indeed, one of the evaluators remarked that she felt the people showed no happiness or anger at all. The scenarios that were given to the actors during the recording of the speech database were not intended to show much variance in anger and happiness; these categories were added later because of interest in Paul Ekman's 6 basic emotions. It is surprising to see, however, that the absolute deviations in the 'pleasant' category are also very low, considering that the database has a wide repertoire of different voices, especially among the actors' recordings.

2.4.1 Choosing A Representation

At this point, 20 evaluation scores exist for each recording in each category. After the normalization of mean and absolute deviation in the individual evaluations, a number of methods were tried to derive one value from these 20 scores that represents the affective intensity in a recording and category, e.g. describes 'how happy does recording 12 sound?'.

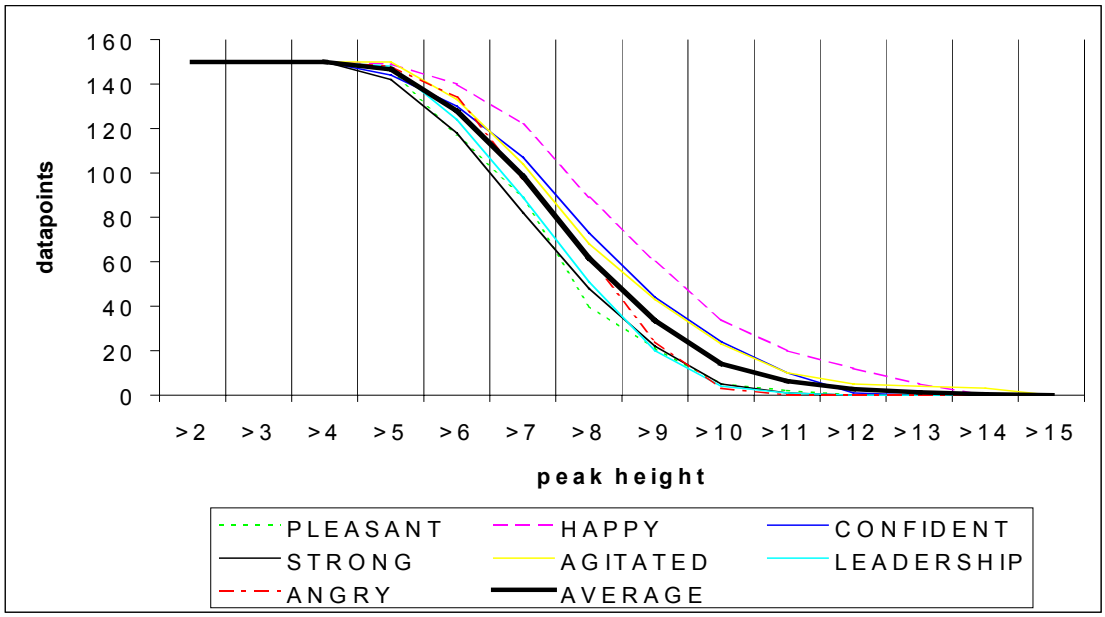


Figure 2.5 Histogram mode height as a measure of agreement. The graph displays how many out of the 150 global maxima in each category are taller than the number indicated on the abscissa. Example: in the CONFIDENT category, 44 out of all 150 histograms had a maximum mode greater than 9.

Of course, the pattern recognition scheme could be trained to learn the probability distributions or histograms of the evaluations, but representing one recording with one value in each category appears less complex. Taking the average of all 20 listeners' responses is fairly sensitive to outliers. A more robust representative is the median, i.e. the center number in the ordered sequence of the evaluations. Alternatively, assuming there is a limited number of outliers with the inner datapoints accurately distributed among the true mean, all points outside the median average deviation (MAD) from the median of the 20 responses (or a percentile of the MAD) can be ignored, and the average of the remaining points is computed. The averages in MADs seem a better representation than the medians or averages, but the question still arises whether it is useful to average in a multimodal case – if a car navigation system was to decide whether to proceed left or right at a forking road, going straight clearly is not in the interest of the passenger. Instead, it seems useful to pick the most probable variant which is what was opted for here after all methods were tried: the 20 votes in each category and recording are represented by the position of the mode, i.e. the histogram's global maximum.

Additionally, the *height* of the maximum is listed as a confidence measure, to show how strongly the listeners agreed in their judgments: if all listeners placed their answers on the same spot, agreement would be high and so would the peak of the histogram. If the scores were all over the scale, the global maximum would be smaller. In this case, the histograms were drawn by convolution of the 20 evaluation scores (as delta-distributions) with a gaussian bell curve of height 1, hence the maximum value a mode could assume is 20. Figure 2.5 gives an overview of the mode heights in the data set. These confidence measures were used during the training of the pattern recognizer as instruction how much the neural network should be influenced by a particular example (see Chapter 4).

2.4.2 American and German Evaluator Responses Compared

Since the scores were obtained in slightly different conditions among the listeners in California than among the German ones, the data is compared only qualitatively here. Some trends, however, are noticeable.

If there were any culture/language specific differences in the communication of affect, these could manifest themselves in different perception of baselines for different impressions. For instance, the sound of German fricatives such as [x] in the German *Dach*[1] are usually perceived as harsh by the American listener, and therefore might always bear a more angry note than speech in his own language.

This evaluation data does not support such generalization. Au contraire, the results are strikingly similar for both German and Californian subjects.

Psycholinguistic category	Average responses American listeners	Average responses German listeners
Pleasant	-0.04 ±0.08	-0.04 ±0.10
Happy	-0.38 ±0.04	-0.26 ±0.06
Confident	0.29 ±0.05	0.20 ±0.06
Strong	0.14 ±0.06	0.15 ±0.05
Agitated	-0.05 ±0.07	-0.01 ±0.06
Leadership	0.05 ±0.08	0.00 ±0.06
Angry	-0.42 ±0.08	-0.31 ±0.13

Table 2.2 "Means of means" of the (original) responses for the American and the German evaluators

[1] or 'r's after consonants – which have a totally different quality in American English pronunciation, namely are spoken as *glides*; voiced and with much more smoothly fricative aspiration

Table 2.2 shows the average responses from all evaluators in each category, computed by averaging the (not normalized) averages of the individual judges in each category (the standard error hence describes how much the mean varied among the evaluators, not among all scores).

The data shows no general bias of the Californian listeners' perception towards angry or happy etc. For each impression, the value's sign is the same for both groups.

The deviations from neutral are consistently larger for the American judges – with the exception of 'pleasant' and 'strong,' where both groups scored almost the same way. Possible reasons for this could be that since the linguistic information was hidden to them, they were more sensitive to the affective content. Another explanation is simply that the American judges tended toward larger scores, which finds further substantiation by the juxtaposition of the different average absolute deviations as listed in Table 2.3. As for Table 2.2, the averages (before normalization) were first computed for each individual subject, and then averaged over all listeners.

Psycholinguistic category	Average absolute deviations American listeners	Average absolute deviations German listeners
Pleasant	0.74 ±0.06	0.69 ±0.05
Happy	0.66 ±0.04	0.42 ±0.04
Confident	0.99 ±0.04	0.78 ±0.06
Strong	0.86 ±0.05	0.70 ±0.06
Agitated	1.04 ±0.05	0.69 ±0.06
Leadership	0.93 ±0.04	0.77 ±0.05
Angry	0.85 ±0.08	0.71 ±0.06

Table 2.3 Average absolute deviations of the (original) responses for the American and the German evaluators

A look at the distribution of the 1050 evaluation histograms of the native English speakers and a comparison with their German counterparts shows a tendency of the German responses towards higher modes, i.e. greater overlap/agreement of the judgments, see Figs. 2.6a-h. The graphs show how many out of the 150 global maxima in each category are taller than the number indicated on the abscissa. The blue graphs symbolize the American answers, the black lines the German ones. The higher peaks could result from a better receiving accuracy among the German listeners when hearing their native language.

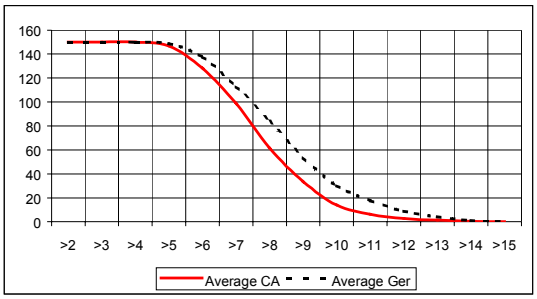


Figure 2.6a Average number of modes that reach the indicated height

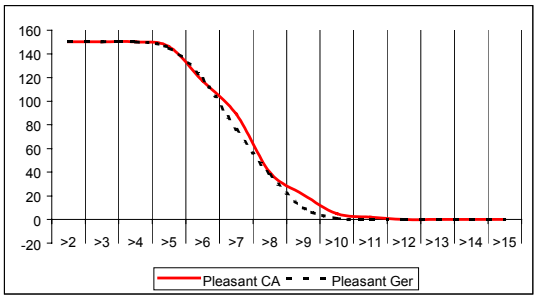


Figure 2.6b Number and height of peaks in the pleasant category

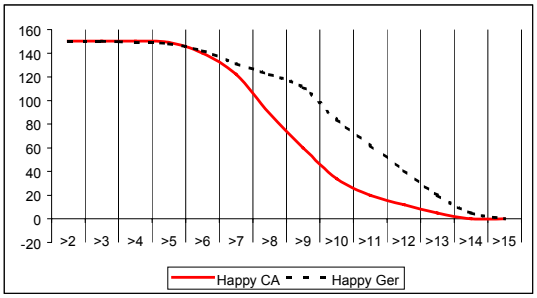


Figure 2.6c Peaks – happy

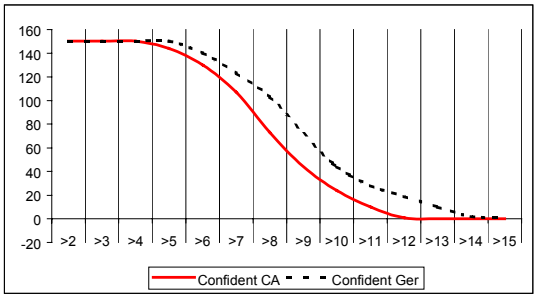


Figure 2.6d Peaks – confident

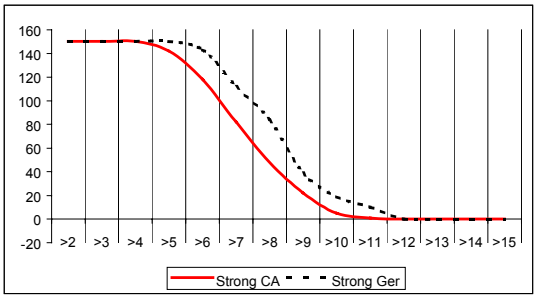


Figure 2.6e Peaks – strong

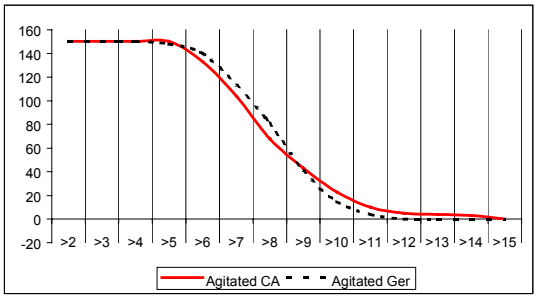


Figure 2.6f Peaks – agitated

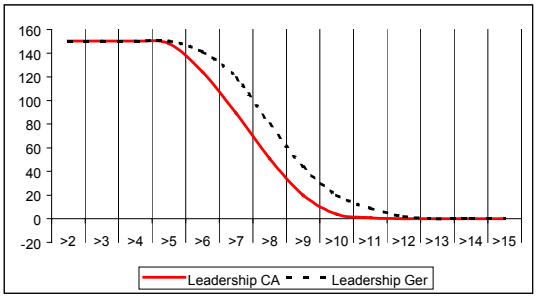


Figure 2.6g Peaks – leadership

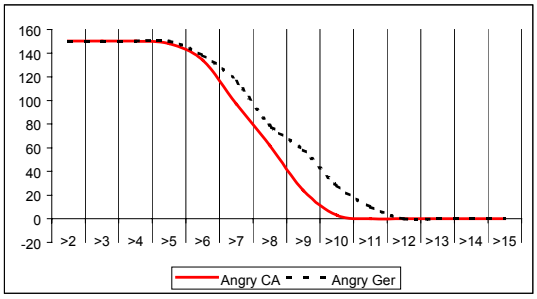


Figure 2.6h Peaks – angry

3 Signal Processing

Whereas the previous chapter describes the *affective* content of vocal communication, this part looks at speech on a lower level and outlines how the *acoustic* parameters that shape the prosody are generated, perceived, and how they can be analyzed. The parameters fall into the categories fundamental frequency, intensity, and spectral composition. To extract these, a pitch tracker is developed, a new psychoacoustically motivated speech loudness model is introduced, and other signal processing tools are build that lead to a total of 18 acoustic parameters representing each recording.

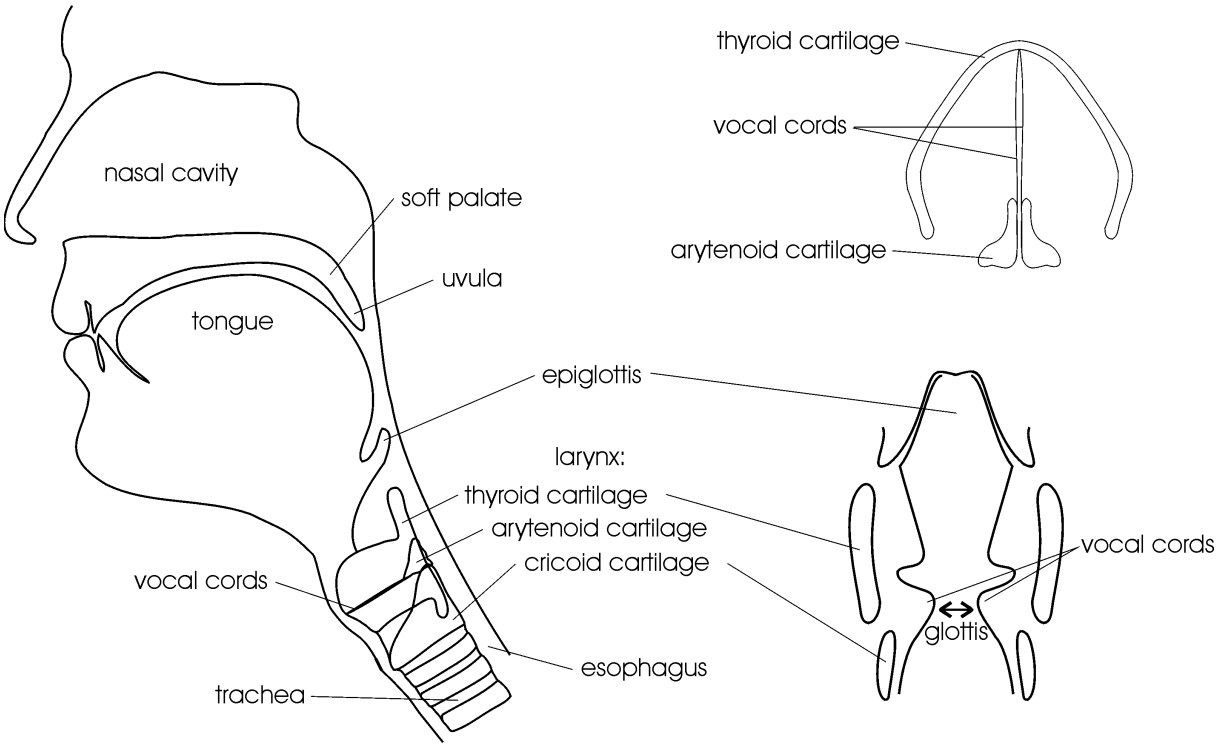


Figure 3.1 The vocal tract: lateral, coronal, and superior view

3.1 The Speech Signal

The sounds of vocal communication are produced by obstructing the airflow from the lungs to the mouth and nose in various places of the vocal tract, see Figure 3.1. In voiced speech, two outside forces interact to create an oscillation of the vocal cords: one is generated by the air pressure that builds up in the larynx underneath the vocal cords when these adjoin. The antagonist force is given by the Bernoulli effect that creates a low pressure at the place of the smallest diameter, the glottis, when the air flow is high. The resting position of the vocal cords during vocalization is determined by the register and therefore by the position of the involved cartilage, muscles, and tendons, especially by the arytenoid cartilage (cf. Klatt and Klatt, 1990). During regular vocalization, the glottis is half-closed at the beginning of a period. The air pressure from below widens the opening until the Bernoulli force becomes strong enough to shut the glottis (van den Berg 1957). At this time, air pressure builds up below the larynx, and the vocal cords open up again. The oscillation of the vocal folds is

supported by the tension of its ligaments and muscles that forces the vocal folds back into the resting position.

As shown in the glottogram in Figure 3.2, the glottal flow is not sinusoidal. The part of the glottal pulse that corresponds to the glottis-closing part of the cycle has a steeper slope than the opening portion which results from the vocal folds' abrupt shutting. Representing the corresponding excitation signal in the frequency domain therefore, since it is not sinusoidal, requires a Fourier series with a fundamental frequency f_0 and harmonics at its integer multiples. These overtones appear as parallel "ripples" in the spectrogram (Fig. 3.3). Note that the harmonics are strongest in the frequency band below 4000 Hz, but the rippled structure remains detectable all the way through frequencies as high as 15 kHz which is a sign of a well-trained vocal apparatus and a glottis that shuts completely during the glottal cycle (the sample shows the voice of a professional actress with regular stage- and radio appearances).

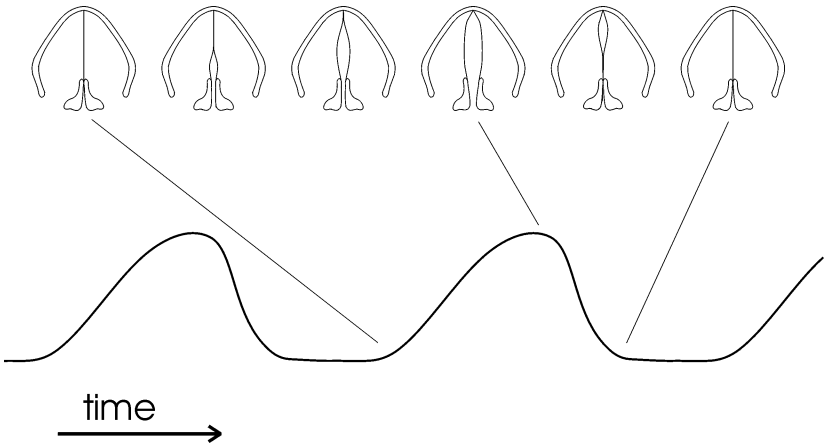


Figure 3.2 The glottal period: oscillating vocal folds (upper row) and the resulting air flow glottogram (bottom row, air volume vs. time)

Besides the voiced, quasiperiodic excitation of the glottal pulse, speech also contains a noise-like quality that stems from the turbulent, non-harmonic airflow through the vocal tract. This unvoiced excitation is employed to produce fricatives like [f] in foot or [ʃ] as in shoe, and plosives where air is released suddenly such as with the aspirate [p] as in put or the affricate [ʦ] in bats. These sounds are visible as vertical bands in high frequency areas between 4-15 kHz in the spectrogram 3.3.

Some speech sounds are both voiced and have a noise excitation element, e.g. [ʒ] as in pleasure, or the voiced 's' [z] in the German word 'singen' at 1.5 seconds in the spectrogram below.

Purely voiced speech sounds are rare; usually, some residual air flow occurs through an opening between the arytenoid cartilage in the larynx. An exception is the so-called *vocal fry*, laryngealized (Klatt and Klatt 1990) or *creak* voice[1]. In this case, the arytenoid cartilage are pressed firmly together, and air is released only between the vocal folds and only during a relatively small fraction of the glottal period (Eckert, Laver 1994; Stevens 1977). The forceful locking of the vocal cords causes an attenuation of the higher harmonics and is revealed by clear frequency periodicity in a spectrogram.

[1] Vocal fry is the most efficient way of speech production since the complete airflow is used to create vocal folds oscillation. It is often accompanied by a low pitch and occurs frequently during the final instances of an utterance. A prominent example of the creak quality is the voice of former U.S. Secretary of State Henry Kissinger.

As opposed to the voiced speech production of regular and creaky speech, *whispering* exploits the nonperiodic fricative excitation of turbulent airflow. The vocal folds are pressed together and air streams through a triangular aperture formed by the adjoining thin ends of the arytenoid cartilage and the stretched-apart base ends of the arytenoids.

Both for the voiced and the breathy excitation, the utterance receives its phonemically characteristic sound as the air passes through the vocal tract (cf. Schroeder 1999). In this passage from the glottis to the lips the airflow is modified through a highly complex set of muscles, bones, cartilage, and ligaments (see Fant, Lumby 1977), which is orchestrated in an even more complex manner. A simplified view of the vocal tract is given in Figure 3.1. The resonances of the vocal apparatus at the poles of its transmission function are called *formants* and shape the frequency response, i.e. the spectral envelope with which the excitation signals' spectra are multiplied (Fant 1970, Schroeder 1999). Differing formant structures are responsible for our ability to produce distinguishable vowels; the most important ones are the first two formants of speech that usually lie below 2 kHz. As the dimensions of the vocal tract vary from speaker to speaker, so do the acoustic characteristics, and a spectrogram such as the one below thus carry information about the speaker's physique, linguistic message, and affective content.

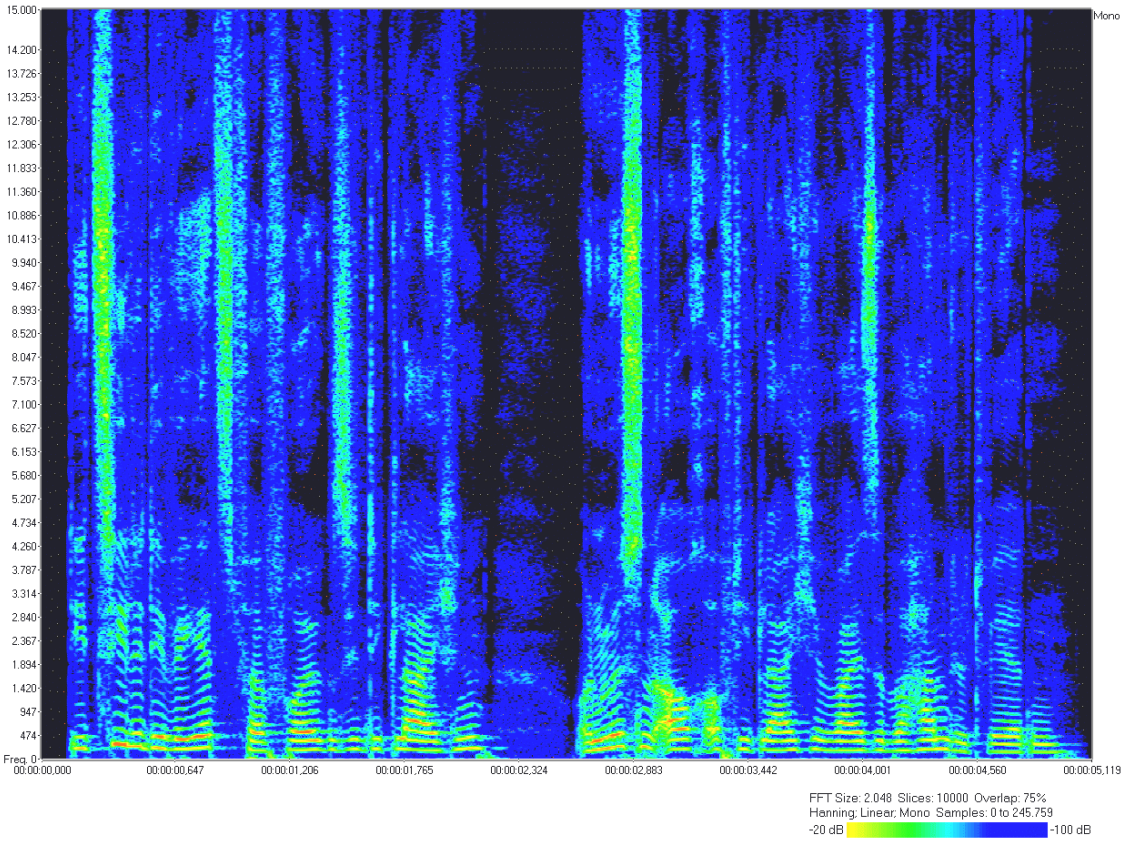


Figure 3.3 Spectrogram of a soft female voice speaking sentence 2 of the database's monologue. Plotted are frequency in Hertz vs. time in milliseconds.

3.2 Acoustic Parameters

The most frequently used acoustic features in the literature on nonverbal speech are fundamental frequency and intensity, followed by spectral properties.

When dealing with these acoustic parameters one has to distinguish the physical quantity, for instance signal power, from the perceived sensation, in this case loudness. For pairs of physical quantity and perceived impression, such as fundamental frequency and pitch, or syllables per time unit and perceived speech rate, the relation is almost linear (above a minimum-change threshold). For other pairs, e.g. spectral structure – timbre, or signal intensity – loudness, the percept depends strongly on the nonlinear processing by our auditory system, see Zwicker, Fastl 1990. The 18 parameters used in the present experiments are mostly related to perception since it is the goal to model impression rather than expression, the stimulus. Besides the aggregate measures that describe values throughout the whole utterance, such as f_0 or loudness median, or average spectrum, some values are picked at prominent points such as the loudest voiced point, or the last voiced point. The speaker's sex is included to see if the network is able to capture the complex interaction of parameters[1] that our auditory system uses to distinguish between male and female voices.

```
*********************************************************
           Statistics for file 44_2
*********************************************************

* File info *
 Samplingrate:              48000 Hz

* F0 - pitch *
 F0Median:                  195.135 Hz
 F0MedianAbsoluteChange:    8.15447 halftones/s

* Spectrum, loudest band set to 50dB *
     0-500Hz:               50 dB
  500-1000Hz:               40.2713 dB
 1000-2000Hz:               28.7416 dB
 2000-4000Hz:               22.1666 dB
 4000-8000Hz:               17.0848 dB

* Loudness *
 LoudMaximum:               9.21783 Sone
 LoudMedian:                6.185 Sone
 LoudMedianAbsoluteChange: 17.3935 Sone/s

* Correlation Loudness [sone] , F0 [halftones] *
 r=                         0.275377

* Prominent points: loudest voiced and last voiced *
 LoudestF0:                 192 Hz
 LoudestLoudness:           9.21783 Sone
 LastF0:                    149.068 Hz
 LastLoudness:              3.07156 Sone
 DifferenceLoudestLastF0:   4.38163 halftones

* Speechrate *
 Speechrate:                7.11461 peaks/s

* Sex *
 Sex:                       f

* Number of Data Points *
 #ofLoudnessPeaks:          30
 #ofVoicedLoudnessPeaks:    26
```

Table 3.1 The acoustic parameters used for pattern recognition in the following experiments

[1] see Mullenix et al. 1995

The loudness and pitch features were extracted at maxima in the loudness contour as explained below in the section on the *lombada* technique (p.46). These loudness maxima are also used to derive the speech rate value as peaks per second. The parameters are collected from one full sentence since it represents a prosodically complete unit.

3.3 Fundamental Frequency/Pitch

The sound quality we perceive as *pitch* is related to the physical stimulus *fundamental frequency* f_0, the inverse of the fundamental period T_0. Although complex tones are in some instances perceived with a *virtual* pitch different to the actual physical fundamental frequency (Terhardt 1979), the impression of pitch for speech corresponds directly to the inverse of the actual fundamental period (Hess 1983).

Various signal properties of the signal can be used to measure f_0. Most pitch detectors operate by finding periodicities either in the time domain (fundamental period trackers) or the frequency domain (fundamental frequency trackers). In the time domain, this process can be as easy as observing the number of zero crossings on the intensity axis (zero crossings analysis basic extractor, ZXABE) or another threshold (threshold analysis basic extractor). Other algorithms consider more complex features of the wave form, or the waveform as a whole, as is done in the *autocorrelation* method explained below.

One problem that pitch trackers need to deal with stems from the fact that an oscillation with harmonics also has peaks at half of the fundamental period T_0, one-third of the fundamental period etc. Schroeder (1968) used this property to detect the fundamental period at the smallest common multiples of the period durations of the individual harmonics in a period histogram. The same method also works as *spectral compression* with a frequency histogram, where the fundamental frequency f_0 is the greatest common divisor of the individual harmonics.

As the previous example does, pitch trackers operating in the frequency domain make use of the overtones that appear in the spectrum as harmonic ripples at integer multiples of the fundamental frequency, see Fig. 3.3. The *cepstrum* tracker built for this work is such an f_0 detector.

Some algorithms explicitly work both in the time and frequency domain, such as in *modulation spectrum analysis* as described by Olaf Schreiner (2000), where fundamental frequency information is used to separate a signal's voiced part (speech) from inharmonic noise.[1]

In this work, two trackers are programmed and the outputs combined to find a reliable measure of pitch. Time windows of 43 ms (2048 samples at a sampling rate of 48 kHz) are extracted from the speech recording, analyzed with an autocorrelation and a cepstrum technique, each yielding 4 candidates for a possible f_0 value. The best one is chosen with respect to system knowledge accumulated from previous values. The time window is then shifted by 17 ms (800 samples), and the process is repeated until the end of the file is reached.

3.3.1 Autocorrelation with Centerclipping

The autocorrelation is used here both to find periodicity in the time domain as well as for voiced/unvoiced or silence detection (similar to the process suggested by Rabiner and Schafer, 1978). The scalar product of the extracted window *w(n)* of length *N* with a time-

[1] For this analysis, a spectrogram with a high temporal resolution is divided into its frequency bands. Another transform in each band reveals the periodicity of each band's temporal structure. All frequencies that show the same modulation are grouped together and considered to belong to the same voiced signal.

shifted copy *w(n+t)* of the same vector is built. The discrete autocorrelation sequence ac $w(t)$ for a signal of finite duration is then expressed as a function of the time shift t (the lag):

$$\text{ac } w(t) = \sum_{n=0}^{N-1} w(n) \cdot w(n+t) \qquad 3.1$$

For periodic signals (or quasiperiodic signals like the voiced speech signal), the autocorrelation peaks at lags that are integer multiples of the fundamental period.

Other, unwanted peaks originate from the harmonics of f_0, especially from the ones attenuated by major formants. This strongly limits the usefulness of the pure autocorrelation as a fundamental period tracking device; common erroneous responses lie at $T_0 \pm T_F$, where T_0 is the actual fundamental period and T_F the period corresponding to a major formant (Schroeder 1970). The misleading influence of the formants can be subdued by means of *centerclipping* the signal vector prior to the autocorrelation. In this procedure, all values whose absolute magnitudes are smaller than a given threshold are set to zero, from the ones with greater magnitude the threshold value is subtracted/added (for values greater/smaller than zero, respectively).[1] To compute the clipping level, the largest (absolute) value in the first third of w is compared to the largest maximum in the last third of w; the smaller one of these two is multiplied by 0.5 and taken as the threshold.

The highest possible value for the autocorrelation function appears at $t=0$. In this case, the vector is simply squared. Dividing this value by the window length yields a power value for this time interval.

For unvoiced signals (essentially noise with a frequency bandwidth of over 10 kHz), the autocorrelation function approaches zero quickly since random processes are uncorrelated. Therefore, this information can be used to classify the signal as unvoiced if after a short time its autocorrelation does not reach a maximum greater than an empirically derived threshold of 30 % of the initial value at $t=0$.

Local maxima that exceed the 30% limit are considered candidates for T_0-values. In case of the standalone-autocorrelation pitch tracker, the first maximum that corresponds to a fundamental period longer than 2 ms is picked. This gives the autocorrelation enough time to decay from the initial high start point, and leaves out the high-pitched formant information. (Consequently, the highest fundamental frequency that can be detected by this procedure is 500 Hz.) For the mixed autocorrelation–cepstrum method, a maximum number of 4 peaks that clear the 30% hurdle are selected as candidates. Their heights are normalized to a sum of 1 and taken as probability measures.

Since only $N-t$ sample-pairs get multiplied (t values are shifted "over the edge" of the vector's copy, see Eq. 3.1) the autocorrelation decreases linearly on average as the lag goes from zero to N. Therefore, the peak picker is biased in favor of low fundamental period values. In case of low voices with long fundamental periods, this could lead to erroneously choosing time values corresponding to higher harmonics. To eliminate this effect, one can either correlate the windowed vector with a vector that begins at the same point as the first one but is of $2N$ length so no multiplications with zero occur. Alternatively, one could multiply the ac value at lag t with a factor compensating the decay. The latter procedure is delicate because as the signal gets smaller, the quantization noise increases in relative size with the multiplication. The first-mentioned procedure, shifting the N-length vector over a 2N-original vector, hasn't shown much effect in tests for this work. Best results were

[1] How well this process eliminates the formants, i.e. the linguistic information, was demonstrated by Licklider and Pollack (1948, quoted in Hess 1983 p. 362) who compared the intelligibility of centerclipped and *infinitely clipped* speech (positive values of the waveform are represented by a value of +1, negative values as -1). Whereas the latter was still understandable despite the quantization noise at 1-bit sampling depth, cutting out even a small intensity center band from the signal crucially reduces intelligibility.

achieved by autocorrelating two equal windows and experimentally adjusting the window length to a size short enough so the speech periodicity is still largely unchanged, and long enough so the autocorrelation does not fall flat too soon for a comfortable majority of recordings. A length of 43 milliseconds (2048 samples) proved to work well.

3.3.2 Cepstrum

The cepstrum (coined as an anagram of "spec-trum") essentially analyzes the signal in the frequency domain (see Schroeder 1999, Hess 1983). This method makes use of the periodicity, the "ripple structure" as seen in Fig. 3.3, in the frequency domain.

The voiced part of the speech signal can be understood as a convolution of periodic excitation pulses *p(t)* occurring with fundamental frequency f_0, with the impulse response *h(t)* of the vocal tract (adapted from Schroeder 1999):

$$s(t) = p(t) \star h(t) = \int_0^\infty p(t-\tau) \cdot h(\tau)\, \mathrm{d}\tau \qquad 3.2$$

A Fourier transform allows to examine the signal in the frequency domain. The convolution in (1) now becomes a simple product,

$$S(\omega) = P(\omega) \cdot H(\omega) \qquad 3.3$$

Before transforming, the data vectors in this work are multiplied by a $(1\text{-}\cos x)^2$ window to avoid the strong spectral splatter that would result from transforming a finite signal with harsh edges (for this short-term analysis, the upper integral bound of Eq. 3.2 is de facto not infinity because one deals with a windowed signal). As seen in Eq. (3.3), the speech signal in the frequency domain can be modeled as the product of the excitation's spectrum and the frequency response of the vocal tract. Taking a logarithm turns this product into a sum,

$$\log S(\omega) = \log P(\omega) + \log H(\omega)\ , \qquad 3.4$$

and a further Fourier transformation returns the viewpoint to the time domain,

$$c(q) = p'(q) + h'(q) \qquad 3.5$$

where the new variable is called *quefrency* and has the dimension of time. Through this, spectral envelope information, which appears mostly in the $0-3$ ms range of the cepstrum *c(q)*, and excitation period information can be separated. The fundamental period appears as a peak in the cepstrum.

Geometrically, the cepstrum can be interpreted as finding the periodicity of the harmonic ripples in the spectrum by means of another FFT. The logarithm in Equation 3.4 smoothes the spectrum, and thus reduces the formants' high-frequency content in the second Fourier transform. It is this spectral flattening that suppresses the low-quefrency vocal tract information and enhances the relative size of the fundamental-frequency peak.[1] Taking the

[1] Other means of spectral flattening instead of the logarithm exist. Ideally, one is able to completely reverse the effect of the vocal tract and separate harmonic structure from spectral envelope – that is, excitation and formant information of Eq. 3.2 – through a deconvolution. In the frequency domain, this *inverse filter* is characterized as the numerical inverse of the frequency response $H(\omega)$, since $P(\omega) = S(\omega)/H(\omega)$. The frequency response and thus the poles and zeros of the inverse filter can be derived from linear predictive coding coefficients (Markel and Gray 1976; Schroeder 1999). For applications where LPC seems too computationally expensive, a simplified inverse filter (SIFT) can be used where a high-frequency range of all but the first formants is blocked

power spectrum – i.e. squared frequency values – synchronizes all phases in the time domain, which adds to the sharpness of the cepstral peak.

The highest glottal-to-noise excitation ratio occurs in the low frequencies of speech, therefore, for this work, the signal is lowpass-filtered in the frequency domain with a cutoff frequency of 5 kHz.

After the cepstrum has been computed, the 4 highest peaks in the quefrency range corresponding to 50–250 Hz are selected. If the global maximum manages to rise above the average of the three next highest peaks by a factor of 1.4, it is accepted as T_0, otherwise the window is classified unvoiced.

3.3.3 Relationship between the autocorrelation and the cepstrum

Although the autocorrelation, working with a time dimension, and the cepstrum, detecting periodicity in the spectrum, appear to be fundamentally different, they are intimately related. As was already used in Equations 3.2 and 3.3, the convolution in one domain corresponds to a multiplication in the other – they are *transform pairs*, i.e.

$$f(x) \star g(x) \;\leftrightarrow\; \mathrm{F}(y) \cdot \mathrm{G}(y) \qquad (\textit{convolution theorem}). \tag{3.6}$$

The (continuous) correlation function corr(*f, g*) and convolution are equivalent except for the numeric sign of the lag, "the direction of the vector shift":

$$\mathrm{corr}\big(f(x), g(x)\big) \equiv \int_{-\infty}^{\infty} f(x+\xi) \cdot g(\xi)\, \mathrm{d}\xi \;, \text{ and}$$

$$f(x) \star g(x) \equiv \int_{-\infty}^{\infty} f(x-\xi) \cdot g(\xi)\, \mathrm{d}\xi \;,$$

which, together with Eq.3.6, results in the *correlation theorem*

$$\mathrm{corr}\big(f(x), g(x)\big) \;\leftrightarrow\; \mathrm{F}(y) \cdot \mathrm{G}(-y)$$

or, for real functions *f* and *g*, where the Fourier transform is given by $G(-y) = G^*(y)$:

$$\mathrm{corr}\big(f(x), g(x)\big) \;\leftrightarrow\; \mathrm{F}(y) \cdot \mathrm{G}^*(y) \;.$$

In the case of $f(x) = g(x)$, the correlation turns into the autocorrelation and yields the *Wiener–Khinchin theorem*,

$$\mathrm{ac}\, f \;\leftrightarrow\; |F|^2 \;, \tag{3.7}$$

which states that autocorrelating a signal in one domain corresponds to building the power spectrum in the other domain.

---

out by a fixed low-pass filter, and formant 1 is suppressed by a time variant band-stop filter (Hess 1976). A spectral flattening function that showed virtually the same performance in this work as that of the logarithm is the square root. Although this method cannot offer an obvious numeric explanation like the logarithm in the cepstrum or inverse filtering/deconvolution, it is geometrically plausible and offers the advantage over the logarithm that it is defined for all nonnegative real numbers and has no further free parameters that might need to be adjusted (M.R. Schroeder, personal communication).

This offers two substantial insights. First, it shows that the autocorrelation can be computed by Fourier-transforming the signal, calculating the power spectrum, and retransforming; a process that seems disadvantageous at first glance, but a look at the computational cost confirms the opposite. Whereas the autocorrelation requires N^2 multiplications, the FFT-variant is of order $N \log N$. At a window size of N=2048 samples, switching from a regular autocorrelation, as was initially implemented, to the FFT-version sped up the whole autocorrelation pitch tracking process by about 80% for a 5 second sentence recorded at 48 kHz and a 2048-sample FFT.

Secondly, the Wiener–Khinchin theorem demonstrates the close relationship between cepstrum and autocorrelation. Without the logarithm in the cepstrum, the two procedures would work the same way: transferring the signal to the frequency domain, building the power spectrum, and returning to the original time dimension. Both methods use a way of formant reduction, namely spectral flattening by means of the logarithm in the cepstrum, and centerclipping in the preprocessing of the autocorrelation.

3.3.4 Combining Cepstrum and Autocorrelation Data

Despite the strong similarity of both functions, their behaviors diverge in some situations. As Wolfgang Hess explains: "Ordinary AC analysis fails when there is a predominant formant (especially F1) at some higher harmonic. Cepstrum analysis, in contrast, fails as soon as the presumption is violated that the signal contain many adjacent harmonics. (Hess 1983, p. 405)" Indeed, in some instances one method is more robust than the other one, and combining the outputs of both trackers is beneficial to overall performance. The major increase in stability however is achieved by a postprocessing that selects the f_0 value from a number of candidates from both the autocorrelation and the cepstrum and takes into account the statistics of the previous speech windows and safely rejects outliers, which are mostly found at half or twice the true f_0 value.

One point to be considered is the change rate of the fundamental frequency. In many natural processes it is helpful to assume that a system state changes slowly over time, and therefore attribute a higher probability to an estimate close to a previous state than to an outlier. This *diffusion* concept, often represented mathematically by Kalman filters, hidden Markov models, diffusion networks (Movellan et al. 1998), or stochastic differential equations, has been successfully applied in a wide range of tools such as face trackers (Shpungin 2000), financial analysis software, radio receivers etc. – and demonstrates that the best estimate for a new value is often the one which preserves the previous position, momentum, acceleration, or higher order derivative. The success of linear predictive coding is another example which shows that more often than not, in nature, everything flows[1]. Hess (1983) quotes studies by Sundberg (1979) that state the maximum rate of change possible for human vocalization is 1%/ms, and for speech ranges from 0.2–0.65%/ms (Black 1967). Therefore, in the postprocessor built for the work presented here, the probabilities of all values within a maximum change rate of 1% are multiplied by 1.5. A priori knowledge of the overall pitch range is included as an interval around the mean rather than as the space between the maximum and minimum of all detected points, because these values are very sensitive to false acceptance of harmonics and subharmonics. Points that deviate from the overall mean by more than 50% undergo a reduction in probability to one-half, the probability of points that lie out farther than at a factor of 1.8 are reduced by another 50%. The mean is updated after every step as $\mu_{n+1}=\mu_n-df_0/10$ to attribute a stronger weight to directly preceding points than to earlier ones.

The total pitch scale of this tracker ranges from 50 to 500 Hz. At higher frequencies, the signal's residual content of formants might get too close to peaks in the autocorrelation or

[1] (Heraclitus 500BC)

cepstrum. The lower bound of 50 Hz is arbitrarily set to allow small window sizes. Hess (1983, p. 64) lists several works describing the normal pitch range of speech: Fairbanks (1940), 65–450 Hz; Risberg (1961), 50–310 Hz; Hadding-Koch (1961), 50–500 Hz; Shaffer (1964), 110–500 Hz; Hollien (1972), 80–300 Hz; Rabiner et al. (1976), 50–500 Hz; Monsen and Engebretson (1977), 110–250 Hz. In vocal music, as noted in (Hess 1983), pitch encompasses 50–1800 Hz, and the total limits of human vocal phonation are reported to be 28 Hz and 3100 Hz.

3.4 Intensity/Loudness

Humans have the ability to recognize a speaker as loud or quiet disregarding how strong the actual stimulus is at our ears, because we are used to a modified vocal effort depending on the speaking situation (*Lombard effect*, see Traunmüller and Eriksson 2000; Junqua, Fincke, Field 1999). We can tell if a person is screaming, or talking barely louder than whispering, whether he/she is amplified by a rock concert size 40kW PA system or whether we hear the speaker through a barely audible radio in the background. In nonverbal speech research, quantities like power or energy have mostly been used to describe a loudness dimension. Using only power neither permits to make assumptions about absolute loudness, nor is it a valid representation of relative perceived loudness for different parts in one recording, since the characteristics of our hearing are ignored. It can merely be an incomplete description of the physical stimulus; incomplete because the perception of loudness in general and specifically speech depends on more than signal power, it is also a function of spectral distribution and range, duration, the soundfield, background noise etc. For speech, it has been shown that the perceived loudness can even be influenced by the visual channel. Glave and Rietveld (1979) synchronized speech recordings with video tracks of speakers that visibly spoke with different vocal effort and found that, although the visual information is unlikely to override the auditory perception, the images have a noticeable effect that is strongest if it supports the heard sounds, i.e. if a loud voice is played together with a film of an even louder speaker. The psychoacoustically motivated loudness model presented here fills this gap and has shown to be a useful tool in the recognition of nonverbal speech information. In addition to this, the possibility for applications arises in fields such as stress level monitoring in air traffic control, speech therapy tools, and automatic gain controls for hearing aids, cell phones, movie theaters.

For this loudness model study, a number of speech samples recorded at different vocal efforts are subjected to a psychoacoustic loudness model developed for this project. The recordings are normalized with regard to their loudest frequencies in a short interval at the beginning of each sample. The data is then transferred to the frequency domain and critical bandwidth filters are placed around spectral peaks for every time frame. With data adapted from the ISO R532B/Zwicker model, the loudness in each band is then calculated and integrated over the whole spectrum to obtain a loudness value for this time window. Compared to other speech intensity dimensions like power, these loudness values as introduced in (Quast 2000) prove to be a better representation of perceived speech volume.

3.4.1 Some Properties of Human Hearing

Human hearing and speech have co-evolved – both in "hard-" and "software" – over a long time span[1] and are thus finely tuned to each other. This fact opens up a number of

[1] Of course it is impossible to precisely date the genesis of language, but it seems likely that it is linked to the advent of our species, Homo sapiens sapiens, 40,000-150,000 years ago (A. Popescu-Belis, personal communication.) Bickerton (1990) believes that in earlier times man used "protolanguage," words occuring in sets, and that the emergence of Homo sapiens sapiens marked the arrival of syntax. Interesting debates in

possibilities for meaningful and efficient speech data representation (cf. Strube's overview of psychoacoustic feature derivation for speech recognition, 1994).

The perceived loudness of an (ideally purely sinusoidal) tone depends mostly on its intensity and on its frequency. The range of human hearing, roughly speaking, covers an interval from 20 Hz to 16 kHz with the highest sensitivity between 2 and 5 kHz, see Fig. 3.4. The physical intensity of sound pressure is described in *decibel* (dB). To rate the perceived volume of sounds with different frequencies, the loudness of a sound is compared to the loudness (in dB relative to the hearing threshold) of a 1000 Hz pure tone that approaches the listener as a plane wave with frontal incident (Schroeder 1999, Zwicker 1990). Its unit is the *phon*. The unit *sone* was created to quantify perceived loudness on a linear scale: 1 sone describes the same loudness as a 40 dB 1 kHz tone, a sound that is sensed twice as loud has loudness 2 sone etc. Zwicker and Paulus (1972) express the relation of stimulus (sound pressure) and perception (loudness) as

$$N' = 0.064\,{}^{sone}\!/_{Bark} \cdot 10^{0.025 L_{ETQ}} \cdot \left[\left(1 + \tfrac{1}{4} 10^{0.1\left(L_G - a_0 - L_{ETQ}\right)}\right)^{0.25} - 1 \right] \qquad 3.8$$

where a sound of level L_G of a (small) frequency band generates a specific loudness N'. The transmission of freefield sound to our hearing system through the head and the outer ear is described as *attenuation* a_0. The attenuation is zero for low frequencies, negative in the interval of highest sensitivity, and positive for higher frequencies a shown in Figure 3.4. The excitation threshold in quiet, i.e. the level necessary to produce an audible sound, is included as L_{ETQ} which starts off with positive values at low frequencies and approaches a constant 4.0 at about 1000 Hz, see Fig. 3.4. The measurement unit of attenuation, excitation threshold, and sound level is dB, the specific loudness is noted in sone/Bark (see below). Slightly different attenuation behavior is observed for diffuse or planar soundfields, but since this effect shows mostly in higher frequencies with little relevance for speech perception, it is ignored in this study.

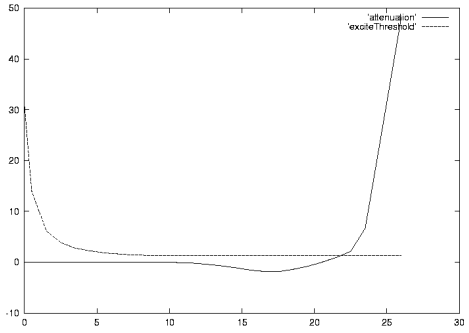


Figure 3.4 Attenuation a_0 and excitation threshold in quiet L_{ETQ} in a free (non-diffuse) soundfield as a function of frequency in Bark, outlining the sensitivity of human hearing

Masking

The term *spectral masking* denotes our hearing's property to weaken or completely suppress a sound if its frequency is close to that of a stronger sound. The stronger sound, called the *masker*, raises the excitation threshold for neighboring frequencies; the closer the frequency, the more so. The masked area, or *specific loudness* for each frequency, depicts an asymmetric cone in the frequency–intensity plane. The slope towards lower frequencies is steeper than towards higher frequencies, and therefore the area under the lower-frequency branch is usually set to zero in psychoacoustic models, as shown in Figure 3.5. Integrating the specific

---

anthropological linguistics rage on the origin of language, whether it is linked to physical changes in the brain, and how it relates to a complex and controllable vocal tract (Hurford, Studdert-Kennedy, Knight 1998).

loudness over frequency yields the total perceived loudness. The application of psychoacoustic effects on the perception of speech has been described by Zwicker (1990), and Glave and Rietveld (1975, 1977).

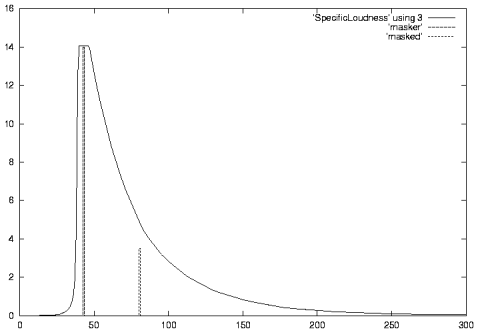


Figure 3.5 Specific Loudness cone created by the masker. The area under the curve depicts total perceived loudness. Sounds below the cone are inaudible (masked), sounds piercing the surface are *partially masked*.

Similar phenomena also occur in the time domain as *temporal masking*, mainly as *postmasking*, when a weak sound closely following a stronger stimulus is completely or partially masked by the first sound. To a certain extent a strong sound can even overshadow a stimulus that occurred earlier. The time gap for this *premasking* is considerably shorter than for postmasking. When defining the size of speech loudness maxima for voiced parts, temporal masking effects can largely be ignored for the purpose of the loudness maximum based data analysis described in Section 4.4 since the average distance between syllables/maxima is about 0.3 s with almost no interval shorter than 0.2 s (see also Zwicker 1990), which is the upper border for temporal masking, and one is interested in the signal properties at the point when the strongest stimulus is created.

Critical Band Rate

As a result of the masking effect shown in Figure 3.5, the overall loudness for two tones heard simultaneously is smaller than the sum of each individual tone if both specific-loudness curves overlap. If the frequency difference between both tones gets smaller than a specific interval, the loudnesses do not add up anymore. This interval is called a *critical band*, measured in *Bark* (named for the German engineer Heinrich Georg Barkhausen). Critical bands roughly correspond to equal length segments on the cochlea in the inner ear. The first 5 critical bands are approximately linear on a frequency scale in Hz, 1 Bark equaling 100 Hz. Above 5 Bark, the intervals grow logarithmically with one Bark equaling approximately 1/5 of its center frequency in Hz (see Zwicker 1990). The whole range of human hearing encompasses about 24 Bark (15.5 kHz). The critical bands are often approximated by third-octave frequency intervals.

3.4.2 Aspects of Speech Production

As pointed out in Section 3.1, two components play a principal role in forming speech: the excitation signal and the vocal tract which acts as a filter. For voiced speech, the opening and closing of the glottis – the aperture created by the vocal cords – generates a periodic excitation that is convolved with the impulse response of the vocal tract. Unvoiced sounds are produced by partially obstructing the airflow in the vocal tract without the periodic modulation of the glottis; in this case the excitation can be modeled as random noise.

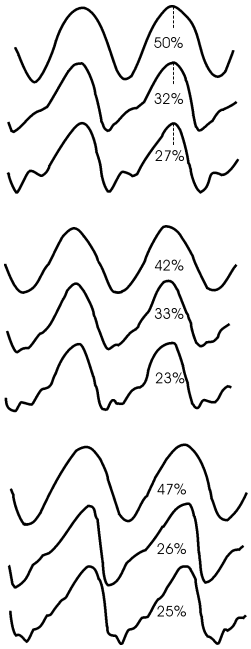


Figure 3.6 Glottis pulse diagrams for a female (center) and two male speakers (from Monsen and Engebretson 1977). The upper row for each speaker displays the glottal pulses for a soft voice, the center row for regular conversation, and the bottom one for loud speech.
The time interval occupied by the downward sloping part of the pulse shape is given as a percentage of the total period.

The glottal pulse shape yields some cues about the vocal effort exercised by the speaker: the steeper the downward slope, the stronger the vocalization, see Fig. 3.6. In this diagram, the time occupied by the downward slope, corresponding to the time interval of glottis closure, is given as a percentage of the total period. With rising vocal effort, this part becomes smaller and the steepness factor (Fant 1979) that describes the downward slope of the decaying glottal pulse becomes greater. The abrupt, forceful closing of the vocal cords results, for increasing vocal effort, in an augmentation of higher frequencies in the source spectrum of the glottis pulse.

Fant (1979) proposed a model that divides the signal $s(t)$ during one glottal period into three segments – glottal opening, falling branch (of the glottogram), and closed glottis – which he describes as trigonometric functions of the pulse rise frequency ω_R and a steepness (of the falling branch in the glottogram) factor K:

Glottal opening:
$s_1(t) = \hat{s} \cdot 0.5(1 - \cos \omega_R t)$ for $0 < t < \pi/\omega_R$

Falling branch:
$s_2(t) = K \cos(\omega_R t - \pi) - K + 1$ for $\pi/\omega_R < t < \frac{1}{\omega_R} \arccos \frac{K-1}{K}$

Closed glottis:
$s_3(t) = 0$ for $\frac{1}{\omega_R} \arccos \frac{K-1}{K} < t < T_0$

The Fourier transform of this signal aptly reproduces the vocal source spectrum and the high frequency attenuation for rising K.

To obtain the shape of the pulse from a recorded waveform, one inverse filters the recorded speech sample to remove the filter effect of the vocal tract (cf. Hess 1983), in the same way that was explained earlier in the context of pitch tracking. For this, the spectral envelope in each time window of the sample is computed. This envelope is then inverted so that each pole or formant now becomes a zero of the inverse (reciprocal) filter's frequency response function, and each zero turns into a pole. Both mechanical and electrical methods exist to inverse filter a speech signal, but for real life applications one is restricted to noninvasive methods such as linear prediction (LPC) (see Schroeder 1999), where the inverse

filter coefficients can be created without effort from the linear predictor in each time window. Unfortunately, glottal pulse slopes do not allow a one-to-one mapping to vocal effort or even perceived loudness, the steepness is also determined by voice quality (Stevens 1977, see also Laver 1980). At the same intensity level, in breathy vocalization the glottis does not close completely, thereby obscuring the glottal pulse shape edges, and in creaky voices (vocal fry) the vocal cords shut very forcefully, creating high frequency energy that is usually associated with screaming.

3.4.3 The Absolute Loudness Model

The model is qualitatively evaluated on the nonverbal-speech database. In order to be able to directly compare data from the same speaker with different vocal effort, a second database of 35 sentences was recorded with 7 people vocalizing one sentence ("I can see you Bob and Steve.") with 5 different degrees of vocal effort. The first degree is whispered, unvoiced speech, the second one low volume talk, level 3 regular face-to-face conversation in a quiet environment, level 4 describes louder conversation as in yelling to a person through a room filled with people, and at the loudest degree it was requested people scream at the top of their voices. The recordings were made in the low-reverberation ISO booth of a professional audio studio on equipment with a signal to noise ratio sufficient to adjust the recording settings to the loudest voice and record all lower volume voices with the same gain and the same mouth-to-microphone distance. In the following, the process of obtaining loudness values is outlined step-by-step with diagrams of a low-volume recording example on the left and a high-volume sample by the same speaker on the right side. The average spectrum is taken as an example for one time window; in the actual process the same steps are applied to each single frame.

Normalization

In the first step of this model, the speech sample is normalized in order to make it invariant to amplifier gain.

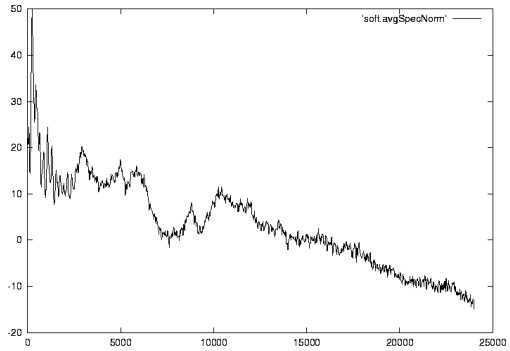


Figure 3.7a dB-Normalized average spectrum for a soft voice. The intensities for frequencies (given in Hz here) above the fundamental-frequency-range peak are small. (The average spectra are only displayed for these examples. In the loudness computation, spectra of single frames are used.)

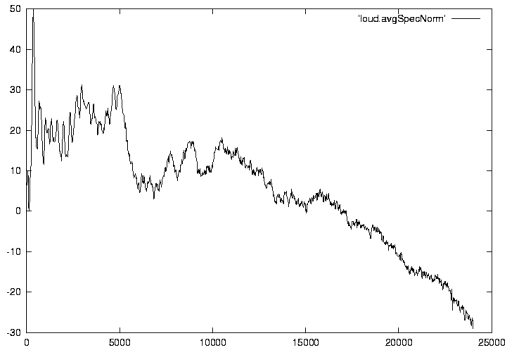


Figure 3.7b Normalized average spectrum for a loud voice. Besides the peak at the fundamental frequency range, a plateau in the harmonics-range of f_0 is now present.

Initially, the recording is transferred to the frequency domain via a short time FFT. The autocorrelation-based voiced/unvoiced classifier is used to pick out the first voiced windows until a number of frames corresponding to a total length of 2.5 seconds of voiced speech is present (or until the end of the file is reached, for short recordings). In each time window, the loudest frequency is selected to compute the average loudest frequency. The normalizer function then returns a value that is added to the amplitude at each frequency in the further processing so that the average loudest frequencies are the same for all recordings. Intensity levels are computed in decibel, so, as an example, the return value for a sample with twice the amplitude of the original recording is 6dB smaller.

Transforming Frequencies to Critical Bands

As outlined above, human frequency perception is based on a (for the most part) logarithmic scale of critical bands measured in Bark. The first 5 critical bands are roughly linear on a frequency scale in Hz, 1 Bark equaling 100 Hz. Above 5 Bark, the intervals grow logarithmically with one Bark equaling approximately 1/5 of its center frequency in Hz. For this model, frequency values in Bark were interpolated from 48 Hz–Bark pair data points as noted by Zwicker (1990). Frequencies below 5 Bark are calculated on a linear scale, higher frequencies are interpolated piecewise logarithmically, i.e. as [Bark] = x ln [Hz] + c with x and c calculated in each interval from the given two adjacent data points.

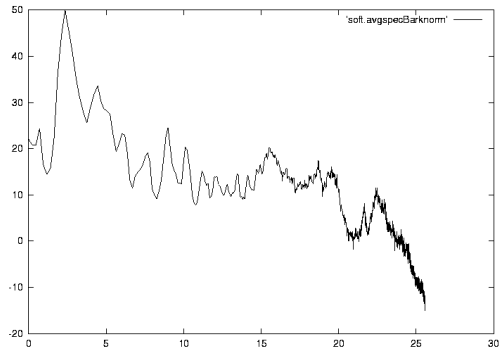


Figure 3.8a Bark-spectrum for a soft voice

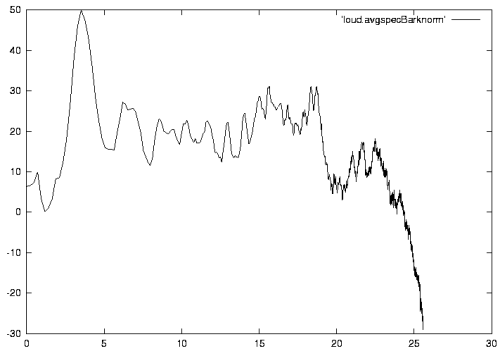


Figure 3.8b Bark-spectrum for a loud voice

The warping of the frequency axis from Hz to Bark now adequately displays which regions are more prominent to human hearing. The harmonics plateau of almost equal loudness as seen in Figure 3.7b now occupies the majority of the frequency range in Figure 3.8b. Holte and Margolis (1987) found that the loudness of third-octave bands of speech is approximately constant from 500 through at least 3150 Hz. This matches the results of this study: up to 300Hz, the loudness is dominated by f_0, usually the loudest frequency. Above that interval, the loudness is determined by the harmonics, filtered by the formants of the vocal tract. The width of this band depends on vocal effort.

Determining Specific Loudness

The sensitivity of our ears with regard to a sound's frequency and intensity is taken into account through specific loudness values N' for each frequency group as returned by the FFT. It is computed according to Equation 3.8 as a function of the intensity level in the particular frequency group, the attenuation, and the excitation threshold level in quiet. Attenuation and

excitation threshold are linearly interpolated for each frequency value from 24 datapoints reported in (Paulus, Zwicker 1972).

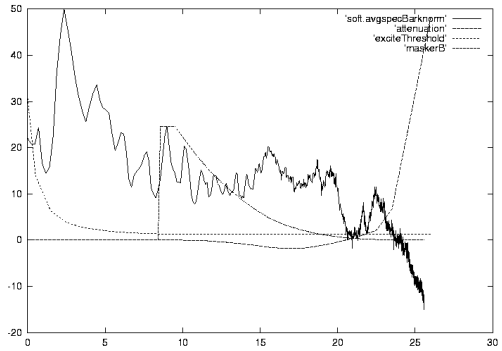


Figure 3.9a Attenuation, excitation threshold, and maskshapes as applied to the Bark-spectrum of the soft voice.

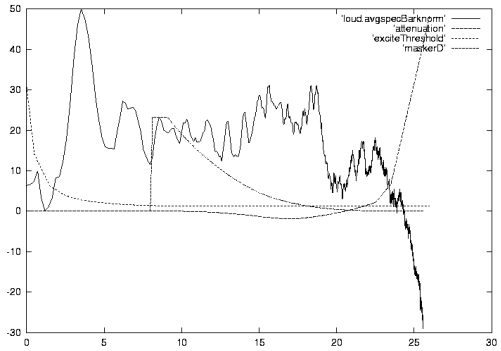


Figure 3.9b Attenuation, excitation threshold, and maskshapes as applied to the Bark-spectrum of the loud voice

Masking

Once the core loudness levels for the whole spectrum are developed, spectral masking phenomena are incorporated. Graphically, this is achieved by placing masking cones over all loudness values as hinted in Figures 3.9a,b from the previous step. This results in the perceived spectrum displayed in Figs. 3.10a,b. In the algorithm, a sweep through the spectrum from low to high frequencies classifies each level value either as plateau point (the ceiling of the masking shape, an interval of 1 Bark) or as slope point.

If a frequency value is in the interval of a plateau, it either assumes the specific loudness of the plateau if the frequency group level is smaller than the plateau level, or it defines a new plateau if the frequency group level pierces the plateau. In the latter case, the smaller frequency positions from the current position to this position minus 0.5 Bark are all set to the new plateau level, additionally, the new plateau borders are established plus/minus 0.5 Bark from the current frequency for the lower/upper border, respectively.

If the original loudness level at the current frequency is not on a plateau but on a downward slope, it either defines a new plateau as outlined above if the core frequency is higher than the specific loudness created by the lower frequency masker, or it is completely masked if it does not reach the slope. In the complete masking case, the new specific loudness level N'_i is determined by the degree of the downward slope dN'/df and the difference f_i-f_{i-1} to the next smaller frequency value:

$$N'_i = \frac{dN'}{df}(f, N'_{i-1}) \cdot (f_i - f_{i-1}) \quad .$$

The decline dN'/df of specific loudness is a function of frequency and intensity. The downward slope is greater for high loudness levels and low frequencies. The slope values are linearly interpolated from data as tabulated by Zwicker (1972). Like in the ISO model (which is based on the Zwicker data), the small area under the lower-frequency slope of the masking-cone is set to zero. This loss is compensated for by slightly adding to the right-side, higher frequency area of the specific loudness, leading to equal results on average.

The spectrum from the previous diagram has been adjusted to the sensitivity of human hearing, also, masking is now accounted for. The actually perceived spectrum is given by the upper, continuous curve.

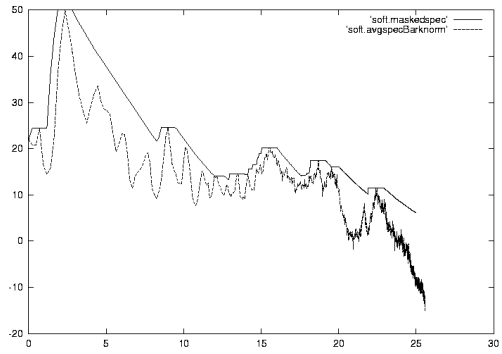


Figure 3.10a Specific loudness spectrogram for a soft voice

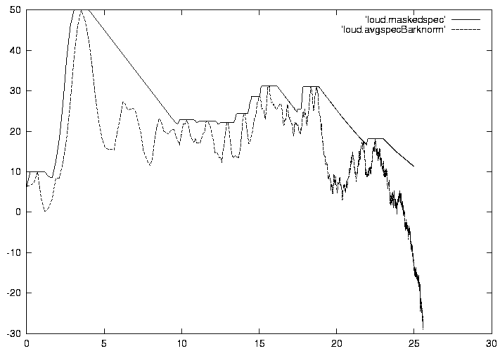


Figure 3.10b Specific loudness spectrogram for a loud voice

Integration

The perceived-loudness value N for the time frame is determined simply by integrating the specific loudness N' over frequency:

$$N = \int_0^{24} N' \, \mathrm{d}f \qquad \text{or rather} \qquad N = \sum N' \, \Delta f \qquad \text{for the discrete case.}$$

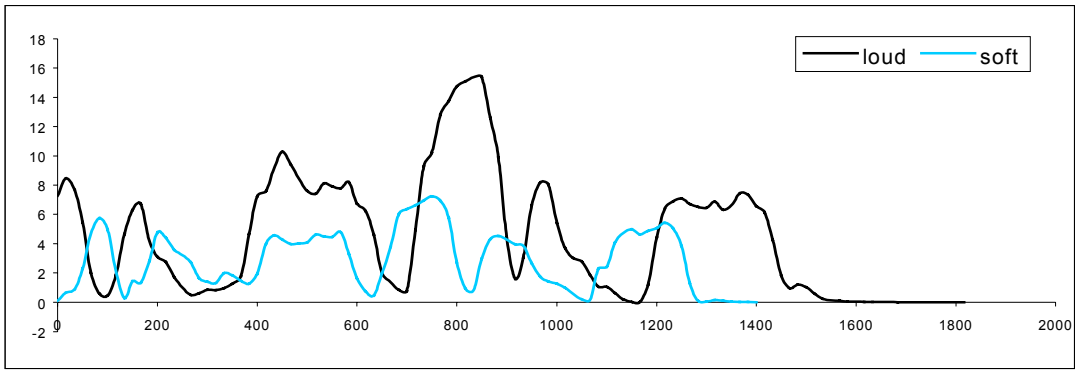


Figure 3.11 Loudness contours (intensity over time [ms]) for the same sentence vocalized with two degrees of vocal effort from the examples above: soft voice (bottom contour) and loud voice (upper).

Plotted as a function of time, these loudness contours give a meaningful representation of perceived loudness. The total loudness impression for an utterance can be described with histogram percentiles, specifically, as the level exceeded 10% of the time (Zwicker 1990).

3.5 Spectral Parameters/Timbre

The spectral parameters used here are fast Fourier transforms from every time window averaged over the course of the whole sentence, as is done for instance in similar experiments by Banse and Scherer (1996) and Pittam et al. (1990). The resulting long-term average spectrum is then divided into the 5 intervals 0–0.5 kHz, 0.5–1 kHz, 1–2 kHz, 2–4 kHz, 4–8 kHz. Each bin is represented by its loudest frequency in decibel. To normalize the data, the entry from the frequency band with the highest intensity is set to 50 dB, all other values are shifted up or down on the loudness axis accordingly. For all 150 speech recordings, the 0–0.5 kHz held the highest intensity. This band holds the fundamental frequency range, which, on average, is the strongest contributor to the perceived loudness. The next higher frequency bins include the formant and harmonic range and are an indicator of vocal effort and vocalization register, as outlined above. Auditory comparison of synthetic loudness and pitch contours with and without spectral information suggests that spectral cues are also to a big extend responsible for the perception of valence (Quast 1999).

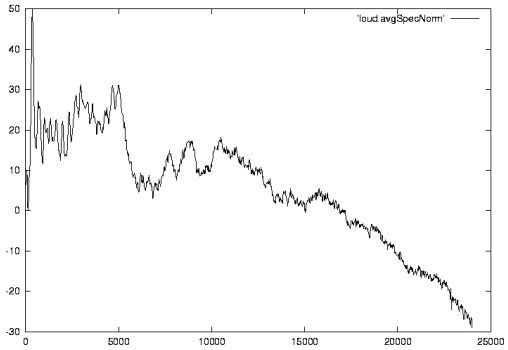


Figure 3.12 A long term spectrum for one sentence

Since the glottal excitation signal is filtered by the vocal tract, the long term spectrum also contains cues about the speaker's vocal tract (Schroeder 1967; Strube 1999; Freienstein, Müller, Strube 1999). The neural network's selection of pattern recognition parameters for the *strong* impression also supports the notion that one is able to obtain an impression of the speaker's physique beyond the vocal tract, see Section 4.5.2.

4 Pattern Recognition

To see if the affective, psycholinguistic value of a speech recording can be represented as a function of its signal processing parameters, neural networks are used as pattern recognition engines to map one set of features onto the other one. Ideally, one wants to be able to enter values like average pitch, loudness as described in Section 3.2, and have the program tell the user "this recording sounded rather happy, not really angry, and would be perceived to have leadership ability". This mapping task is given to a neural network type called multilayer perceptron. It is trained by showing it example pairs of acoustic and affective data, and, if it has learned successfully, the network is able to generalize, i.e. see new examples, and be able to assess its affective content. Since listeners' agreement strongly fluctuated in the evaluation of the speech recording, the neural networks are programmed to consider the quality of a data point when learning from it. A new psychoacoustically motivated data representation technique called *lombada* is introduced that stores the prosodic information at a fraction of the space occupied by the original recording.

4.1 Neural Networks

Neurocomputing describes a wide field of information processing systems that have the fascinating capacity to learn tasks autonomously by monitoring a process and adapting internal states – without algorithmic or programmed computing or knowledge that explicitly states how the problem is to be solved (as is the case in *expert systems*). Major advantages of neurocomputing are the universal ability to generalize, noise tolerance, adaptability, and fault tolerance. Neural networks have found application in areas such as pattern classification, probability density estimation, coding/decoding, denoising, incomplete-data restoring, function estimation, and control tasks.

Neural networks have also found their way into many areas of speech processing (Rahim 1994) such as formant extraction, voiced/unvoiced classification, and the mapping of phonetic to articulatory features (Freienstein, Müller, Strube 1999), also speech synthesis (Sejnowski, Rosenberg 1987; Scordilis, Gowdy 1989). Another task in speech processing where the application of neural network decision making seems meaningful is the peak picking process in the postprocessing of autocorrelation and cepstrum, see Section 3.3.

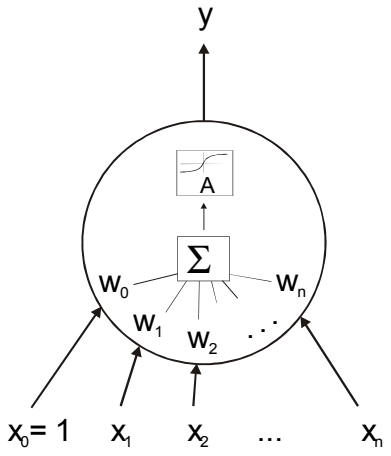


Figure 4.1 A single neuron. The entries of the input vector are multiplied with their respective weights and summed up. The output y is then formed by the activation function A of the scalar product.

4.1.1 *Definition*

Originally motivated by the brain's massively parallel distribution of neurons, neural networks have developed into many different directions over the years. The common ground is summarized by Robert Hecht-Nielsen as follows:

> A *neural network* is a parallel, distributed information processing structure consisting of processing elements (which can possess a local memory and can carry out localized information processing operations) interconnected via unidirectional signal channels called *connections*. Each processing element has a single output connection that branches ("fans out") into as many collateral connections as desired; each carries the same signal – the *processing element output signal*. The processing element output signal can be of any mathematical type desired. The information processing that goes on within each processing element can be defined arbitrarily with the restriction that it must be completely local; that is, it must depend only on the current values of the input signals arriving at the processing element via impinging connections and on values stored in the processing element's local memory. (Hecht-Nielsen 1991, pp. 2–3)

Like in the brain, knowledge is acquired by the network through a learning process, and the influence of one cell/neuron on another one is determined by a synapsis'/connection's strength, in the artificial neural network symbolized as a *weight*.

4.1.2 *Decision Making*

In the simplest case, the output y of a neuron is produced simply by scalar multiplication of the input vector $\vec{x} = \{1, x_1, ..., x_n\}$ with the weight vector $\vec{w} = \{w_0, w_1, ..., w_n\}$. The output is the neuron's estimate of the function it wishes to model. This processing element is sometimes referred to as *Adaline* (adaptive linear element, Widrow and Hoff 1960). The factor w_0 is always multiplied with an input of 1, the *bias*.

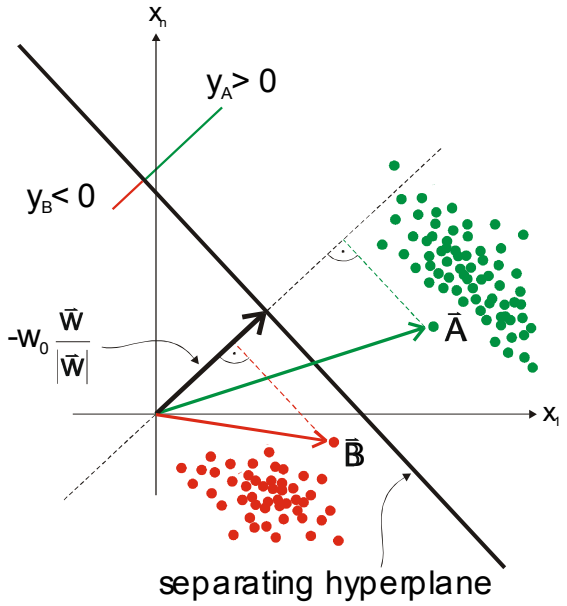


Figure 4.2 A classification task: the weight vector $\vec{w}$ defines the separating hyperplane perpendicular to $\vec{w}$ that divides the total space into two areas of red or green points. Each data vector is projected onto the (unit-length) vector $\vec{w}/|w|$, and the bias term w_0 is added. The sign of the solution y determines the class.

A classification into two groups can be done according to the sign of the result: an activation function A (see Figure 4.1) puts every data point $\vec{x}$ that yields a positive output y into one bin, an $\vec{x}$ with a negative result into the other one. An activation function that meets these requirements is the *Heaviside step function* which returns 0 for negative inputs and 1 for inputs greater than or equal to 0. This classification task can easily be elucidated geometrically as shown in Figure 4.2.

In other cases, one might wish to know how far away an input point is from the separating hyperplane, or it is the goal to model a function that returns continuous, not binary, values. In this case, the scalar product of input and weights is modified by a different type of activation function. Desirable properties are the ability to limit the output to a certain range, for instance –1 to +1, or 0 to 1, but also to show good separation in places of high data concentration. Such an activation is given by sigmoid (s-shaped) functions. Assuming the data consists of two classes $C_{1,2}$ with Gaussian distributions having equal covariance matrices, it can be shown that the *logistic function,*

$$s(x) = \frac{1}{1 + \mathrm{e}^{-x}} \quad , \tag{4.1}$$

is tuned to this task and additionally allows the outputs of this discriminant to be interpreted as posterior probabilities $p(C_k|x)$ that point x falls into category C_k (Bishop 1995). As opposed to the Heaviside function, the sigmoid is differentiable, a fact that will be important in the derivation of the backpropagation learning algorithm.

4.1.3 Learning

The different types of learning fall in one of the categories *supervised training*, *graded* or *reinforcement training*, and *self-organization*. In supervised training, the network is presented with an input/output vector pair, and the system adapts its inner states (the weights) to find a mapping that is the best representation of the output values as a function of the input, judged by an arbitrary error function. Once the network has converged to an optimal state, the weights are 'frozen,' and the net is now ready to work as a pattern recognizer, an input vector is run through the network and the output is the net's estimate of the function to be modeled. During reinforcement training, the network is not supplied with the correct answer, but only receives information how good or bad the response was. Unsupervised or self-organized training is the most autonomous process; here, the network receives only input values and organizes the data in a desired manner, e.g. to model the distribution or to build clusters. In this work, for example, the pattern recognizer is presented the acoustic parameters as input and the listeners' evaluations as output, hence the learning is supervised.

In the case of the Adaline with a Heaviside activation function, a simple learning, i.e. updating of the weights, can be achieved by

$$\vec{w}_{\mathrm{new}} = \vec{w}_{\mathrm{old}} + \delta\vec{x} \quad ,$$

where δ is the difference of the correct class number y' (0 or 1) and the network's estimate y with the same possible values. This construct, named *perceptron* by its developer Frank Rosenblatt, has the ability to find a hyperplane that correctly and completely divides two linearly separable sets, and it converges to this state in a finite number of steps (Rosenblatt 1958, see also Haykin 1994).

More generally, it is desirable to find the set of weights $\vec{w}$ that minimizes output error (often measured as mean squared error) $F(\vec{w})$. It can be shown (Hecht-Nielsen 1991) that F, under very general assumptions, is differentiable with respect to w_i,. Thus, the direct, "greedy" way to minimize error (used, for instance, by the backpropagation learning algorithm below) is to descend the error surface towards a minimum in the direction of the negative gradient:

$$\vec{w}_{\mathrm{new}} = \vec{w}_{\mathrm{old}} - \alpha\nabla_w F(\vec{w}) \tag{4.2}$$

where the cofactor α, the *learning rate*, limits the "step-length" of the descent.

4.2 Multilayer Perceptrons

Much like real, biological neural networks, artificial neural networks gain their power and robustness through the massively parallel combination of simple units. The multilayer perceptron (MLP) arranges single neurons in *layers* and passes through information in a *feedforward* manner, that is, the input vector is entered at the first layer, sent through the next layer(s) and processed by its neurons, until it reaches the output layer. Multilayer perceptrons have three distinct features:

1. The network has at least three layers: an input layer which does no data processing but holds the input vector, one or more hidden layers of neurons with nonlinear activation, and an output vector without activation functions.
2. The activation functions are nonlinear and differentiable. Usually, a sigmoid function such as the logistic in Equation 4.1 is applied. (If the activation were linear, the network could perform no better than a single-layer array of neurons.)
3. All neurons are fully interconnected (at least in the initial state); every neuron receives the outputs of all neurons in the previous layer as input.

The general layout looks as follows:

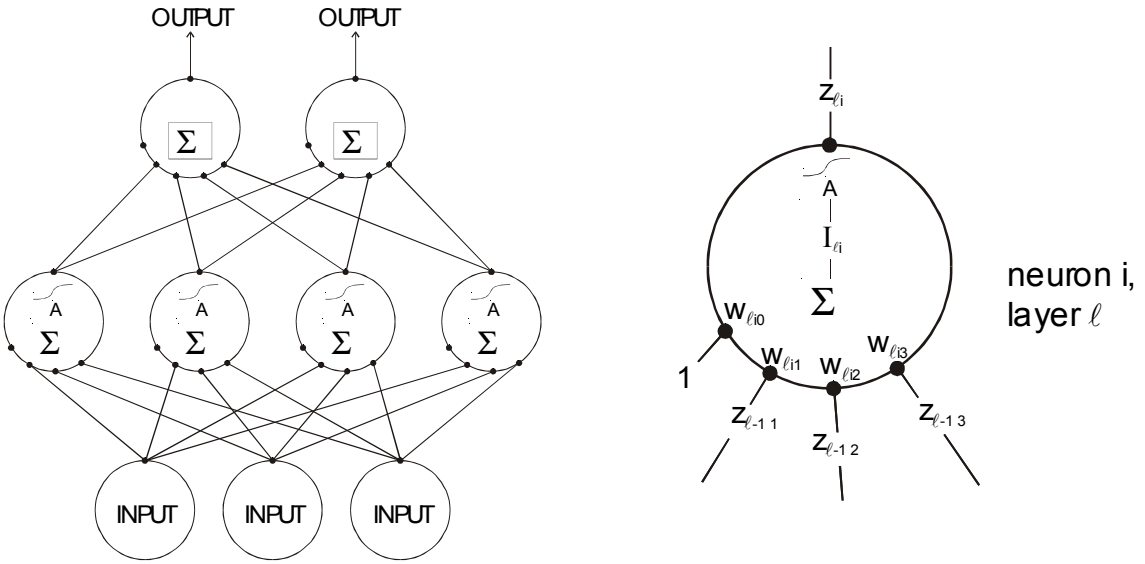


Figure 4.3 Left: an example of a three-layer MLP. The three-dimensional input layer holds the data for processing in the four-neuron hidden layer. The two output neurons simply return the scalar product of the weight vector and the input from the preceding layer without a nonlinear activation function. The single neuron on the right illustrates the symbols used in the derivation of the backpropagation algorithm. The indices l, i, j denote the number of the layer, neuron, and weight, respectively. I_{li} marks the product of input $\vec{z}_{l-1}$ and weights $\vec{w}_{li}$ which is then processed by the sigmoid activation function s and returned as the output z_{li} of the neuron.

4.3 Backpropagation

Multilayer perceptrons are closely connected to a learning paradigm called backpropagation. This powerful method was introduced and rediscovered numerous times in various versions throughout the last half of the 20th century (e.g. Robbins, Monro 1951; Werbos 1974; Parker

1986), but the big breakthrough came in 1985/86 when David Rumelhart, Ronald Williams and others at UCSD's PDP group advanced backprop to a technique that could easily be implemented on computers (Rumelhart, Hinton & Williams 1986). In this supervised training algorithm, the network receives an output-input data pair. The input data is fed through the layers and returns the neural net's estimate of the output. Now the deviation from the true output is computed, and this error is propagated layer by layer from the output to the input neurons. The next paragraphs outline how the weight change of each neuron is expressed as a function of the change in the next higher (towards the output) level (adapted from Hecht-Nielsen 1991).

As mentioned in Eq. 4.2, backpropagation descends the error surface along the gradient to find the minimum-error set of weights. This direction is given by

$$-\nabla_w F(\vec{w}) = \frac{\partial F(\vec{w})}{\partial w_{lij}} = \frac{\partial F(\vec{w})}{\partial I_{li}} \frac{\partial I_{li}}{\partial w_{lij}} \qquad 4.3$$

and can be expressed through the dependence of I_{li} (the linear combination of that neuron's inputs z_{l-1} with its weights, see Figure 4.3) on that neuron's weights w_{lij} by using the chain rule.

Introducing $\delta_{li} = \frac{\partial F}{\partial I_{li}}$ and using the definition $I_{li} = \sum_q w_{liq}\, z_{l-1,q}$, Equation 4.3 turns into

$$\frac{\partial F(\vec{w})}{\partial w_{lij}} = \frac{\partial F(\vec{w})}{\partial I_{li}} \frac{\partial I_{li}}{\partial w_{lij}} = \delta_{li} \frac{\partial}{\partial w_{lij}} \left(\sum_q w_{liq} z_{l-1,q} \right) = \delta_{li} z_{l-1,j} \quad . \qquad 4.4$$

At the output layer, which has no sigmoid activation function, z_{li} equals I_{li}, and thus

$$\delta_{li}^{\text{output layer}} = \frac{\partial F}{\partial z_{li}} = \frac{\partial}{\partial z_{li}} \sum_p \left(y_p - z_{lp} \right)^2 = -2\left(y_i - z_{li} \right) \qquad 4.5$$

for a mean squared error F and the output value y_i at output-layer neuron i. Thus, the deltas for the output layers can be computed – directly after the datapoint has been fed through the network – from the true value y and its estimate z_{out}.

Now that the deltas for the last layer are known, one can explicitly state the functional dependency of the hidden-layer deltas on the ones in the output. By expressing a hidden-layer delta through values in the next higher, already computed, layer, the error is backpropagated layer by layer from the output to the first hidden layer. In the hidden layers, whose neurons contain an activation function s, one obtains

$$\delta_{li} = \frac{\partial F}{\partial I_{li}} = \frac{\partial F}{\partial z_{li}} \frac{\partial z_{li}}{\partial I_{li}} = \frac{\partial F}{\partial z_{li}} \frac{\partial s(I_{li})}{\partial I_{li}} = \frac{\partial F}{\partial z_{li}} s'(I_{li}) \quad . \qquad 4.6$$

To connect layer l with layer $l+1$, the partial derivative is expanded with the chain rule to read

$$\frac{\partial F}{\partial z_{li}} = \sum_{p=1}^{M_{l+1}} \frac{\partial F}{\partial I_{l+1,p}} \frac{\partial I_{l+1,p}}{\partial z_{li}} \qquad \text{with } M_{l+1} \text{ : number of neurons in layer } l+1 \qquad 4.7$$

Substituting the results of Equation 4.7 into Eq. 4.6 yields

$$\delta_{li}^{\text{hidden layer}} = s'(I_{li}) \sum_{p} \delta_{l+1,p} w_{l+1,pi}$$

These deltas and the neuron-outputs z_{li} can now be used to compute the gradient in 4.4 and update the network's weights. Calculating the actual gradient requires averaging over a (theoretically infinite) number of datapoints, but it suffices to estimate it from an appropriate number of points (*batch learning*). The network will even carry out a gradient descent if the weights are updated after every datapoint is observed. This *jump-every-time learning* variant is used in this work, the weights are updated according to

$$w_{lij}^{\text{new}} = w_{\text{lij}}^{\text{old}} - \alpha \delta_{li} z_{l-1,j} \qquad 4.8$$

every time a point is trained. The learning rate α is usually adjusted empirically to a small positive value.

The backpropagation-trained multilayer perceptrons introduced here are comfortably robust and efficient. Although for special pattern recognition problems other schemes can be superior, the MLPs most of the time still show good performance, and its uncomplicated versatility renders it applicable for a wide range of tasks. As Robert Hecht-Nielsen (1989) showed,

> Given any $\varepsilon>0$ and any L_2 function $f:[0,1]^n \rightarrow \mathbb{R}^m$, there exists a three-layer backpropagation neural network that can approximate f to within ε mean squared accuracy. (quoted from Hecht-Nielsen 1991, p.132)

Functions f belong to L_2 if each of f's coordinate functions is square-integrable on the unit cube which includes all continuous or even piecewise continuous (on a finite number of subsets in the domain) functions. The unit-hypercube constraint can easily be relaxed to any compact (closed and bound) set.

4.4 Training on the Psycholinguistic and Signal Processing Data

4.4.1 General considerations

When trying to find affective patterns in speech, the least biased solution would be training a network on a running speech signal that has been evaluated according to its nonverbal content, to assure that no acoustic cues are missing in the representation. However, the huge flow of data and the high intrinsic dimensionality obviously render this approach unpractical.

Even if the pattern recognizer focused on tracking a small number of parameters like pitch and loudness over time, the plethora of possible f0 and intensity contours would still make the training process, if at all possible, extremely tedious. Cohn and Katz (1998) tried to classify pitch contours by modeling them as the best fit (in an rms-error sense) of one of 16 function types. However, it appears doubtful that a sinusoid for instance can capture the intricacies of an actual f_0 contour. Moreover, similar sentences with different lengths, for example, "I can see the dog" and "I can see the big yellow ugly dog," can be spoken with the same affective content and, for the syllables that appear in both sentences, with the same stress in both phrases. The 5 additional syllables in "big yellow ugly" would however strongly change the contour fit with a sinusoidal. Here, hidden Markov models might lead to a more flexible, warp-invariant contour classification solution. As the example above with two contours of different lengths hints, it might not be necessary to consider the whole contour for the pattern recognition, but merely concentrate on values at important points – such as the focus of a sentence, the last or the first syllable – and see how they relate.

Most studies have described a sentence's acoustic parameters with aggregate measures such as mean pitch, standard deviation of loudness etc. that don't attempt to model the dynamics but globally summarize values.

The 18 parameters used in this work are also mostly non-dynamic, see Table 3.1. In addition to the aggregate dimensions, pitch and loudness data is also extracted at two prominent points, the loudest one and the last voiced one.

4.4.2 Loudness Maximum Based Data Analysis

The loudness contours displayed in Figure 3.11 demonstrate how strongly the intensity varies throughout an utterance. It seems plausible that a pitch value in a loud time frame is perceived with more emphasis than pitch when the loudness is low. Due to temporal masking, some frames might not be perceived at all. In this case, a sentence's pitch contour might sound very similar to a synthetic contour where each loudness peak is represented by only one f_0 value. To verify this assumption, sound files were prepared in which a sine wave was frequency modulated[1] to play a speech recording's fundamental frequency contour. The amplitude is the same as in the actual speech recording. From this original loudness/pitch contour, a second file with the same loudness profile is prepared. The pitch values at loudness maxima are picked from the original file, but in between two loudness peaks, the instantaneous frequency is linearly approximated from the fundamental frequency at the two adjacent peaks. Even with this crude approximation, the original and the synthetic contours had the same affective quality and sounded almost alike. At slow speech rates below 2 peaks per second the difference in perceived affect for these soundfile pairs was still small (Quast 1999).

This finding can be applied to store the affective content carried by prosody very efficiently by only extracting data at voiced loudness maxima. One loudness maximum usually corresponds to one phonetic unit, so this *loudness maximum based data analysis* (lombada) can be thought of as syllable based data extraction. Especially in slow speech, more than one loudness peak can occur per linguistic syllable (e.g. "hello-o"), but since the prosody is not carried by *verbal* units but *vocal* events, recording two or more points per syllable in this case does justice to the affective data representation.[2]

With the lombada technique, the data flow is reduced by a factor of 4000 from the 96000 Bytes per second of a DAT-quality speech recording to an average of 24 Bytes/s for loudness and pitch values and their positions collected at 4 peaks per second – a big benefit for real-time applications. The data refinement also takes one step of finding a suitable prosody representation away from the neural network. It seems plausible that this complexity reduction should ease the learning process. At these low data rates of 5–30 vectors per sentence, pattern recognition on these discrete f_0/loudness contours even seem possible again. Hidden Markov models because of their warp invariance (see Rabiner 1989), and *convolution networks* (LeCun et al. 1990) that can find functional elements at different parts of a contour (shift invariance), as used in handwritten character recognition, are prime candidates for future research in this area.

---

[1] as $s(t) = \sin(2\pi \int_0^t f_0 \, \mathrm{d}\tau)$

[2] If one is indeed interested in finding syllables in a *linguistic* sense as opposed to phonetic units, a slightly different algorithm was already described by Paul Mermelstein in 1975 that correctly separated linguistic syllables in more than 90% of 400 cases. The goal in this work is different, but it was the observation how well loudness maxima and linguistic syllables coincide that led to the lombada idea. Extracting voiced loudness points that are global maxima in a 68 ms (4 frames) segment as done here automatically picks linguistic syllables ca. 80% of the time.

4.4.3 Using Evaluator Agreement to Adapt the Network's Learning Rate

The multilayer perceptron learns by examples of acoustic parameters–affective evaluation pairs. In regular backpropagation training, however, the network is not able to see how "trustworthy" such a data point is. It cannot tell if all 20 evaluators unanimously decided 'this recording sounds happy' or if the responses were distributed over the whole scale. Since all examples are treated the same way, the pattern recognizer would be led astray by recordings were listeners didn't agree.

To remedy this dilemma, the MLP in this work was programmed to receive information about the quality of the example it is trained with. This was chosen to be done through the backpropagation learning rate α of Eq. 4.8 which here becomes a function of the evaluation histogram's mode height h:

$$\alpha(h) = \begin{cases} 0.01 & \text{for } h > 10 \\ 10^{\frac{h}{2}-7} & \text{otherwise} \end{cases} \qquad 4.9$$

This way, for all recordings where listener's agreement yielded a peak height greater than 10, the points are treated the same. For smaller peaks, the influence of these recordings on the weight updating decays quickly as the peak height decreases.

The same alpha-weighting is also used in the network performance measurement, see Section 4.5.1.

4.4.4 Network Training

Once the network structure is set up, the internal state of the network is initialized by setting the weights to small values in a gaussian distribution with mean 0. The 18 dsp-parameters and the evaluation data were previously written to a file where the entries in every dimension were linearly transformed to fall, approximately, in a range of 0–1 or –1 through +1. Before the training starts, this data is read into RAM. The training for each affective category is done separately, in different networks, because the evaluators' agreement and therefore the learning rate α (a global network feature) differs between the affective categories for the same recording. The speech data is split into three subsets: a training set from which the backprop examples are chosen, a runtime test set which is used to compute the training error, i.e. evaluate the progress in performance during the training, and a final, hold out test set that is left untouched until the learning process is finished, and is used to assess the network's generalization ability. The training set is the biggest one and contains 80% of all points, the test sets each 10%.

The training is organized in cycles. During each cycle, 10 times the number of all datapoints, i.e. 1500, examples are backpropagated through the MLP. The acoustic/dsp parameters are entered as input, the corresponding psycholinguistic evaluation result given as the true output. The points are picked at random from the training set. After one cycle, each point in the runtime test set is sent through the net and the mean squared error computed from the deviation of the true value (the listener data affective value) and the network's estimate, the MLP output. The network is given epochs of at most 100 cycles to lower the training error. If in these 100 cycles the training error was never lower than at the beginning of the epoch, the training is stopped. In the second case, if the network manages to find a set of weights that results in a training error that is lower than the one at the beginning of the epoch, the weights and general network configuration is stored in a file as a potential best MLP, the current epoch is stopped, and the network gets another maximum of 100 attempts in a new epoch. Figure 4.4 shows how the error develops during training.

Two types of performance values are computed: the mean squared error (MSE) as an internal measure of the learning progress, and the number of points correctly classified by the

net, see Fig. 4.4. The network output is counted as a correct response if it is placed into a 2-MAD interval around the true value of the 20 listeners' responses for that recording. By this criterion, the MLP has to get closer to the histogram max-mode position than 50% of the human evaluators to score a hit (on average). Both the MSE- and the 2-MAD values are weighted with the learning rate since it describes listener's agreement and how strong the datapoint's influence on the learning would have been.

To remove a bias that might result from the arbitrary selection of the training/runtime error/final error data sets, the complete process is executed 20 times with reshuffled sets.

4.5 Results

The 7 affective categories are clearly separated into 2 groups: pattern recognition either unmistakably works with that impression, or it doesn't at all. There are no dubious cases where one might benevolently interpret the network's performance as poor but existent generalization. The learning progress can be outlined by tracking the training error as done for Figure 4.4. For the 4 categories that the MLP was able to learn, the accuracy improved as outlined in the left graph. These impressions are strong, agitated, leadership, and confident. The remaining three, pleasant, happy, and angry, show an error contour like the center graph in Fig. 4.4.

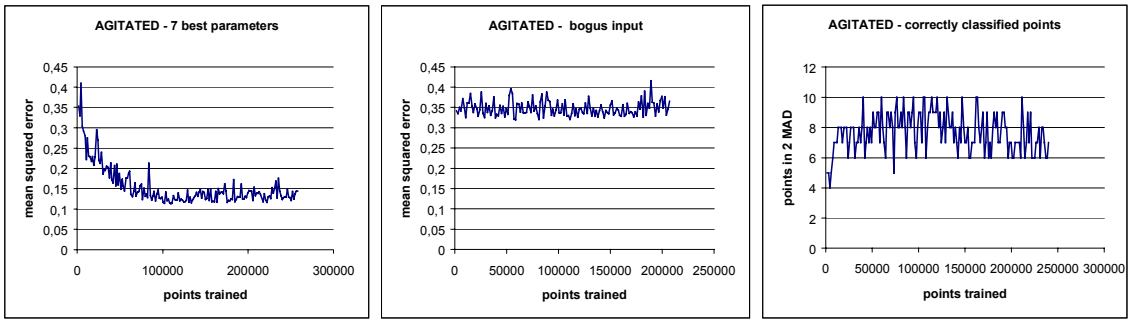


Figure 4.4 Training progress example for the agitated category. The left graph displays the decreasing mean squared error during training on the acoustic and evaluation data. In the center graph, the same network was supplied with the evaluation data, but the acoustic parameters were substituted by fake, random input. The curve on the right shows how many out of the 15 test-set points were estimated correctly (placed in the 2-MAD interval, here without α-weighting; training on proper acoustic data). The x axis indicates how many data points were backpropagated through the multilayer perceptron.

The failure of the pattern recognition in the categories pleasant, happy, and angry is not particularly discouraging. The impression of pleasance is an entirely subjective one and might depend on the listener's sex, age, and personal liking, thus it might be impossible to derive and learn one value that represents the scores of all evaluators. When the speech database was recorded, it was not planned to evaluate the samples on a 'happy' and 'angry' scale. These were added at a later time because of the interest in basic emotions. The situations that were interpreted by the actors were not intended to show strong happiness or anger, therefore, the database's variance in these categories is low. One of the evaluators wrote in the comments field of the evaluation webpage that the speakers "were neither happy nor angry". In fact, the three categories that the network was unable to learn are the ones with the three lowest average absolute deviations, see Table 2.1, p.18.

In the visual domain, Matt Dailey, Gary Cottrell, and Ralph Adolphs (2000) achieved classification behavior comparable to human subjects when modeling the perception of facial expressions corresponding to Ekman's 6 basic emotions from still images – using a very simple six-unit, one-layer neural network classifier on top of a Gabor filter lattice feature extraction. It would be surprising to find that the two basic emotions happiness and anger are coded in a much more complex manner in speech. In a speech data set with higher happiness variance, an indicator of happiness might be as simple as the spectral characteristics of smiling, such as the attenuation of the second formant (Tartter and Brown 1994).

The good classification ability in the categories strong, agitated, leadership, and confident promises usefulness for future applications.

4.5.1 Quantifying Generalization Ability

It is hard to give an intuitive performance measure for the pattern recognition, since, as opposed to forced-choice classification, this recognizer operates on a continuous scale, and a mean squared error value doesn't allow much insight. Requiring the classifier to place its estimate inside a 2-MAD interval around the maximum mode to score a hit, as is done here, is a fairly strict criterion, and the network usually manages to place 4-8 out of 15 recordings (not weighted with the learning rate α, depending on the quality of the data in the subset) of the previously unseen final test set into the correct range for the 4 categories that were learned. This is comparable to the agreement of the human subjects and surprisingly good considering that only a small fraction[1] out of these 15 points has the maximum learning rate associated with it (see Eq. 4.9), and therefore the majority of this would have had a low influence on network training.

The performance measure p plotted in the graphs of Section 4.5.2 describe how many out of the n points x_i in the untouched test set were correctly placed into the 2-MAD interval centered at the mode. Since the training is weighted by the learning rate α (see Sect. 4.4.3), the same is done in the measurement of the network's generalization ability. With a classification measure $c(x_i) = 1$ for estimates correctly placed into the point's 2MAD interval $\Delta_{2\mathrm{MAD}_i}$, and $c(x_i) = 0$ otherwise, the weighted absolute performance measure p^* can be noted as

$$p^* = \frac{\sum_{i=1}^{n} \alpha_i c(x_i)}{\sum_{i=1}^{n} \alpha_i} = \frac{\sum_{\text{correct}} \alpha_i}{\sum_{i=1}^{n} \alpha_i} \quad . \tag{4.10}$$

To show how much better than chance the network performs, this measure is divided by the value p^*_{chance} which one would achieve by randomly placing points into the interval of possible answers – the scale width Δ_s described below – to give the relative performance measure p used in the following graphs

$$p = \frac{p^*}{p^*_{\text{chance}}} = \frac{\sum_{\text{correct}} \alpha_i}{\sum_{i=1}^{n} \alpha_i} \Bigg/ \sum_{i=1}^{n} \frac{\Delta_{2\mathrm{MAD}_i}}{\Delta_s} \quad . \tag{4.11}$$

p is averaged over 20 runs, see below, to make sure it does not only represent one specific test set but the whole data.

[1] Only 56 out of the 600 points in these 4 categories have a histogram peak height greater than 10 and therefore the maximum learning rate of 0.01 associated with it, see Figure 2.5.

To find a good number of input features, parameters are added sequentially starting with one and testing all the way through to a complete 18 dimensional input layer. For the first feature, an MLP with one input neuron is created and trained with the first dsp-parameter as input. After the performance with this measure has been determined, another network with one input neuron is created that tries to map the second parameter to the affective output. Again, the performance is tested, and the process is repeated with the next acoustic parameter until the network has examined the usefulness of each measure individually. The one of the 18 parameters with the highest classification ability is fixed as the first input value. In Figure 4.5a below, the network functioned best with parameter 'loudness at loudest voiced point'. The next goal is to find the one of the 17 remaining features that constitutes the best 2-input-neuron network together with the first parameter. For the *pleasant* category below, 'speechrate' was added as the second parameter. Then, 16 three-input-neuron MLPs are set up to find the measure that complements the first two parameters in the best way, and so forth until the performance of all networks, with all input sizes from 1–18, has been determined. Except for the size of the input layer, which is naturally determined by the number of input parameters, the setup is kept constant with one hidden layer of 6 units and one output neuron.

It is tempting to think of the sequence in which the parameters were chosen as the order of importance for classification. However, the network's classification ability and its measure depends to such a strong degree on the random distribution of points in the training and test sets and on the order in which they are backpropagated through the network, that the performance fluctuation can be as strong as 50% for the same affective category and network layout. Even after shuffling the distributions and starting over 20 times for each setup[1], one cannot exactly derive which parameter is the most important one to classify a category; although it frequently occurred that a feature which was among the first 5, say, to be picked ended up in the best 5 again after shuffling and restarting.

The measure p in the following 7 graphs indicates how much better than chance the network performs. As noted above, *chance level* is defined as the probability that a point, placed at random within the bounds of the evaluation scale, falls into the 2-MAD interval in which a response is scored as correct (i.e., the ratio of evaluation-2-MAD size for a particular datapoint, and the size of the whole evaluation scale in that category). The scale width describes the total interval in which evaluation scores could have been placed and is determined by the normalization of the psycholinguistic evaluation data, see Sect. 2.3.3. The rating scale during the evaluation ranges from –2 through +2. During the postprocessing of the scores, the average absolute deviations d_i in each category are set to 0.5 for each listener (by dividing through the average absolute deviation and multiplying by one half, therefore, the answers at –2 would be moved to $-1/d_i$, +2 to $+1/d_i$.) Hence, if a category didn't have much deviation from the 0-baseline, the scale would have been expanded more than in a category with higher variance. The second normalization performed in the postprocessing is the adjustment of means. The more the listeners' average responses $\bar{x}$ for one impression varied, the farther the limits of the overall scale are pushed. This scale width is therefore not only needed to compute the chance level, it is also a useful (inverse) indicator of evaluation variance and agreement in a category.

[1] The computational effort for this incremental parameter search in every affective category included $18+17+...+1=19\times9$ network runs, each with 20 restarts/shuffling, at an average of 300000 backpropagation runs for each network training, yielding an estimated 10^9 backpropagated points. This took about 12 hours of computation for every category on a contemporary pc.

The lower and upper bounds of the scale, b_l and b_u, respectively, are given by

$$b_l = \min_{i=1\ldots20 \text{ evaluations}} \left(\frac{-1}{d_i} - \bar{x}_i \right) \qquad b_u = \max_{i=1\ldots20 \text{ evaluations}} \left(\frac{1}{d_i} - \bar{x}_i \right) \qquad 4.12$$

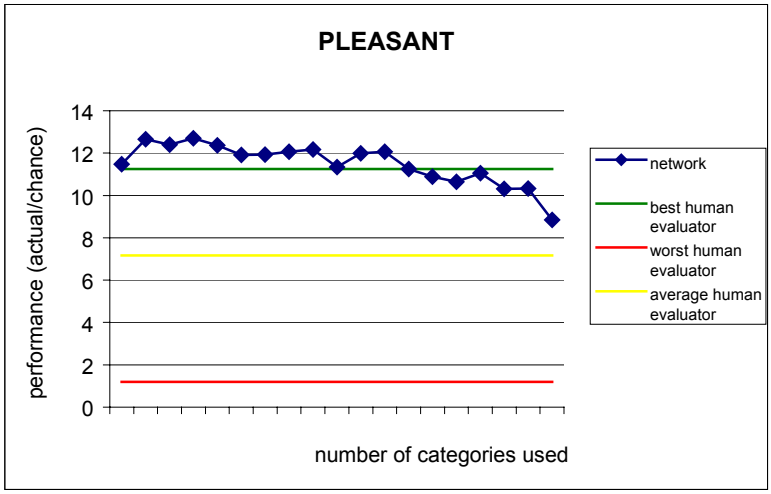
with the average absolute deviation $d_i = \overline{|x - \bar{x}|}$.
This yields the category's scale width Δ_s

$$\Delta_s = b_u - b_l \qquad 4.13$$

To show how much the human evaluators agreed and how the performance of the neural network classifier compares to the human subjects', the same measure which is used for the multilayer perceptron is applied to each of the 20 evaluations. Here, p was measured as how many out of all 150 responses were placed into the 2MAD interval for each point. Again, the performance is α-weighted.

Even if the pattern recognizer does not learn the affective content represented by the acoustic parameters, its responses still seemingly rate well above chance level because the listeners' scores might have only fallen into a small interval on the relatively large scale, and the network is able to model the distribution of points and randomly output points that fall into the area of highest density (the evaluation responses are not explicitly whitened over the whole response scale). As a reference, to show how well the network models the output data distribution if provided with meaningless input, the very first points on the following 8 graphs describe the network's performance when provided with fake, random data instead of an acoustic parameter.

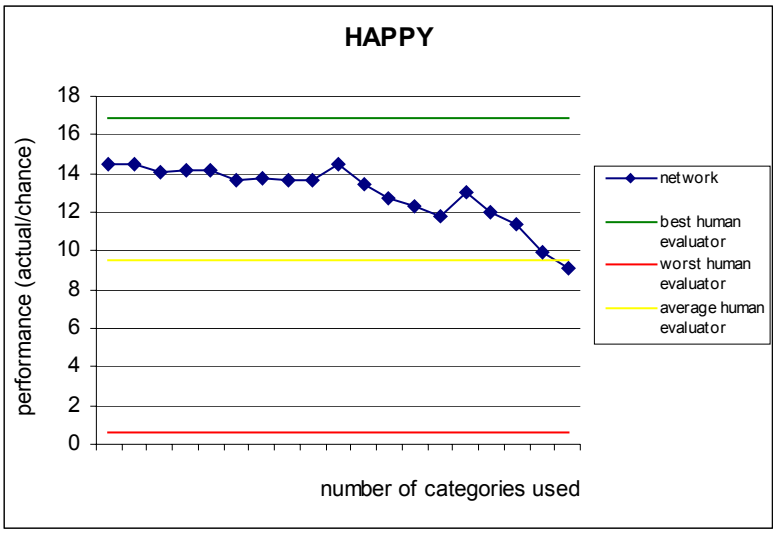
4.5.2 Incremental Parameter Selection Results



scale width: 5.92

1. loudness at loudest voiced point
2. speechrate
3. loudness at last voiced point
4. median f_0 change
5. median loudness change
6. median f_0
7. f_0 at loudest voiced point
8. sex
9. Δf_0 loudest-last voiced point
10. spectrum 1–2 kHz
11. median loudness
12. spectrum 2–4 kHz
13. spectrum 0–0 . 5 kHz
14. spectrum 4–8 kHz
15. maximum loudness
16. spectrum 0.5–1 kHz
17. f_0 at last voiced point
18. correlation pitch–loudness

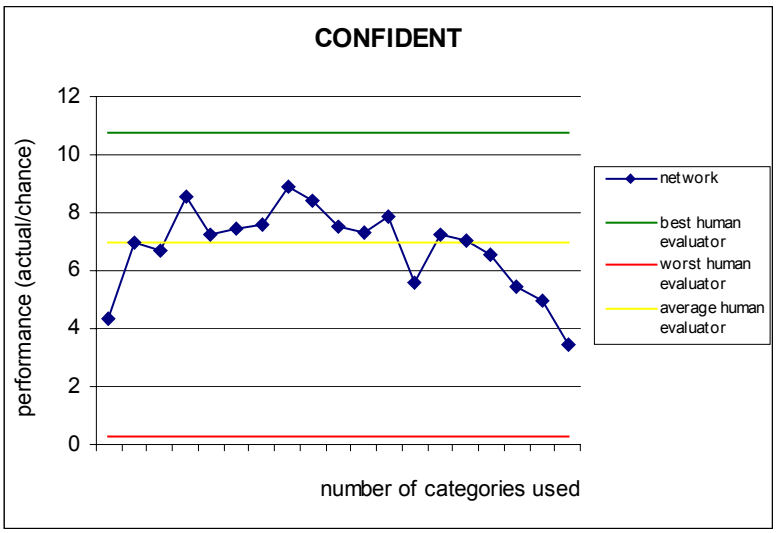
Figure 4.5a Performance vs. input size for the *pleasant* category (not learned). Since it has a large scale (the third widest in these experiments), and probably evaluation scores clustered in a smaller interval, the MLP puts its responses in the area of maximum likelihood to model the output distribution. A look at the individual error graphs however confirms that no learning is taking place, the contours resemble the center graph in Fig. 4.4.



1. speechrate
2. median loudness
3. f_0 at last voiced point
4. f_0 at loudest voiced point
5. sex
6. spectrum 2–4 kHz
7. spectrum 0–0 . 5 kHz
8. maximum loudness
9. spectrum 0.5–1 kHz
10. correlation pitch–loudness
11. median f_0
12. Δf_0 loudest-last voiced point
13. loudness at last voiced point
14. spectrum 1–2 kHz
15. spectrum 4–8 kHz
16. median f_0 change
17. loudness at loudest voiced point
18. median loudness change

scale width: 6.07

Figure 4.5b Performance vs. input size for the *happy* category (not learned). This impression shows the worst generalization ability, lowest average absolute deviation, and second widest scale of all categories[1]. As the error contours confirm, no learning takes place.

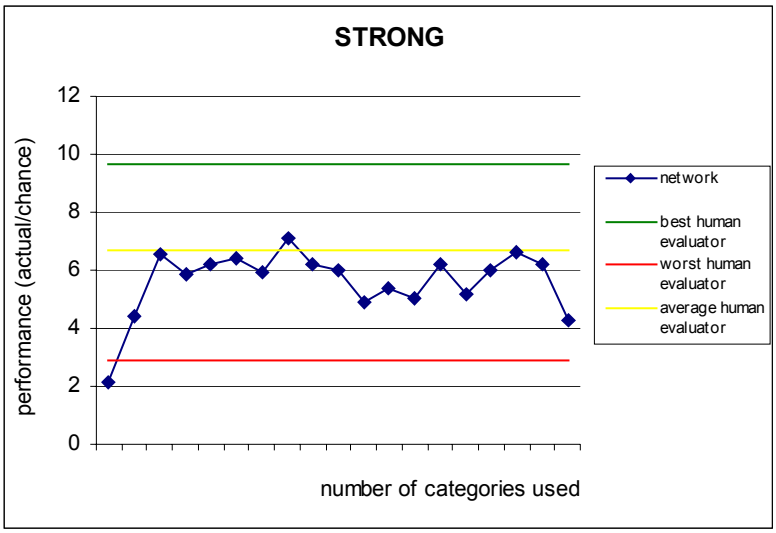


1. loudness at loudest voiced point
2. median f_0 change
3. spectrum 2–4 kHz
4. maximum loudness
5. loudness at last voiced point
6. median loudness
7. spectrum 0.5–1 kHz
8. correlation pitch–loudness
9. spectrum 4–8 kHz
10. median f_0
11. spectrum 0–0 . 5 kHz
12. spectrum 1–2 kHz
13. median loudness change
14. f_0 at last voiced point
15. Δf_0 loudest-last voiced point
16. f_0 at loudest voiced point
17. sex
18. speechrate

scale width: 4.87

Figure 4.5c Performance vs. input size for the *confident* category (learned). This is one of the two higher-level impressions requiring a relatively high degree of interpretation on the listener's side. The graph indicates that the growing complexity inherent with increasing input dimensionality is detrimental to generalization ability. The performance could probably be improved with a more custom-made setup than the generic x-6-1 MLP, for instance with more hidden level neurons and possibly a smaller learning rate.

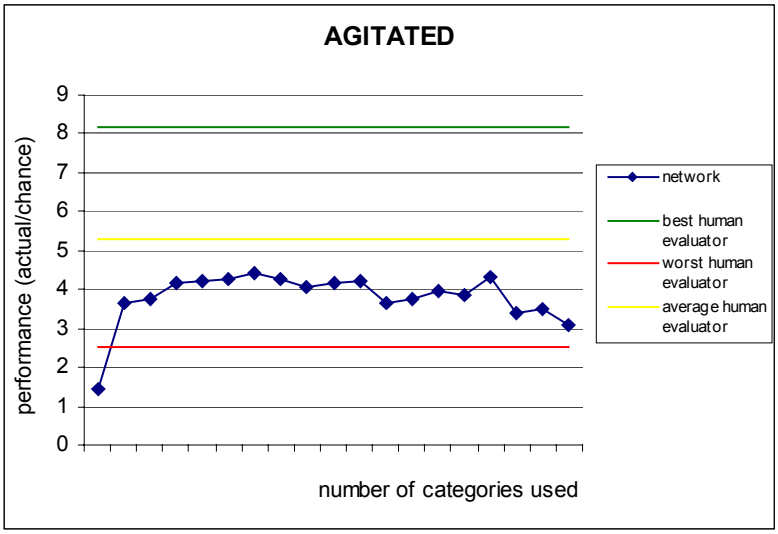
[1] Incidentally, 'happy' was also the first category on which the neural network was tested, so a considerable effort was undertaken to set up a functioning system. It is unlikely that further parameter tweaking would have any positive effect.



1. spectrum 2–4 kHz
2. spectrum 1–2 kHz
3. spectrum 0.5–1 kHz
4. correlation pitch–loudness
5. loudness at loudest voiced point
6. median loudness
7. f_0 at last voiced point
8. spectrum 4–8 kHz
9. median f_0 change
10. median f_0
11. median loudness change
12. spectrum 0–0 . 5 kHz
13. speechrate
14. maximum loudness
15. f_0 at loudest voiced point
16. sex
17. Δf_0 loudest-last voiced point
18. loudness at last voiced point

scale width: 5.48

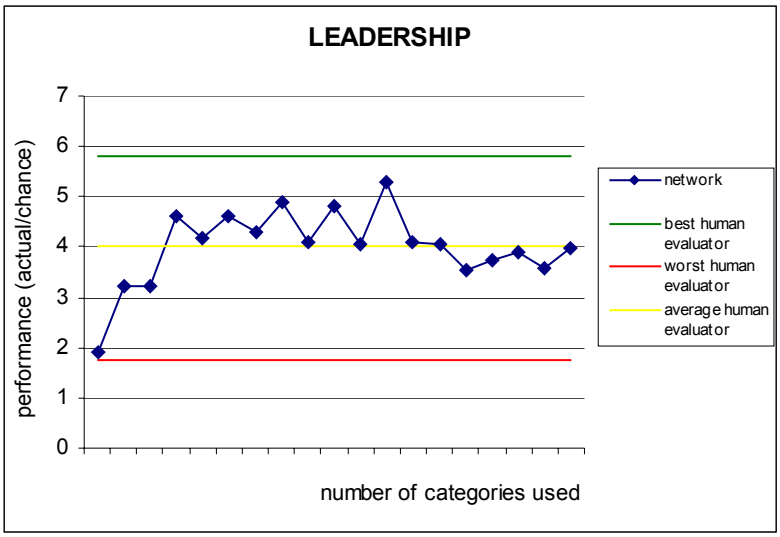
Figure 4.5d Performance vs. input size for the *strong* category (learned). This impression yielded the best results. At the 7-input-neuron network configuration, the performance is 3.3 times higher than for the random-input run. The first values suggest that spectral values, i.e. timbre, is a good indicator of the listener's impression of physical strength.



1. spectrum 4–8 kHz
2. sex
3. Δf_0 loudest-last voiced point
4. loudness at last voiced point
5. spectrum 0.5–1 kHz
6. f_0 at loudest voiced point
7. median f_0
8. f_0 at last voiced point
9. correlation pitch–loudness
10. spectrum 0–0 . 5 kHz
11. spectrum 1–2 kHz
12. spectrum 2–4 kHz
13. loudness at loudest voiced point
14. maximum loudness
15. median loudness
16. median f_0 change
17. median loudness change
18. speechrate

scale width: 3.61

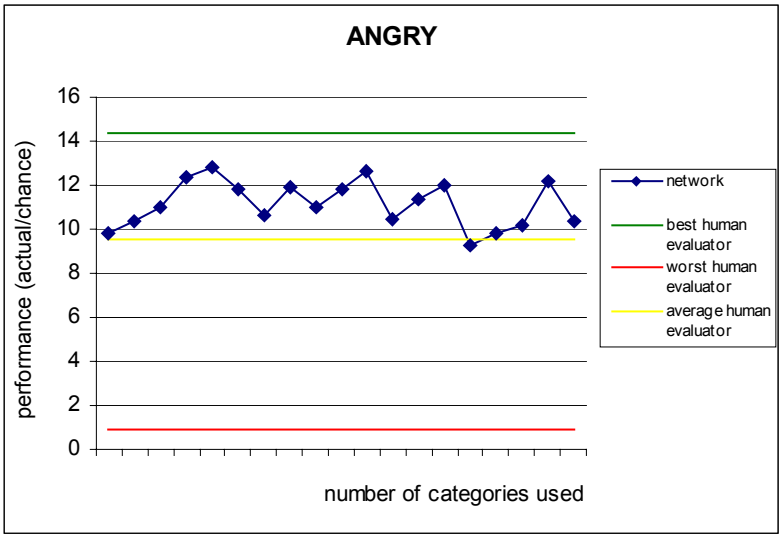
Figure 4.5e Performance vs. input size for the *agitated* category (learned). The second best category.



1. median f_0
2. spectrum 0.5–1 kHz
3. spectrum 0–0 . 5 kHz
4. f_0 at last voiced point
5. loudness at loudest voiced point
6. median loudness
7. Δf_0 loudest-last voiced point
8. sex
9. median loudness change
10. loudness at last voiced point
11. maximum loudness
12. spectrum 4–8 kHz
13. spectrum 2–4 kHz
14. f_0 at loudest voiced point
15. correlation pitch–loudness
16. spectrum 1–2 kHz
17. speechrate
18. median f_0 change

scale width: 3.25

Figure 4.5f Performance vs. input size for the *leadership* category (learned). The second higher-level category the network managed to learn.



1. median loudness
2. median loudness change
3. median f_0
4. spectrum 4–8 kHz
5. sex
6. loudness at loudest voiced point
7. spectrum 1–2 kHz
8. spectrum 2–4 kHz
9. speechrate
10. median f_0 change
11. loudness at last voiced point
12. spectrum 0–0 . 5 kHz
13. Δf_0 loudest-last voiced point
14. spectrum 0.5–1 kHz
15. maximum loudness
16. f_0 at last voiced point
17. correlation pitch–loudness
18. f_0 at loudest voiced point

scale width: 7.91

Figure 4.5g Performance vs. input size for the angry category (not learned). The evaluation for this impression yielded the highest scale width but also higher variance than *pleasant* and *happy*, so the psycholinguistic data has probably the most incoherent data of all categories.

4.5.3 All Categories Compared

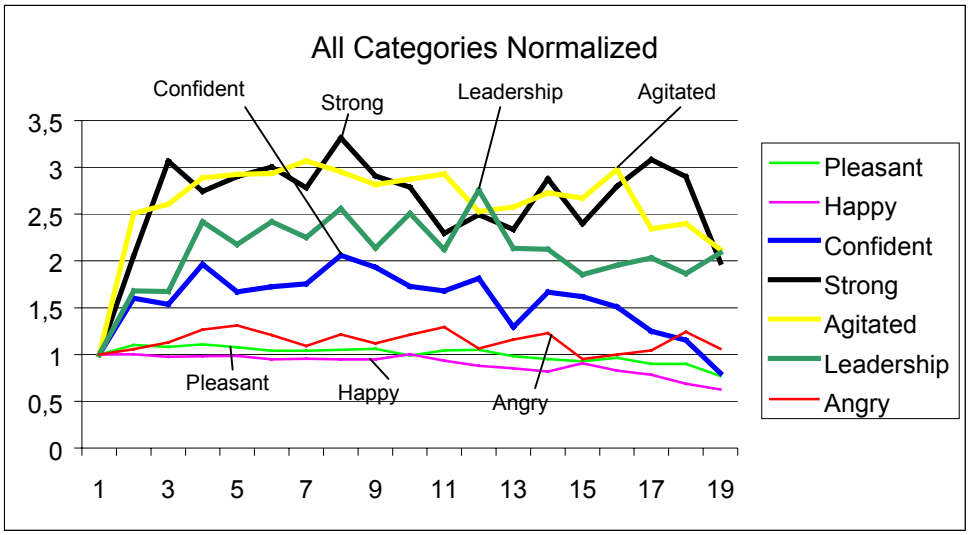
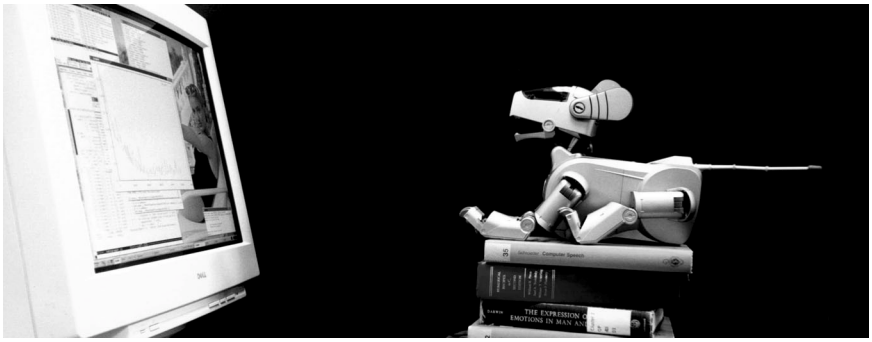


Figure 4.6 Generalization ability in all categories, normalized with respect to the performance of the random-input network

The contours above outline how well the networks were able to model the individual impressions. The results from graphs 4.5a–g were divided by each network's performance with the fake, random input (i.e., divided by the performance when the networks learned to model the data distributions, but not the mappings from input to output). The low-level impressions, *strong* and *agitated*, are learned well, followed by the higher-level categories *leadership* and *confident*. The network – with one hidden layer of 6 neurons, scalar output, and input dimensionality depending on the number of parameters – does best in the 7-12 input neuron range.

5 An Application: A Nonverbal Speech Interface for a Robot Dialogue System



A computer terminal is not some clunky old television with a typewriter in front of it. It is an interface where the mind and body can connect with the universe and move bits of it about.

Douglas Adams,
Mostly Harmless

...and one day in the future, they will most likely be true interfaces indeed. At this point however, human-computer interaction, for the most part limited to the use of a "typewriter, a clunky television tube," and a mouse, is fairly clumsy. A robust speech dialogue system would be a major advancement towards a natural interface. Progress in the field of multimodal speech processing and language recognition justifies hope we will be able to build these in the next decades.

As pointed out in the introduction, human speech recognition works in a multimodal manner and on multiple interacting levels. We combine a deep knowledge about grammar, the meaning of previously said words, the speaker, the culture, and so forth with the heard sounds, the prosody, and gestures to parse a speech stream and extract meaning. If one took away a listener's knowledge about meaning, grammar, and use of prosody of a language – say by having a person spot words from a dictionary in a language the listener doesn't understand – experience shows the ability to transcribe speech to text falls short of a native speaker's. Current computer speech recognition (or, to be more precise, word recognition) systems for

the most part suffer from the same impediment, they only exploit the verbal channel. As a consequence, both their performance and functionality as natural human–computer interfaces are strongly limited. If a computer wanted to interact with a user, it seems reasonable it would make use of the many modalities that human communication offers, such as the *linguistic* (verbal) data but also information about the speaker's state, if he was serious or joking, emotionally agitated or calm, etc. A good speech recognition system could use the affective message of an utterance to assign probabilities to words in cases where the right selection based on acoustic clues is difficult. For instance, if a software is to decide between the words "fun" or "gun" and the nonverbal content signals a hectic atmosphere, the latter candidate appears to be the rational choice.

The use of prosody becomes even more crucial in *meaning* recognition. In irony, for example, the words usually carry the opposite message of their literal content. If a human–computer speech dialogue system assesses the machine's behavior considering only the verbal information of a user's contemptuous "Well done computer, you just deleted my most important file!", this clearly does not capture the intended message.

Motivated by the artificial intelligence features of SONY's pet robot Aibo, the findings described in the last chapter were applied to see if the data representation and pattern recognition can be used in a speech dialogue interface. Aibo's actions are steered by a behavioral model that develops over time as the dog interacts with the environment. To do this, he uses two microphones, a color video camera, an infrared edge finder, and various tactile sensors.

To interact with the dog, the owner gives commands via a remote control that plays a sequence of tones understood by Aibo. In a feasibility study for a more natural nonverbal speech interface, it was investigated if vocal utterances can be classified as prohibition or positive reinforcement by prosody only. Thus, in this case, *ex*pression is modeled rather than *im*pression.

Malcolm Slaney (1998) built the pattern recognition system *Baby Ears* that placed utterances of parents to their infants in one of the three categories approval, attention, and prohibition. He achieved classification rates as high as 65% for the best 318 (based on listener's agreement) out of 500 utterances. The parallels in both tasks are striking: in both cases, the addressee doesn't understand the linguistic, verbal content of the utterance, and the message must be brought across by prosody. Slaney used mostly pitch values, also mel-frequency cepstral coefficients (cepstrum on a psychoacoustic (Bark) frequency scale) and energy variance.

For the Aibo study, utterances in the two categories prohibition and encouragement are recorded from 11 people of 5 nationalities (U.S. American, Spanish, Norwegian, French, and German). All speakers were asked to produce 4 sentences, one given sentence each for the prohibition and the positive reinforcement categories, and additionally one sentence that they were asked to extemporize for each category. Some used more sentences, so the total number of recordings is 69 (36 prohibition, 33 encouragement). The recordings were made at UCSD's Center for Research in Language. Since this was a feasibility study, not an implementation task, the speakers were recorded through a regular microphone, not through Aibo's sensors.

The data is nicely separable with the dsp-parameters introduced in Chapter 3. For all speakers and utterances, the network achieves a recognition rate of 78.4% (reshuffled 40 times). For the recordings of each single user, the data can be well classified just by discriminating in the three dimensions maximum perceived loudness (95.7% correct), median loudness (88.4%), or median f_0 (81.2). Network training was not attempted because of the small number of datapoints for a single speaker.

Advances in robot communication as seen by Mike Peters

6 Discussion, Conclusions, and Outlook

The results of the pattern recognition experiments nourish confidence that the framework presented here has potential for application in real life tasks.

6.1 The Database

The first goal as stated in the introduction, to create a versatile database that contains a number of natural sounding speech recordings that can be used to generate a variety of different impressions in a listener, as well as the psycholinguistic, affective evaluation of the recordings, has worked well. Even though the semantic differential evaluation, as opposed to forced-choice, adds complexity to the judging, the collected data proved meaningful in the pattern recognition. In the three of the seven classes that didn't allow modeling of the percept, namely pleasant, happy, and angry, this failure was hinted by the low average absolute deviations for these impressions.

Collecting the data over the internet has also shown to work well. It is arguable whether it is more appropriate to run all evaluation in the same controlled conditions as done with the Californian listeners, or whether it is beneficial to include a variety of different environments by allowing users to do the experiment wherever they want, thereby possibly eliminating any biases from uniform conditions. Since the agreement in German scores appear comparable to that of the American listeners, none of these methods should be ruled out.

The database contains a vast stock of information that can be analyzed beyond the use in the present experiments, e.g. if the age of the listener influences personal liking, or if the sex of the listener and the evaluator biases perception of affect, as Aronovitch found (1976). It can be investigated if the receiving accuracy changes due to fatigue during the course of the 1.5 h evaluation. Or if it makes a difference whether the evaluator had had language training in German or not. Another useful investigation would be an analysis of variance between the databases to see how the German and the Californian scores compare, and of course pattern recognition also on the German evaluation data. It would also be interesting to see if the pattern recognition accuracy improves or decreases if, instead of two values representing 20 evaluators, the scores of a single person are used. Especially for very subjective categories, like *pleasant*, it might be more meaningful to model a single liking than an average one which possibly doesn't exist.

6.2 Acoustic Parameters and Data Representation

The biggest part of time of this thesis work was spent on the second goal as stated in Section 1.2, to extract dsp parameters that possibly carry the nonverbal information in the speech samples and lead to a memory efficient data representation. With the exception of the FFT[1], no preprogrammed software was used; all code runs in real time and was written in C++ for this work with the goal of robustness and automation in mind. The pitch tracker developed for this work gains its strength from combining two elementary pitch detectors, and the biggest benefit comes from including system history knowledge, which makes the tracker fairly robust without compromising accuracy.

The new loudness model has proved useful in the recognition of affect. The main benefit is that one is able to supply pattern recognition machines with a straightforward value that can describe the loudness baseline of utterances – whether a person speaks loudly or softly – a task values like power or energy are unable to perform. Glottal pulse shape examination as

[1] The fast Fourier transform was taken from Numerical Recipes in C.

pointed out in Sect. 3.4.2 yields related information, namely about vocal effort, but that parameter is only a description for the sender's intensity, describing the expression, rather the impression that the perceiving listener has. Speakers exercising the same vocal effort aren't in general perceived as speaking with the same loudness since the influence of the vocal tract is not accounted for, neither are the different attenuations and excitation thresholds at different speakers' frequency ranges. Another advantage of the model presented here is that it yields accurate relative loudness values for one recording. Speech loudness variance and fluctuation are parameters that are used in nearly all studies on the statistics of nonverbal speech, but power contours and absolute loudness contours are not linearly related to each other, so values other than absolute loudness are an imprecise description for relative parameters such as variance and fluctuation if they claim to model perception. The discrepancy is especially noticeable for unvoiced sounds that have most of their energy distributed in higher frequency regions where the human ear is less sensitive. Thus, their perceived loudness is smaller. Compare for instance the affricates' and sibilants' intensities in Figure 3.3 to those of the periodic, voiced sounds. Both classes seem to carry similar energy values, but the ear's attenuation in the high-frequency bands of the noise-like consonants is lower, and therefore they are perceived as softer. As a practical byproduct of this, virtually all loudness peaks now correspond to voiced sounds, usually with one maximum per syllable. This allows to attribute a high prior probability to the voiced case for a voiced/unvoiced classification system, adding to the precision of the classifier.

This psychoacoustic model – since it computes specific loudness values for each frequency – has the advantage of a much finer frequency resolution over the Zwicker-ISO model which works in fixed critical bands. In the latter scheme, the audible spectrum is divided into 24 frequency bins corresponding to critical bands. The loudest value in each bin then determines the frequency group loudness level for its interval. For example, two sounds with the same excitation level that are, say, 0.2 Bark apart on the frequency scale, would fall into the same band in the ISO model if they had frequencies 2.7 and 2.9 Bark. Therefore, the overall loudness would be the same as the loudness for just one sound. Now consider the same sounds with their frequencies shifted to 2.9 and 3.1 Bark. In this case, the core loudnesses would fall into two different bins and the levels would add up. Clearly, in this case, reality is not accurately represented. In the absolute-loudness model, both instances would more appropriately be treated in the same manner, with a masking cone that has a plateau width of 1.2 Bark. For a finer frequency resolution, a bank of (more than 24) gamma filters can be used, but the center frequencies would still be fixed in this case, as opposed to the dynamic placing of masking cones at each peak in the Bark spectrum.

As the glottal pulse steepness-of-decay data shows, the glottal pulse source spectrum for high vocal effort utterances exhibits an augmentation of higher frequencies. Therefore, if the vocal tract transfer function remains somewhat constant or at least doesn't counteract this effect, one would expect the absolute loudness model to work for all voiced sounds. As confirmed by these experiments, the normalization works well for an intensity range from soft to loud talk. For whispering, i.e. unvoiced speech with a flatter spectrum distributed mostly in higher frequencies, one would expect the normalization to yield no values or err on the low side, and therefore make the overall loudness stronger than perceived. This is the case for three of the 5 whispered recordings. In the other two, the normalization process still yielded a value that correctly placed the loudness below that of the low volume voiced recording; possibly due to very brief (accidentally) voiced parts that were sufficient to correctly calibrate the normalization constant. For screaming, not only do the high frequencies get a stronger boost than lower frequencies, but the intensity in all frequencies is increased. Therefore, the normalization is unable to grasp the actual loudness and misclassifies the intensity to a level below that of loud talk.

Future work could determine if a combination of a speaking-related model returning vocal effort data as computed by inverse filtering and glottal pulse shape in conjunction with this hearing-based loudness model is able to also grasp the screaming case. An additional quantity that can give information on vocal effort and should be investigated in this context is the duty factor, the ratio of glottal pulse to total fundamental period length, which varies from 0.3–0.7 depending on vocal effort and the type of register used.

Another approach that can be tested is not only normalizing with regard to loudest frequency, but also to the level of unvoiced sounds whose intensities do not change as dramatically for higher vocal effort. However, this would work only on clean speech signals since it is almost impossible to separate random noise from unvoiced speech which lacks the periodic structure of voiced utterances.

Higher vocal effort usually coincides with higher breathiness, therefore the glottal-to-noise excitation ratio should also be investigated if it is useful for normalization.

The normalization, which our ear performs in a fraction of a second, would probably improve substantially if the phoneme of the utterance was also known. Glave and Rietveld (1975) report a difference of 5.2 dB for the average loudness level as judged by 5 listeners who assessed recordings of the vowels [i], [u], and [ɛ], all normalized to the same rms levels. In the present work, collecting intensities over the course of 2.5 seconds of voiced frames yields 10–15 voiced phonemes to compute an average. If the phonemic information was presented to the normalizer, an even higher accuracy of average loudness estimation could probably be achieved in a much shorter time.

The normalization essentially sets the loudest frequencies equal and in the further process loudness is judged by the width of the frequency band that contains the formants, below the fundamental frequency range. Of course, the intensity of the fundamental frequency band does not stay constant, but is also augmented for higher vocal effort. Thus, in order to not only have a loudness scale that is monotonically correct, it is desirable to have a linear loudness scale, where a sound that is perceived twice as loud as a reference receives a measure twice as high (like in the transition from the *phon* to the *Bark* unit). To do so, it must be investigated how the fundamental frequency band's intensity changes with regard to higher vocal effort. It is also necessary to find an absolute reference point to which to calibrate the scale. Finally, this scale will have to be compared to data from human listeners judging loudness for validation. This can be done by having listeners rank the recorded database in order of their loudnesses and comparing the order to results from the model.

The loudness model also leads to the lombada data representation. This efficient data representation has shown good performance, and as the listening tests in (Quast 1999) suggest, most likely preserves the vast majority of audible prosody.

Besides the acoustic parameters used here, a list of other features could be derived in future work to describe the stimulus, for instance the dimensions of the Göttingen Hoarseness Graph that is used to describe pathologic voices, see Michaelis, Fröhlich, and Strube (1998) Breathiness, for instance, given as the glottal to noisy excitation ratio, is a quantity that is hardly used in nonverbal vocal speech research. Information about the vocal tract could be obtained from the spectrum (Schroeder 1967; Strube 1999; Freienstein, Müller, Strube 1999) and included in the decision process. This task is related to speaker normalization and phoneme recognition, which would be a very useful addition to the information for the pattern recognizer: if the network knew what phoneme and word belongs to an utterance, it could tell whether a variation in vowel length is due to linguistic information (such as [u] and [u:] in the German K<u>u</u>ss and M<u>u</u>ße, respectively), or is due to the speaker's speechrate.

6.3 Pattern Recognition

The last goal was to perform pattern recognition on the psycholinguistic and the signal processing data to see if the nonverbal content of speech is indeed carried by the extracted dsp-parameters and can be learned by a pattern recognition scheme and compare favorably to the agreement of the human listeners. Recognition worked for the 4 categories confident, strong, agitated, and leadership, it didn't for pleasant, happy, and angry. As pointed out earlier, the failure to learn the latter three impressions should not lead to give up hope for these, plausible reasons for the failure are the low variance in the database and the very subjective nature of the pleasant category. One might speculate that the 4 functioning categories – which are most likely not orthogonal – measured an underlying impression, and the choice of parameters selected for training in Section 4.5.2 is not fundamentally different enough to rule out this possibility. This needs to be further investigated, possibly with analysis of variance.

Crucial for the learning success was the new adaptation of the learning rate method, that allowed to tell the network how much to "trust" a point. Without this information, learning didn't take place.

As the nonverbal speech interface feasibility study in Chapter 5 showed, the model is adaptable and can be applied in a straightforward manner. The whole speech interface study was programmed in a mere 4 hours.

Acknowledgements

Working in an interdisciplinary field that not only includes the physics of speech but also psycholinguistics, pattern recognition, social psychology, as well as digital signal processing, has been an exciting challenge that I gladly delved into. I am grateful to many people who supported me and made these thematic crossovers possible, especially to those who have helped me knowing their professional work would not be rewarded other than with a thank you.

I am fortunate to have Manfred Schroeder, who wrote the book on computer speech, as my thesis advisor. He gave me the opportunity to assist him with his writing work, a rewarding experience from which I gained many insights into speech processing that came to use in this thesis. Professor Schroeder's support opened many doors on both sides of the Ocean.

Terrence Sejnowski was so kind to invite me as a visiting scholar to the Institute for Neural Computation at UCSD. Terry's INC is the prime address for interdisciplinary research on neural data processing, and it has been a privilege to be a part of it.

My abode at UCSD is Javier Movellan's Machine Perception Lab, home of the robot dog and a vibrant origin of human-computer interaction research. Javier made me feel at home right away with his humorous and energetic manner, and so did all my other friends at the Machine Perception Lab, Marni, John, Jonathan, Tim, Etienne, Ian, Evan and Boris. Javier gave me the resources to do this work, many good counsel on statistics, and financed this work's evaluation experiment from his funds.

I am especially indebted to the actors of the Deutsches Theater in Göttingen, who lend me their voices free of charge to put together the speech database. Watching them impersonate a shy schoolchild and, a minute later, present an impressive interpretation of the country's chancellor has been one of the most fun parts of this work. I am also thankful to all the non-actors who let me record them for this database, and to the volunteers who took the time and evaluated the speech database.

My home at the University of Göttingen is in the Speech and Neural Network group led by Professor Schroeder and Dr. Hans Werner Strube. My thanks for their support goes to them as well as my friends Olaf, Matthias, Dirk, Jan, Knut, Heiko and Ioannis; especially to Olaf, who helped me on countless occasions, was my right hand in Göttingen while I was in San Diego, and also made valuable suggestions after proofreading this thesis; and to Matthias, who, ever since I joined the group, has always taken the time to give very useful ideas and critique. Matthias and Jan also did the speech evaluation, thanks!

Dr. Karl Lautscham and Wolfgang Ebrecht of our institute's electronics lab built a custom preamplifier for me that made the high quality audio recordings possible.

Many colleagues and friends helped me to produce excellent quality recordings in San Diego, namely Robert Buffington, Fred Dick, Jonathon Vance, Chris Mercer and Peter Otto.

Suitbert Ertel at the Georg-Elias-Müller Institut für Psychologie took the time on two afternoons to give me an excellent personal crash course on some very useful techniques in social and personality psychology, for instance on the semantic differential. Professor Ertel also gave me the pointer to Klaus Scherer.
Professor Scherer, the authority on nonverbal speech research, sent me a very useful literature list that led the direction in the beginning stages.

Among the people that shared their ideas and made valuable suggestions are Robert Hecht-Nielsen, Gary Cottrell, Bhaskar Rao, John Mullenix, Jeff Cohn, Gary Katz, and Te-Won Lee. Robert and Gary at UCSD taught the first courses on pattern recognition I heard and introduced me to the coolest thing I learned during my university level studies: neurocomputing.

Douglas Adams helped me find that nice quote of his on human-computer interaction again that I only vaguely remembered. So Long Douglas, and Thanks.

A thanks to my friends and scuba buddies, especially my good friend Dave Maley, and for good times shared on both sides of the ocean to my closest friends Kristin Hünerbein and Tina Bäse.

Jackie Ferretti has been with me on all ways in California and deserves credit for many good line of code in my software, especially in the evaluation program.

My biggest thanks goes to my family, my brother Thorsten, and my parents, Helmut and Heide Quast, who have supported most all my daring endeavors, including this one. Thorsten, a frequent visitor on San Diegan shores, has helped me in many ways, not least as financial adviser and computer hardware expert. My parents have supported me in by far more ways than could be mentioned here, especially during the last weeks of preparing this thesis. My family has been a constant source of energy throughout the past 27 years, and I would not have been able to get where I am now without them.
To all of them goes my heartfelt gratitude.

References

Adams, D.N.: *Mostly Harmless* (Ballantine Books, New York 1993)

Aronovitch, C.D.: The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. The Journal of Social Psychology **99**, 207–220 (1976)

Banse, R., Scherer, K.R.: Acoustic Profiles in Vocal Emotion Expression. Journal of Personality and Social Psychology **70**, No. 3, 614–636 (1996)

Berg, J. van den, Zantema, J.T., Doornenbal, P.: On the air resistance and the Bernoulli effect of the human larynx. J. Acoust. Soc. Am 29, 626–631 (1957)

Bezooyen, R. van: Characteristics and Recognizability of Vocal Expressions of Emotions. (Foris, Dordrecht 1984)

Bickerton, D.: *Language and Species* (1990)

Bishop, C.M..: Neural Networks for Pattern Recognition (Oxford University Press 1995)

Black, J.W.: The magnitude of pitch inflection. ICPhS 6, 177–181 (1967)

Cairns, D.A., Hansen, J.H.L: Nonlinear analysis and classification of speech under stressed conditions. J. Acoust. Soc. Am. **96**, 3392–3400 (1994)

Cahn, J.E.: Generating expression in synthesized speech. Master's thesis (Massachusetts Institute of Technology 1989)

Cohn, J.F., Katz, G.S.: Bimodal Expressions of Emotion by Face and Voice. Workshop on Face/Gesture Recognition and their Applications, The Sixth ACM International Multimedia Conference, Bristol, England (1998)

Clore, G.L.: Cognitive phenomenology: Feelings and the construction of judgment. In Martin, L., Tesser, A. (eds.): The Consruction of Social Judgments (Lawrence Erlbaum Associates, Hillsdale 1992)

Dailey, M., Cottrell, G., Adolphs, R.: A Six-Unit Network is All You Need to Discover Happiness. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (2000)

Damasio, A.R.: Descartes' Error: Emotion, Reason, and the Human Brain (Putnam Press, New York 1994)

Darwin, C.: *The Expression of the Emotions in Man and Animals* (University of Chicago Press, Chicago, London 1965. Original publication 1872)

Eckert, H., Laver, J.: *Menschen und ihre Stimmen* (Beltz Psychologie-Verlags-Union, Weinheim 1994)

Ekman, P.: An argument for basic emotions. Cognition and Emotion **6**, 169–200 (1992)

Fairbanks, G.: Recent experimental investigations of vocal pitch in speech. J. Acoust. Soc. Am. **11**, 457–466 (1940)

Fant, G.: Glottal source and excitation analyis. Speech Transmission Laboratory – Quarterly Progress and Status Report **1**, 85–107 (1979)

Fant, G., Lumby, H.: Laryngeal Mechanisms and Features. Introductory remarks for the proceedings of the 8th International Congress of Phonetics Sciences in Leeds. Phonetica **34**, 249–255 (1977)

Fant, G.: *Acoustic Theory of Speech Production* (Mouton, The Hague, 1970), 2nd ed.

Fernandez, R., and Picard, R.: Analysis and Classification of Stress Categories from Drivers' Speech. MIT Media Lab Perceptual Computing Sect. Tech. Report No. 513 (2000)

Fiukowski, H.: *Sprecherzieherisches Elementarbuch* (VEB Bibliographisches Institut Leipzig 1984)

Freienstein, H., Müller, K., Strube, H.W.: Vocal-tract parameter estimation from formant patterns. In Proceedings of the Joint Meeting ASA/EAA/DEGA, Berlin (1999)

Frick, R.W.: Communicating Emotion: The Role of Prosodic Features. Psychological Bulletin **97**, No.3, 412–429 (1985)

Friend, M., Farrar, M.J.: A comparison of content-masking procedures for obtaining judgments of discrete affective states. J. Acoust. Soc. Am. **96**, 1283–1290 (1994)

Glave, R.D., Rietveld, A.C.M.: Bimodal cues for speech loudness. J. Acoust. Soc. Am. **66**, 1018–1022 (1977)

Glave, R.D., Rietveld, A.C.M.: Is the effort dependence of speech loudness explicable on the basis of acoustical cues? J. Acoust. Soc. Am. **58**, 875–879 (1975)

Groening, M.: Lost our Lisa. The Simpsons, 5F17 (1998)

Hadding-Koch, K.: Acoustico–phonetic studies in the intonation of southern Swedish. Research Reports, Institute of Phonetics, University of Lund (Gleerup, Lund 1961)

Haykin, S.: Neural Networks: A Comprehensive Foundation (Macmillan, Englewood Cliffs 1994)

Hecht-Nielsen, R.: *Neurocomputing* (Addison-Wesley 1991)

Hecht-Nielsen, R.: Theory of the backpropagation neural network. In *Proceedings of the International Joint Conference on Neural Networks* **1**, 593–611 (IEEE Press, New York 1989)

Heraclitus: *On Nature* (Ephesus, ca. 500BC)

Hess, W.J.: *Pitch Determination of Speech Signals: Algorithms and Devices* (Springer, Berlin, Heidelberg, New York 1983)

Hess, W.J.: A pitch-synchronous digital feature extraction system for phonemic recognition of speech. IEEE Trans. ASSP **24**, 14–24 (1976)

Hollien, H.: Three major vocal registers: a proposal. ICPhS **7**, 320–331 (1972)

Holte, L., Margolis, R.H.: The relative loudness of third-octave bands of speech. J. Acoust. Soc. Am. **81**, 186–190 (1987)

Hurford, J.R., Studdert-Kennedy, M., Knight, C. (eds.): *Approaches to the evolution of language* (Cambridge University Press 1998)

Junqua, J.-C., Fincke, S., Field, K.: The Lombard Effect: A Reflex to Better Communicate with Others in Noise. In *Proceedings ICASSP '99*, 2083–2086 (1999)

Klatt, D.H., Klatt, L.C.: Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. **87**, 820–857 (1990)

Laver, J.: *The Phonetic Description of Voice Quality* (Cambridge University Press 1980)

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In Touretzky, D.S.

(ed.): *Advances in Neural Information Processing Systems 2 – NIPS '90* 396–404 (Morgan Kaufmann, San Mateo 1990)

Leinonen, L., Hiltunen, T.: Expression of emotional-motivational connotations with a one-word utterance. J. Acoust. Soc. Am **102**, 1853–1863 (1997)

Licklider, J.C.R., Pollack, I.: Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. J. Acoust. Soc. Am. **20**, 42–51 (1948)

Markel, J.D., Gray, A.H., Jr.: *Linear Prediction of Speech* (Springer, Berlin, Heidelberg, New York 1976)

Marty, A.: *Untersuchungen zur allgemeinen Grundlegung der Grammatik* und Sprachphilosophie (Niemeyer, Halle/Saale 1908)

Mayer, J.D., Salovey, P.: The intelligence of emotional intelligence. Intelligence 17, 433–442 (1993)

Mermelstein, P.: Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. **58**, 880–883 (1975)

Michaelis, D., Fröhlich, M., Strube, H.W.: Selection and combination of acoustic features for the description of pathologic voices. J. Acoust. Soc. Am. **103**, 1628–1639 (1998)

Monsen, R.B., Engebretson, A.M.: Study of variations in the male and female glottal wave. J. Acoust. Soc. Am **62**, 981–993 (1977)

Movellan, J.R., Mineiro, P., Williams, R.J.: Learning Path Distributions Using Non-equilibrium Diffusion Networks. In Jordan, M.I., Kearns, M..J., Solla, S.A. (eds.): *Advances in Neural Information Processing Systems 10 – NIPS '98* (MIT Press, Cambridge 1998)

Mullenix, J.W., Johnson, K.A., Topcu-Durgun, M., Farnsworth, L.M.: The perceptual representation of voice gender. J. Acoust. Soc. Am. **98**, 3080–3095 (1995)

Murray, I.R., Arnott, J.L.: Synthesizing emotions in speech: is it time to get excited? In *Proceedings of the Fourth International Conference on Spoken Language Processing* (1996)

Murray, I.R., Arnott, J.L.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J. Acoust. Soc. Am. **93**, 1097–1108 (1993)

Niedenthal, P.M., Kitayama, S. (eds.): The *Heart's Eye: Emotional Influences in Perception and Attention* (Academic Press, San Diego 1994)

Osgood, C.E., Snider, J.G.: *Semantic Differential Technique: A Sourcebook* (Aldine Publishing Co., Chicago 1969)

Parker, D.B.: A comparison of algorithms for neuron-like cells. In Denker, J. (ed.): *Proceedings of the Second Annual Conference on Neural Networks for Computing* **151**, 327–332 (Am. Inst. of Physics, New York 1986)

Paulus, E., Zwicker, E.: Programme zur automatischen Bestimmung der Lautheit aus Terzpegeln oder Frequenzgruppenpegeln. Acustica **27**, No.5 253–266 (1972)

Picard, R.W.: *Affective Computing* (MIT Press, Cambridge, Massachusetts, 1997)

Pittam, J., Scherer, K.R.: Vocal Expression and Communication of Emotion. In Lewis, M., Haviland, J.M. (eds.): *Handbook of Emotions* (Guilford Press, New York 1993)

Pittam, J., Gallois, C., Callan, V.: The long-term spectrum and perceived emotion. Speech Communication **9**, 177–187 (1990)

Protopapas, A., Lieberman, P.: Fundamental frequency of phonation and perceived emotional stress. J. Acoust. Soc. Am. **101**, 2267–2277 (1997)

Quast, H.: Recognition of Nonverbal Speech Features. In *Proceedings of the 6th Joint Symposium on Neural Computation, Caltech* (INC, San Diego1999)

Quast, H.: Absolute Perceived Loudness of Speech. In *Proceedings of the 7th Joint Symposium on Neural Computation, USC* (INC, San Diego 2000)

Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Waibel, A., Lee, K.-F. (eds.): *Readings in Speech Recognition* (Morgan Kaufmann, San Mateo 1989)

Rabiner, L.R., Schafer, R.W.: *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, New Jersey 1978)

Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., McGonegal, C.A.: A comparative study of several pitch detection algorithms. IEEE Trans. ASSP **24**, 399–413 (1976)

Rahim, M.G.: *Artificial Neural Networks for Speech Analysis/Synthesis* (Chapman & Hall, London, 1994)

Risberg, A.: Statistical studies of fundamental frequency range and rate of change. Speech Transmission Laboratory – Quarterly Progress and Status Report **4**, 7–8 (1961)

Robbins, H., Monro, S.: A stochastic approximation method. Annals of Math. Stat. **22**, 400–407 (1951)

Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev. **65**, 386–408 (1958)

Ross, E.D., Edmondson, J.A., Seibert, G.B.: The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice. Journal of Phonetics **14**, 283–302 (1986)

Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In Rumelhart, D.E., McClelland, J.L.: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* **1**, 318–362, (MIT Press, Cambridge, MA, 1986)

Scherer, K.R., Banse, R., Wallbott, H.G., Goldbeck, T.: Vocal cues in Emotion Encoding and Decoding. Motivation and Emotion **15**, 123–148 (1991)

Scherer, K.R.: On the symbolic functions of vocal affect expression. Journal of Language and Social Psychology **7**, 79–100 (1988)

Scherer, K.R.: Vocal Affect Signaling: A Comparative Approach. In Rosenblatt, J., Beer, C., Busnel, M., Slater, P.J.B. (eds.): *Advances in the Study of Behavior*, Vol 15 189–244 (Academic Press, New York 1985)

Scherer, K.R.: On the nature and function of emotion: A component process approach. In Scherer, K.R., Ekman, P. (eds.): *Approaches to emotion* 293–318 (Erlbaum, Hillsdale, NJ 1984)

Scherer, K.R.: Vocal cues to speaker affect: Testing two models. J. Acoust. Soc. Am. **76**, 1346–1356 (1984)

Scherer, K.R. (ed.): *Vokale Kommunikation: nonverbale Aspekte des Sprachverhaltens* (Beltz, Weinheim, Basel 1982)

Scherer, K.R.: Personality inference from voice quality: the loud voice of extroversion. European Journal of Social Psychology **8**, 467–487 (1978)

Scherer, K.R.: Randomized Splicing: A note on a simple technique for masking speech content. J. Exp. Res. Pers. 5, 155–159 (1971)

Schreiner, O.: Ein Vergleich verschiedener Frequenzzerlegungen zur Modulationsfilterung von Sprache. Diploma Thesis (Drittes Physikalisches Institut, Georg August Universität Göttingen 2000)

Schroeder, M.R.: *Computer Speech: Recognition, Compression, Synthesis* (Springer, Berlin, Heidelberg, New York 1999)

Schroeder, M.R.: Parameter estimation in speech: a lesson in unorthodoxy. Proc. IEEE **58**, 707–712 (1970)

Schroeder, M.R.: Period histogram and product spectrum: new methods for fundamental-frequency measurement. J. Acoust. Soc. Am. **43**, 829–834 (1968)

Schroeder, M.R.: Determination of the geometry of the human vocal tract by acoustic measurements. J. Acoust. Soc. Am. **41**, 1002–1010 (1967)

Scordilis, M.S., Gowdy, J.N.: Neural Network Control for a Cascade/Parallel Formant Synthesizer. Proc. IEEE Conf. on Acoustics, Speech, Signal Processing **1**, 297–300 (1990)

Sejnowski, T., Rosenberg, C.R.: Parallel Networks that Learn to Pronounce English Text. Complex Systems **1**, 145–168 (1987)

Shaffer, H.L.: Information rate necessary to transmit pitch-period durations for connected speech. J. Acoust. Soc. Am. **36**, 1895–1900 (1964)

Shpungin, B.E., Movellan, J.R.: A Multi-Threaded Approach to Real Time Face Tracking. UCSD MPLab Technical Report 2000.02 (2000)

Slaney, M., McRoberts, G.: Baby Ears: A Recognition System for Affective Vocalizations. Proc. IEEE Conf. on Acoustics, Speech, Signal Processing Seattle, WA (1998)

Steeneken, H.J., Hansen, J.H.: Speech under stress conditions: overview of the effect on speech production and on system Performance. In Proc. IEEE Conf. on Acoustics, Speech, Signal Processing **10**, 2079–2082 (1999)

Stevens, K.N.: Physics of Laryngeal Behavior and Larynx Modes. Phonetica **34**, 264–279 (1977)

Strube, H.W.: Acoustic Theory and Modeling of the Vocal Tract. In (Schroeder 1999)

Strube, H.W.: Psychoakustisch orientierte Merkmalsbildung in der Vorverarbeitung der Spracherkennung. In *Fortschritte der Akustik – Proceedings of DAGA* (1994)

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

Tartter, V., Braun, D.: Hearing smiles and frowns in normal and whisper registers. J. Acoust. Soc. Am. **96**, 2101–2107 (1994)

Teager, H.M., Teager, S.M.: Evidence for nonlinear sound production mechanisms in the vocal tract. In Hardcastle, W.J., Marchal, A. (eds): Speech Production and Speech Modeling, 241 – 261 (Kluwer 1990)

Tembrock, G.: Die Erforschung des tierlichen Stimmausdrucks (Bioakustik). In Trojan, F. (ed.): *Biophonetik* (Bibliogr. Inst, Mannheim 1975)

Terhardt, E.: Calculating virtual pitch. Hearing Res. **1**, 155–182 (1979)

Tolkmitt, F.J., Scherer, K.R.: Effect of Experimentally Induced Stress on Vocal Parameters. Journal of Experimental Psychology: Human Perception and Performance **12**, 301–313 (1986)

Traunmüller, H., Eriksson, A.: Acoustic effects of variation in vocal effort by men, women, and children. J. Acoust. Soc. Am. **107**, 3438–3450 (2000)

Werbos, P.J.: Beyond regression: New tools for prediction and analysis in the behavioral sciences. Doctoral Dissertation, Harvard University (1974)

Wilde, O.: *The Importance of Being Earnest*. (1899)

Widrow, B., Hoff, M.E.: Adaptive switching circuits. In *IRE WESCON Convention Record* Vol. 4, 96–104 (1960)

Zwicker, E., Fastl, H.: *Psychoacoustics: Facts and Models*. (Springer, Berlin, Heidelberg, New York 1990)

Name and Subject Index

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

Z