3.1 Allgemeine Angaben zum Teilprojekt A3

3.1.1 Titel:

Kurztitel: Speech
Incremental specification in speech

3.1.2 Fachgebiete und Arbeitsrichtung:

Acoustic Phonetics, Statistical Signal Processing

3.1.3 Leiter:

Yang, Bin, Prof. Dr.-Ing., 07.05.1963
Lehrstuhl für Systemtheorie und Signalverarbeitung (LSS),
Fakultät für Informatik, Elektrotechnik und Informationstechnik,
Universität Stuttgart, Pfaffenwaldring 47, 70550 Stuttgart
Telefon: 0711/685-7330
Telefax: 0711/685-7311
E-Mail: bin.yang@LSS.uni-stuttgart.de

Dogil, Grzegorz, Prof. Dr., 24.12.1952
Institut für Maschinelle Sprachverarbeitung (IMS),
Philosophisch-Historische Fakultät,
Universität Stuttgart, Azenbergstraße 12, 70174 Stuttgart
Telefon: 0711/121-1379
Telefax: 0711/121-1366
E-Mail: dogil@ims.uni-stuttgart.de

Wokurek, Wolfgang, Dipl.-Ing. Dr. techn. (TU Wien), 17.04.1961
Institut für Maschinelle Sprachverarbeitung (IMS),
Philosophisch-Historische Fakultät,
Universität Stuttgart, Azenbergstraße 12, 70174 Stuttgart
Telefon: 0711/121-1383
Telefax: 0711/121-1366
E-Mail: wokurek@ims.uni-stuttgart.de

Ist die Stelle des Leiters/der Leiterin des Projektes befristet?

☑ nein ☐ ja, befristet bis zum _____
☐ eine weitere Beschäftigung ist vorgesehen bis zum _____

3.1.4 In dem Teilprojekt sind vorgesehen:

Untersuchungen am Menschen oder am menschlichen Material

☐ ja ☑ nein

Die erforderliche Zustimmung der zuständigen Ethikkommission liegt dem Antrag zum Teilprojekt in Kopie bei

☐ ja ☐ nein

klinische Studien im Bereich der somatischen Gentherapie

☐ ja ☑ nein

Tierversuche

☐ ja ☑ nein

gentechnologische Untersuchungen

☐ ja ☑ nein

Untersuchungen an humanen embryonalen Stammzellen

☐ ja ☑ nein

Die gesetzliche Genehmigung liegt vor

☐ ja ☐ nein

3.1.5 Beantragte Förderung des Teilprojektes im Rahmen des Sonderforschungsbereichs (Ergänzungsausstattung)

Haushaltsjahr	Personalmittel	Sachmittel	Investitionsmittel	Gesamt
2006/2			0	
2007			0	
2008			0	
2009			0	
2010/2			0	

(Beträge in Tausend EUR)

3.2 Zusammenfassung

Short summary
This SFB project A3 will focus on the incremental specification of speech at the acoustic, phonetic, and prosodic level. Our goal is to develop some general principles and computational procedures for incremental specification, which we will then apply to the specification of the distinctive features of speech such as landmarks and pivots, and of the speaker's specific features in order to investigate the phonetic and prosodic representation of speech.

German summary
Dieses SFB-Projekt A3 untersucht die inkrementelle Spezifikation der Sprache auf der akustischen, phonetischen und prosodischen Ebene. Unser Ziel ist die Entwicklung von allgemeinen Prinzipien und Rechenverfahren zur inkrementellen Spezifikation, die wir dann auf die Spezikation der distinktiven Merkmale der Sprache wie Landmarks und Pivots und

auf die Spezikation der sprecherspezifischen Merkmale anwenden wollen zur Erforschung der phonetischen und prosodischen Sprachrepresentation.

Extended summary

This SFB project A3 will focus on the incremental specification of speech at the acoustic, phonetic, and prosodic level. Our goal is to develop some general principles and computational procedures for incremental specification, which we will then apply to the specification of the distinctive features of speech such as landmarks and pivots, and of the speaker's specific features in order to investigate the phonetic and prosodic representation of speech.

When we listen to speech we follow two parallel perceptual goals. The first goal is the fast and robust lexical access. This goal is best achieved by building underspecified phonological representations which scan the speech signal for a small set of distinctive features, which are the building blocks of phonological words. The other goal is to identify the speaker, to recognize her accent to get hints at her attitudes and emotions from the details of prosody and voice quality. This goal is achieved by considering very richly specified representations of speech. These two goals have been considered for a long time as complementary, or even independent (cf. the distinction between phonetics and phonology). However, recent evidence shows that processing of speech to achieve these two goals is not only parallel but that it is also highly interactive. Laboratory Phonology, which has become the leading framework for studying speech in the last decade, has provided substantial evidence for phonological and phonetic underspecification and complete signal specification as required and interacting modes of representation. Our project will apply models and methods from linguistic phonetics and from statistical signal processing to model these interactions in an explicit computational framework. The key ingredient of the model is the concept of context. The listener increasingly takes various types of context information into account as it becomes available. The context information comprises the immediate phonetic context, the prosodic context, the linguistic context, the discourse context, as well as speaker specific information. Our research will design a formal account of how such context information is taken into account as the listener abstracts away from acoustic signal to extract relevant features from speech.

3.3 Ausgangssituation des Teilprojekts

3.3.1 Stand der Forschung

Incremental specification framework

In this project, the main focus will be placed on methods of specification in phonetics which supplement the well known arguments for phonological underspecification. When Trubetzkoy (1939) formulated criteria to determine which of the universally possible correlations among speech sound are relevant for lexical distinction, the first symbolic system of distinctive features was established. The non-distinctive features were irrelevant and were never specified by structural phonology. In generative phonology the first attempts at incremental specification were undertaken in that the unspecified features were allowed to be filled in by an ordered set of rules during phonological derivation, yielding completely specified symbolic surface representations. The formalism of generative phonological rules

allowed only the immediate segmental context of the underspecified features to be taken into account. However, it possessed an explicit mechanism for incremental specification ("disjunctive ordering"). Autosegmental phonology allowed for the extensions of the specification context (Ringen, 1975; Goldsmith, 1976) to segmental and tonal features, and lexical phonology (Kiparsky, 1985; Archangeli, 1988) presented some solutions to the problem of which features are universally available for insertion into unspecified lexical entries. However, the symbolic attempts of specification in phonology never reached the degree of formal precision which would allow them to go beyond the observation level (cf. Clements 1987; Steriade 1987; McCarthy & Taub 1992, for critical reviews of such models). The job of fully specifying speech has been passed on to phonetics.

Linguistic phonetics requires the concept of underspecification to explain the phenomena of coarticulation. Keating (1988) in her window model of coarticulation suggested that certain segments are more or less transparent to phonetic rules of coarticulation: "When phonetic rules build trajectories between segments, an unspecified segment will contribute nothing of its own to the trajectory" (Keating, 1988, 281). Others have considered the phenomenon from the opposite angle, in that they looked at cases of "coarticulatory resistance" in segments which block intersegmental trajectories (Byrd, 1996). Although numerical and statistical procedures were used to describe the phonetic data, no formal numerical model of specification in context has been attempted which would go beyond the description of selected examples (cf. Pierrehumbert & Beckman 1988; Cohn 1990, 1993; Vollmer 1997; Zsiga 1997).

The most ambitious program of specification in speech has been inspired by models of signal processing as used in acoustic phonetics. The "landmark specification model" as developed by Stevens and associates (cf. Stevens 1989, 1998; Stevens et al. 1986; Stevens 2005) assumes that the articulatory-acoustic-auditory relations are *quantal*. The articulatory-acoustic relations are quantal in the sense that the acoustic pattern shows a change from one state to another as the articulatory pattern is varied through a range of values. The regions in speech where the acoustic parameter undergoes large changes for relatively small manipulations of the articulatory parameter form the basis of distinctive features *and* the basis for underspecification. The model does not make a distinction between phonology and phonetics. It rather uses numerical and symbolic representations to refer to same areas in speech representation. Stevens suggests that a phonetic and phonological theory should specify only those areas of the signal when a feature *change* is implemented. These areas, to which Stevens refers as *landmarks* in the acoustic steam, have to be fully specified phonetically and labelled with a symbolic distinctive feature mark. The rest of the characteristics of the signal can be interpolated from the landmark areas. Thus in the speech signal there will be an alternation between narrow landmark regions marked by acoustic events where there are rapid changes, and temporal regions where the acoustic parameters remain relatively steady. The idea behind landmark specification theory is that the temporally broad areas of stability can be predicted from the temporally narrow areas of change. Although a lot of research has been devoted to the detection of landmarks and distinctive features, much less research has been directed towards the specification process itself. Attempts at contextual models of feature enhancement (Stevens et al., 1986) have been abandoned, and incremental specification in speech as a subject on its own has not been systematically studied up to now. In our research we want to supplement this shortcoming and provide formal models of specification which could be extended to speech. Formal models of in-

cremental specification are well established in statistical signal processing and incremental specification of simple mathematical objects does exist.

A simple geometrical example is to approximate a given vector by a linear combination of an increasing number of other vectors (Golub & Van Loan, 1983). A similar mathematical problem is the representation of a function by means of a number of other kernel functions. In statistics, a sequence of random variables or random vectors is used to predict another random variable or vector (Kay, 1993). In information theory, the information amount is measured by the entropy. The part of information which can not be specified by other random events, is called the conditional entropy (Cover & Thomas, 1991). One example from statistical signal processing is a linear prediction of the current signal sample $x(n)$ from its N past samples $x(n-1), \ldots, x(n-N)$ (Haykin, 1995). Another example is the low-rank approximation of a full rank correlation matrix for a random vector process (Scharf, 1991). In all examples, there exist formal computational procedures to incrementally specify an object by other ones. In addition, all these incremental specifications share some common approaches like innovation, analysis by synthesis, information feedback and interaction between static and dynamic incremental specifications. The interesting research question is whether and to which extent these principles and methods are applicable or extensible to speech.

Incremental specification of perceptually critical regions of speech

Articulator-bound distinctive features are a phonological representation of speech utterances (Stevens, 1998). They enable us to search for discrete symbols that are the key concept of communication (Shannon, 1948) in the continua of speech acoustics. An articulatory feature is distinctive if it can potentially be used to distinguish between words of a language. A source of practical problems for articulatory feature based speech processing is mirrored in the word "potential". Acoustic cues often are extracted straightforward from the signal. But if articulatory gestures overlap or more than one articulatory gesture could have produced the speech acoustics, the feature extraction procedure is still unknown (Stevens, 2002).

Landmarks were defined to identify times when the acoustic manifestations of the linguistically motivated distinctive features are most salient (Liu, 1996). They were proposed to be used in an automatic speech recognition system. Algorithm for automatically detecting the following acoustically abrupt landmarks were given in (Liu, 1996): stop closures and releases, nasal closures and releases, and the point of cessation of free vocal fold vibration due to a velopharyngeal port closure at a nasal-to-obstruent juncture. The detection algorithm divides the wideband spectrogram (6ms) into 6 frequency bands. The energy in each band is processed with respect to two time scales. The coarser scale is used for detection and the finer scale for temporal localization. The landmarks glottis, sonorant and burst are detected based on the peak changes of the band energies. Conclusions are that about 8% of the landmarks were missed (not detected) and that adding prosodic information could improve performance. Algorithms to detect the nonabrupt landmarks (for semivowels) and the vocalic landmarks (for vowels) still need to be developed.

Based on the experimental and theoretical results from psycholinguistics, Lahiri & Marslen-Wilson (1991) and Lahiri & Reetz (2002) have developed the Featurally Underspecified Lexicon (FUL) model. FUL assumes that phonological representations consist of

maximally underspecified features. The acoustic signal is analyzed for coarse acoustic characteristics like bandpass energy measurements and LPC based formant estimates (Reetz, 1999) which are then transformed into phonological features and mapped onto lexical representations. If the features extracted from the signal are found in the lexicon, a match is obtained. No match occurs if the extracted features are not represented in the underspecified lexical representations. A mismatch occurs if the features conflict with the lexical representations. In this case, the mapping is rejected.

FUL is similar to the ISC procedure followed in our project in that it operates on maximally underspecified features for the purpose of lexical access. However, the ISC procedure is more general, as it specifies procedures for full specification as required for recognition of extralinguistic parameters of speech. The major difference between the two models lies in the role of context. FUL´s feature extraction procedure is explicitly context independent.

Finite-state models and event logics are proposed to deal with the recognition of new words (Carson-Berndsen, 2000). The system uses features extracted from speech with fine grained statistical models that were trained on speech corpora. The features are combined in multilinear tiers and parsed with an finite-state and event logic component. Prosodic context is not ruled out principally but currently not incorporated.

Incremental specification of speaker specific features

In most linguistic theories idiosyncratic aspects of speech (voice features, ambient noise, etc.) are considered noise and disregarded. However, new experimental evidence collected mainly in the area of laboratory phonology and exemplar theory suggest that perceptual details of speakers voice are stored in memory and are integral to later perception (Goldinger, 1996). Detailed phonetic knowledge that native speakers have about the categories of their language includes speaker specific features and thus has to be modelled by a theory (cf. Pierrehumbert 2001, and motivation for exemplar-based speech representation in project A2). Generally, we distinguish between acoustic and non-acoustic speaker specific features. Examples of acoustic features are gender, prosody, emotion, voice quality, age etc. Non-acoustic speaker specific features include mimic and gestures. By evaluating the speaker specific features, the listener obtains information about the physical, psychological, and emotional characteristics of the speaker. In this SFB proposal, we focus on the acoustic speaker specific features only.

In contrast to speech recognition where the linguistic content is of interest, speaker specific features are a less explored field. Nevertheless, speaker specific features and in particular emotion recognition in human-computer interaction have attracted much attention in the last years. Different speech features (pitch, energy, duration, formant, cepstral coefficients, time variations, voice quality parameters, ...) have been tested and different classifiers (hidden Markov model, neural network, linear discriminant analysis, Gaussian classifiers, ...) have been evaluated for emotion recognition (Haderlein & Nöth, 2003; Batliner et al., 2000; Petrushin, 2000; Cowie et al., 2001; Huber, 2002; Lee & Narayanan, 2005). In comparison, speaker specific features other than emotion have been less studied. Their incremental specification in context (ISC) has not been addressed yet.

3.3.2 Eigene Vorarbeiten

Incremental specification framework

We have extensively used incremental specifications in statistical signal processing. The purpose was to approximate a given signal object with an increasing accuracy in terms of other objects. Below we briefly describe two examples.

In the first case, the object to be incrementally specified is a discrete-time signal $d(n)$ at the time instant n. We use a linear combination of the most recent N observations $x(n), x(n-1), \ldots, x(n-N+1)$ of a second signal $x(n)$ to approximate $d(n)$. Let $w_1, w_2, \ldots, w_N$ be the parameters involved in the linear combination. The approximation for $d(n)$ is $y(n) = w_1 x(n) + w_2 x(n-1) + \ldots + w_N x(n-N+1)$. The difference $e(n) = d(n) - y(n)$ denotes the approximation error. The goal is to choose the parameters w_i in such a way that the approximation error $e(n)$ is minimized in some sense, i.e. the approximation of $d(n)$ by $x(n)$ is optimized.

Starting with an initial rather inaccurate approximation of $d(n)$ by only one sample $x(n)$, we gradually increase the number of samples N until a certain approximation accuracy is achieved. In addition, we distinguish between static and dynamic incremental specification. In the former case, the relationship between the signals $d(n)$ and $x(n)$ is fixed and we only need to compute the optimum approximation once. In the latter case, $d(n)$ depends on $x(n)$ in a time varying way and the approximation process becomes adaptive. We have developed efficient algorithms for this computation (Yang, 1991; Yang & Böhme, 1992; Yang, 1995c). Note that linear prediction is a special case of this approximation problem with $d(n) = x(n+1)$. Linear prediction has been successively applied to speech compression in mobile phones, internet telephony etc.

In the second example, the object to be specified is the correlation matrix $\mathbf{R}$ of a random vector process. Given a sequence of observations, the problem is to compute a sequence of incremental underspecifications of $\mathbf{R}$ in terms of its eigenvectors and eigenvalues. We have developed efficient algorithms for this purpose (Yang, 1995b; Kavčić & Yang, 1996b; Yang, 1996). If the number of required eigen-components corresponding to the depth of incremental specification is unknown, we have designed computational procedures to estimate this number based on statistical criteria (Yang, 1995a; Kavčić & Yang, 1996a).

Incremental specification of perceptually critical regions of speech

Pivots are a concept proposed to describe the variety and robustness of human speech communication. Dogil (1986, 1988) has provided a framework which accounts for the fact that even in most difficult circumstances (foreign accent, loud environment etc.) listeners comprehend speech quickly and efficiently. There must be some signals in the phonetic string which are particularly easy to grasp and to process. These signals are called "pivots" and parsers working with these signals are called "pivot parsers". The pivot parser proposed by Dogil (1988) recognizes the unmarked prototypical consonant-vocal parts of syllables which were found sufficient for sentence recognition during comprehension experiments. We have continued the research on the detection and classification of pivots at IMS.

As preparatory work for the detection of perceptually relevant acoustic speech structures, we view the time-frequency analysis of speech (Wokurek et al., 1987), the robust sound class detector (Wokurek, 1995), the automatic classification of pitch accents (Fach

& Wokurek, 1995), the time-frequency analysis of glottal structures (Wokurek, 1997), the speech segmentation based on entropy rate (Wokurek, 1999), and the estimation of voice quality parameters (Classen et al., 1998; Wokurek & Pützer, 2003; Pützer & Wokurek, 2006).

Incremental specification of speaker specific features

We have also studied the speaker specific features of speech. In particular, we focused on the voice quality parameters: open quotient gradient (OQG), glottal opening gradient (GOG), skewness gradient (SKG), rate of closure gradient (RCG), and incompleteness of closure (IC) (Stevens & Hanson, 1998) and their applications. These parameters are closely related to the glottal excitation of speech and thus describe well the speaker characteristics. They are estimated from the speech signal by spectral gradients. It has been shown that the voice quality parameters allow the distinction between different speaking groups like gender (Wokurek & Pützer, 2003), pathological and non-pathological speakers (Wokurek & Pützer, 2003), and word stress (Classen et al., 1998; Classen & Wokurek, 2003). It was also demonstrated that they can be used to classify different emotions and voice qualities for individual speakers (Lugger & Yang, 2006). Another interesting result is that some of the voice quality parameters are quite insensitive to acoustic disturbances like background noise and room reverberation (Lugger et al., 2006). This makes the voice quality parameters a robust tool for specifying speaker specific features.

Another type of speaker specific features is the perspective category. It contains channel effects (reverberation) and the speaker location (distance and direction). We have developed algorithms for estimating the speaker location from the time differences of arrival of microphone pairs and studied the geometrical impact of the microphone array to the localization accuracy (Yang & Scheuing, 2005, 2006; Scheuing & Yang, 2006).

3.3.3 Liste der publizierten einschlägigen Vorarbeiten

Referierte Veröffentlichungen – in wissenschaftliche Zeitschriften

G. Dogil and B. Möbius, "Toward a perception based model of the production of prosody," *Journal of the Acoustical Society of America*, vol. 110, no. 5, Pt. 2, p. 2737, 2001.

G. Dogil, K. Claßen, M. Jessen, K. Marasek, and W. Wokurek, "Stimmqualität und wortbetonung im deutschen," *Linguistische Berichte*, vol. 174, pp. 202–245, 1998.

B. Yang, "Asymptotic convergence analysis of projection approximation subspace tracking algorithms," *Signal Processing, Special Issue on Subspace Methods, Part I: Array Signal Processing and Subspace Computations, eds. M. Viberg and P. Stoica*, vol. 50, pp. 123–136, 1996.

A. Kavčić and B. Yang, "Adaptive rank estimation for spherical subspace trackers," *IEEE Trans. Signal Processing*, vol. 44, pp. 1573–1579, 1996.

A. Kavčić and B. Yang, "Subspace tracking with adaptive threshold rank estimation," *Special Issue on Array Optimization and Adaptive Tracking Algorithms, Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 14, pp. 75–91, 1996.

B. Yang, "An extension of the PASTd algorithm to both rank and subspace tracking," *IEEE Signal Processing Letters*, vol. 2, pp. 179–182, 1995.

B. Yang, "Projection approximation subspace tracking," *IEEE Trans. Signal Processing*, vol. 43, pp. 95–107, 1995.
B. Yang, "A QR multichannel least squares lattice algorithm for adaptive nonlinear filtering," *AEÜ International Journal of Electronics and Communications*, vol. 49, pp. 171–182, 1995.
B. Yang and J. F. Böhme, "Rotation based RLS algorithms: Unified derivations, numerical properties and parallel implementations," *IEEE Trans. Signal Processing*, vol. 40, pp. 1151–1167, 1992.
G. Dogil, "Der Pivot Parser: Eine Hypothese zur Sprachwahrnehmung," *Linguistische Berichte*, vol. 106, p. 456–470, 1986.

Referierte Veröffentlichungen – auf wesentlichen Fachkongressen

M. Lugger, B. Yang, and W. Wokurek, "Robust estimation of voice quality parameters ander real world disturbances," in *Proc. IEEE ICASSP*, 2006.
M. Lugger and B. Yang, "Klassifikation verschiedener Sprechergruppen mit Hilfe von Stimmqualitätsparametern," in *7. ITG-Fachtagung Sprach-Kommunikation*, 2006.
B. Yang and J. Scheuing, "A theoretical analysis of 2d sensor arrays for tdoa based localization," in *Proc. IEEE ICASSP*, 2006.
B. Yang and J. Scheuing, "Cramer-Rao bound and optimum sensor array for source localization from time differences of arrival," in *Proc. IEEE ICASSP*, vol. 4, pp. 961–964, 2005.
A. Schweitzer, N. Braunschweiler, G. Dogil, and B. Möbius, "Assessing the acceptability of the SmartKom speech synthesis voices," in *Proceedings of the 5th ISCA Speech Synthesis Workshop (Pittsburgh, PA)*, pp. 1–6, 2004.
B. Möbius and G. Dogil, "Phonemic and postural effects on the production of prosody," in *Proceedings of the Speech Prosody 2002 Conference* (B. Bel and I. Marlien, eds.), (Aix-en-Provence), pp. 523–526, Laboratoire Parole et Langage, 2002.
G. Dogil and B. Möbius, "Towards a model of target oriented production of prosody," in *Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark)*, vol. 1, pp. 665–668, ISCA, 2001.
G. Dogil, "Acoustic landmarks and prosodic asymmetries," in *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS)*, (San Francisco), pp. 128–134, 1999.
G. Dogil, M. Jilka, and G. Möhler, "Rules for generation of tobi-based american english intonation," *Speech Communication*, vol. 28, no. 2, pp. 83–108, 1999.
G. Dogil, J. Kuhn, J. Mayer, G. Möhler, and S. Rapp, "Prosody and discourse structure: Issues and experiments," in *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications*, (Athens, Greece), pp. 99–102, 1997.

Referierte Veröffentlichungen – in Monographien

G. Dogil, "Understanding prosody," in *Psycholinguistics: An International Handbook* (G. Rickheit, T. Herrmann, and W. Deutsch, eds.), pp. 544–566, Berlin: Mouton de Gruyter, 2003.
G. Dogil, "How we say when it happens," in *Intonation of Aspectual Meaning: Remarks on NOCH in German* (H. Kamp and U. Reyle, eds.), pp. 1–17, Tübingen: Max Niemeyer Verlag, 2002.
G. Dogil, "The phonetic manifestation of word stress," in *Word Prosodic Systems in the Languages of Europe* (H. van der Hulst, ed.), pp. 273–334, Berlin: de Gruyter, 1999.

G. Dogil and M. Jessen, "Phonologie in der nähe der phonetik: Die affrikaten im polnischen und deutschen," in *Phonologie* (M. Prinzhorn, ed.), pp. 223–279, Opladen: Westdeutscher Verlag, 1989.
B. Yang, A. Pèrez-Neira, and M. A. Lagunas, "Adaptive arrays for communication," in *Digital Signal Processing in Telecommunications* (A. R. F. Vidal, ed.), ch. 3, pp. 93–170, Berlin: Springer-Verlag, 1996.
B. Yang, *Rotationen verwendende Parallelalgorithmen für hochkomplexe Signalverarbeitung und systolische Rechenfelder*. Düsseldorf: VDI-Verlag, 1991.
G. Dogil, *Linguistic Phonetic Features: A study in systematic phonetics*. Duisburg: Linguistic Association University of Duisburg, 1988.
G. Dogil, *The Pivot model of speech parsing*. Wien: Verlag der Österreichischen Akademie der Wissenschaften, 1988.

3.4 Planung des Teilprojekts (Ziele, Methoden, Arbeitsprogramm)

3.4.1 Fragestellung

This SFB project will focus on the incremental specification of speech in the acoustic, phonetic, and prosodic context.

Incremental specification framework

By incremental specification we understand a sequential process of providing information gain for a certain object. Starting with an initial highly underspecified representation τ_0, we provide successively additional information $i_1, i_2, \ldots$ about the object resulting in a sequence of more detailed specifications $\tau_1, \tau_2, \ldots$, see Fig. 3.1. Due to the incremental nature of this process, the information amount is growing and each specification τ_n contains its predecessor τ_{n-1} as a subset. The additional information i_n is determined from the context.

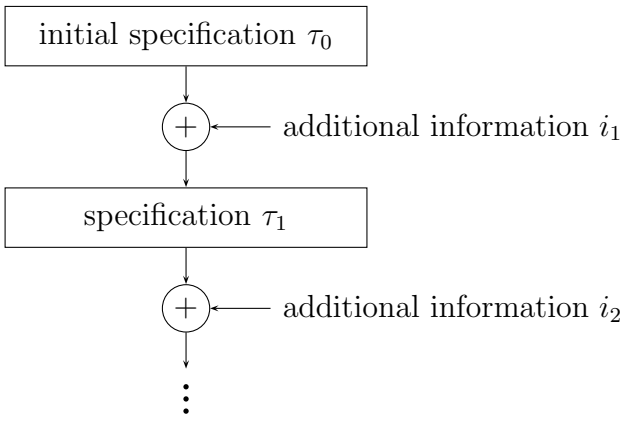

Abbildung 3.1: Incremental specification

We understand speech as fully specified when we have a complete acoustic wave (signal) of the speech. It is a particular acoustic realization of a particular text by a particular speaker at a particular time. From this full specification (or exemplar, cf. project A2), we are theoretically able to extract all desired information at different linguistic and paralinguistic levels. For a particular application, however, such a full specification is highly inefficient because of its information redundancy, hidden features, and large memory requirement. Therefore, appropriate underspecifications of speech dedicated to particular applications are highly desirable. Examples are speech compression, speech recognition, speaker identification, detection of gender, emotion, prosody etc.

Despite some successful implementations of these underspecifications in practice, we are still far away from a thorough understanding of speech communication. In this project, we are interested in a general framework of incremental specification in speech:

- How to derive a sequence of underspecifications at different levels of details? What is the relationship between adjacent specification levels?
- What is the impact of the context to the specification process? Are there some general principles and computational procedures for the incremental specification?
- A general principle from mathematics and statistical signal processing dealing with incremental specification is the so called innovation approach (Kailath, 1970; Kailath et al., 2000; Haykin, 1995). The basic idea is to provide completely new (innovative) information to the previous specification at each step. Is this approach applicable to incremental specification in speech?
- We distinguish between static and dynamic incremental specification. In the former case, we deal with a fixed information amount. In the latter case, the information amount and thus the context is growing in time. Are there common approaches for the static and dynamic incremental specification? What are their differences?
- Depending on the time scale used, the context can be local and global. What is their relationship to the static and dynamic incremental specification?
- Incremental specification is a sequential process of adding details. It is also important to study the reverse process of incremental abstraction in which we extract a few key features from the full specification. What are the relationships and interactions between incremental specification and incremental abstraction?

In addition to the study of the general framework of incremental specifications, we will also apply this framework to two particular areas of speech representation: a) distinctive features of speech derived at landmarks and pivots at the phonetic level, b) speaker specific features at the prosodic level. Both issues will be addressed below.

Incremental specification of perceptually critical regions of speech

We know that speaker and listener both exploit acoustic together with structural qualities of speech and language for communication. Phonation and articulation continuously produce acoustic structures where communication relevant information is not distributed equally over time and frequency but is placed at some characteristic points. This fact is described

in speech research as events like acoustic cues, landmarks, pivots or more abstractly as phonetic and phonologic features. This is a highly underspecified representation of speech. What we want to know are: How are such events and features optimally chosen, detected and specified with respect to the processing of context? How is the incremental nature of the phonetic context taken into consideration best?

Incremental specification of speaker specific features

Human communication consists of two channels, the explicit channel and the implicit channel. While the explicit channel transmits explicit messages like the linguistic content of a speech, the implicit channel transmits implicit messages or speaker specific features. In speech communication, speaker specific features include gender, prosody (the speaker specific part of prosody), emotion, voice quality, age of the speaker, social background, health, dialect etc.

From this broad spectrum of speaker specific features, there has been, from the application point of view, a strong focus on emotion recognition in the last years due to widely used man-machine interfaces. In comparison, the other types of speaker specific features have been less explored although it is by now clear that they play a very important role in language processing (cf. the empirical finding of Exemplar Theory discussed in detail in A2). Our intention in this SFB project is not a new emotion recognizer, but rather a better understanding of how speaker specific features change in context. In particular, we consider these speaker specific features which can be measured by the voice quality parameters, see section 3.3.2:

- Which speaker specific features correlate with the voice quality parameters?
- How can we extract voice quality parameters from the acoustic signal?
- What is the impact of the semantic context to the speaker specific features? How do the voice quality parameters change in context?
- How can we apply the incremental specification framework to the speaker specific features? Is the innovation approach applicable to this problem?
- How do linguistic and speaker specific features interact and how do they influence each other?

3.4.2 Ziele

This SFB project will focus on the incremental specification in speech. We address three closely related issues: a) an incremental specification framework, b) incremental specification of the distinctive features of speech, c) incremental specification of the speaker specific features in terms of the voice quality parameters. The goal of the first study is to better understand the general incremental specification process and to develop some formal computational procedures. One main idea is to look at the well established incremental specifications in other disciplines like signal processing and mathematics and to study how can we extend these principles to the much more complicated speech representation. Then

we apply this framework to two particular areas of speech representation: distinctive features of speech as a highly underspecified representation and speaker specific features. The goal of these studies is to gain a deep insight into how the different information channels of speech are coded, what their relationships and interactions are, and how can we model and describe them.

Fig. 3.2 illustrates the goals of this research. The three issues mentioned above are assigned to three work packages AP1, AP2, and AP3 (AP: Arbeitspaket). We observe that AP1, as it addresses the incremental specification framework, is a broad topic and crosses the acoustic, phonetic and prosodic level. In comparison, AP2 and AP3 dive into the details of the phonetic and prosodic representation. Fig. 3.2 shows that it is not our intention to extend AP1 beyond the prosodic level within the first four years. The reason is quite simple: At these high linguistic levels, we are dealing more with symbols and rules than with numbers and statistics as it is the case in mathematics, signal processing, and information theory. Nevertheless, extending the framework of incremental specification to the higher linguistic levels is our long-term goal.

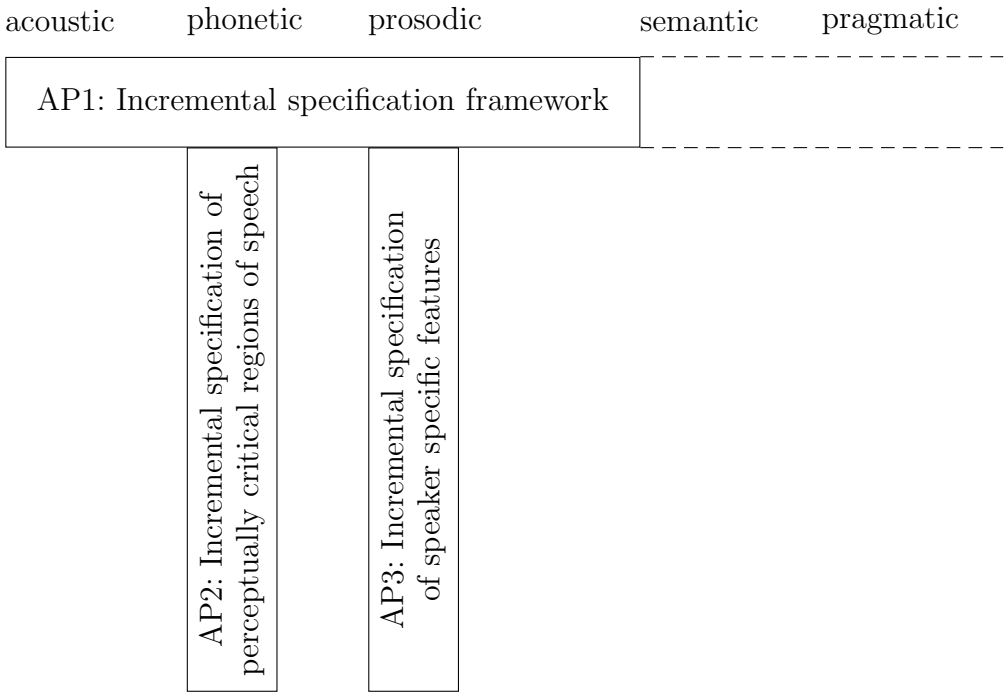

Abbildung 3.2: Three goals of this SFB project

The aim of AP2 is the development of a new model for the acoustic and phonetic layer of speech communication based on information theory. The new notion is that the information is sophistically distributed among the various channels that are established between speaker and listener by means of phonation, articulation and perception. The strategies used for the distribution of the information by the speaker and those used for information extraction by the listener shall be investigated in the framework of incremental specification in phonetic and prosodic context. That means the information extracted from the speech signal in a given context situation will be specified. We view this as the incremental step which can be repeated to form a process of iterated information extraction that are useful in various application fields outside this project including context dependent speech

recognition frontends and speaker identification.

The results of this work are structured models for perceptually relevant acoustic speech events including their use to carry speech information in phonetic and prosodic context.

3.4.3 Methoden und Arbeitsprogramm

A unified set of models will be constructed extending and generalizing the known methods of detecting and processing landmarks, phonetic and phonological features, and prosodic and voice quality parameters. These will comprise of signal processing together with hierarchies of continuous and symbolic probabilistic models ruling the distribution of information to the available channels of speech communication under phonetic prosodic contexts. These models shall be dynamic to adapt to changes in the communication channels like ambient noise or competing speakers. They should be flexible to be speaker independent and apply to many different communication environments and allow a specification to a particular speaker listener pair. Finally the form of these models, their changes and the information extracted in specific phonetic and prosodic contexts like stress, will be studied and documented.

Table 3.1 shows the work plan of this SFB project. It consists of three work packages **APx** corresponding to the three goals in Fig. 3.2. Each package consists of a number of subpackages. We have also defined a total number of six milestones **Mx**. Below all packages and subpackages will be described in details.

AP1.1: Incremental specification in phonology, phonetics and in some other disciplines

In this subpackage we will study and review various attempts to formulate principles of incremental specification in phonetics and phonology. Furthermore, we will study well known formalisms of incremental specification from other disciplines like signal processing, statistics, information theory, and mathematics.

Phonology: Disjunctive rule principles in the classical generative phonology restricted the context of rule application to the immediate neighborhood of the affected distinctive feature. However, the rules are allowed to apply again if their structural description is met by a broader context. In this way incremental growth of context was possible. Similar formal principles are finding their way into constraint evaluation of current optimality theory (Karttunen, 1998; Eisner, 2000)

Feature enhancement regulates the information gain between distinctive and redundant phonological features. For example, the feature [back] in vowels is reinforced by the feature [round] in the same context, and the feature [voice] is correlated with the feature [spread glottis]. The formal relations between correlated and uncorrelated features have been addressed, but no formal model has been developed to our knowledge.

Phonetics: The landmark specification theory assumes that only areas of the quantal acoustic change have to be fully specified. The rest of the acoustic characteristics of the signals can be interpolated from the landmark areas.

The DIVA model (for Directions Into Velocities of Articulators, Guenther et al. 1998; Dogil & Möbius 2001) assumes that speech is represented by a set of regions in perceptual space. Guenther has shown that only the *directions* in the perceptual space have to be

	Jahr	06		07				08				09				10	
	Quartal	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
AP1	Incremental specification framework	x	x	x	x	x	x	x	x	x	x						
1.1	Incremental specification in phonology, phonetics and in some other disciplines	x	x	x													
1.2	Incremental specification in speech processing			x	x	x											
M1	Studies of known incremental specifications finished					•											
1.3	The innovation approach					x	x	x									
1.4	Incremental specification and incremental abstraction						x	x	x								
1.5	Static and dynamic incremental specification							x	x	x							
1.6	Context dependent incremental specification								x	x	x						
M2	Studies of the general incremental specification framework finished										•						
AP2	Incremental specification of perceptually critical regions of speech	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
2.1	Implementation of existing landmark detectors	x	x	x													
2.2	Study of these landmark detectors on phonemes and syllables			x	x	x											
M3	Study of existing landmark detector finished					•											
2.3	Generalize the matched filter detector to time-frequency structures						x	x	x								
2.4	Apply this generalized detector to critical regions									x	x	x					
M4	Critical region model/detector designed											•					
2.5	Structured critical regions model											x	x	x			
2.6	Dependency of the structured critical regions model on word stress, phrase stress, and interpersonal context												x	x	x	x	x
M5	Critical region model extended to context																•
AP3	Incremental specification of speaker specific features									x	x	x	x	x	x	x	x
3.1	Speaker specific features and voice quality parameters									x	x	x					
3.2	Context dependent changes of speaker specific features based on the MapTask database										x	x	x	x			
3.3	Creation of a project specific database												x	x			
3.4	Context dependent changes of speaker specific features based on the new database													x	x	x	x
M6	Studies of speaker specific features finished																•

Tabelle 3.1: Arbeitsprogramm

invariantly specified. Once the direction that the perceptual parameter is moving towards is known, its acoustic and articulatory targets can be predicted by a general modelling algorithm (in case of DIVA an unsupervised neural network).

Signal Processing: An important problem from signal processing is the prediction of a signal $d(n)$ from the samples $x(n), x(n-1), \ldots$ of another signal. In other words, $d(n)$ is specified incrementally by $x(n), x(n-1), \ldots$. Another example is the low rank approximation of a correlation matrix. It relies on the observation that a few components of the correlation matrix is sufficient to represent or approximate the original random process.

Statistics: One central topic in the Bayes theory is the relationship between random events. Some random events are closely related. It is thus possible to describe one event by others. Other random events are independent to each other. In this case, there is no way to specify one event by others.

Information theory: A direct application of the Bayes theory can be found in the information theory. Here the information amount of a random event is measured by entropy which reflects its uncertainty. The information amount of an event which is new and not contained in other events is called conditional entropy. Only this innovative part of the information contributes really to the incremental specification.

Mathematics: Given a vector in a higher dimensional space. One problem is to approximate that vector in terms of an increasing number of other vectors. The so called Gram-Schmidt orthogonalization (Golub & Van Loan, 1983) solves this problem efficiently by using the innovation approach. The same method can be applied to function approximation.

In all areas above except phonetics and phonology, there exist well studied computational procedures for incremental specification. We will identify their common principles and possible restrictions. The results of this research will form a good starting point for incremental specification in speech (possibly after modifications and extensions).

AP1.2: Incremental specification in speech processing

The principle of incremental specification is not completely new to speech representation. Some speech processing algorithms do have an incremental specification component. Two such examples are described below, both for speech compression (Markel & Gray, 1976; Spanias, 1994; Vary et al., 1988).

The first example is a waveform coder based on adaptive differential pulse code modulation (ADPCM). This speech coder is used in the cordless telephone (DECT) and reduces the speech data rate from 64 kb/s to 32 kb/s. Since the speech signal is correlated, the current sample $x(n)$ of the speech is predicted by a linear combination of the past samples $x(n-1), x(n-2), \ldots$ of the same signal. The speech compression is achieved by transmitting the prediction error signal instead of the original signal $x(n)$. Since the prediction errors have a much smaller amplitude than the original samples, we need less bits for coding. At the receiver side, the original signal is reconstructed from the transmitted prediction error signal without any perceptional degradation of the speech quality. In this example, the original signal $x(n)$ is incrementally specified by a linear prediction of its own past samples.

The second example is a more advanced speech compressing technique, vocoder (voice coder and decoder). While ADPCM is a nonparametric approach, a vocoder relies on a pa-

rametric model of the speech signal: the glottal excitation signal formed by the vocal tract. While the vocal tract determines the major part of the linguistic content of the speech, the glottal excitation contributes mainly to the speaker characteristics. During the coding process, the speech signal undergoes an LPC (linear predictive coding) analysis resulting in an estimated vocal tract and an estimated glottal excitation. Both the estimated vocal tract and part of the estimated glottal excitation are transmitted. This kind of vocoders has been widely used in mobile phones, internet telephony, MPEG etc. They achieve a higher speech compression rate than ADPCM. Here the incremental specification of speech consists of two steps: specification of the linguistic content by the vocal tract and specification of the speaker specific features by the glottal excitation. Of course, this is rather a simple model. Recent studies have indicated that vocal tract and glottal excitation are not completely independent. While individual variations of the vocal tract do contribute to the speaker characteristics, the glottis also produces relevant prosodic features from the linguistic point of view.

In this subpackage, we will look at these speech processing algorithms and study them more detailed from the view point of incremental specification.

AP1.3: The innovation approach

The process of incremental specification is not unique. For a given object, there exist many different ways of specification. Some of them are costly and some others are more efficient in terms of information amount and computational effort.

An interesting observation is that many incremental specifications from statistical signal processing, information theory, and mathematics are characterized by the so called *innovation approach* (Kailath, 1970; Kailath et al., 2000). Its basic idea is to provide a sequence of innovations at each step of the specification process. By innovation we understand completely new (sometimes called independent or orthogonal) information which was not available before. In other words, the additional context information i_n and the previous specification τ_{n-1} in Fig. 3.1 have no part in common. In geometry, innovation describes orthogonal vectors. In information theory, innovation means independent random events and variables. In signal processing, the idea of innovation is expressed by the *orthogonality principle* (Haykin, 1995).

One big advantage of the innovation approach is the zero information redundancy. It makes the incremental specification computationally efficient. It also facilitates the context evaluation at each specification step. We will review systematically all innovation approaches from different areas. We will identify their common concepts and limitations. Is this approach applicable and extensible to speech? The key question will be how to define the metric "orthogonal" for speech representations in different contexts. In this issue our project supplements project D5 "Biased learning for syntactic disambiguation" which looks for a "similarity metric" (i.e. non-innovative) for disambiguation. From the information theoretical point of view, the joint entropy (full specification) of the observations $X_1, X_2, X_3, \ldots$ can be incrementally specified by a sequence of conditional entropies (innovations)

$$H(X_1, X_2, X_3, \ldots) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots$$

This general principle has been successively applied in various language and speech processing methods like grammar modelling and hidden Markov model based speech recognition.

AP1.4: Incremental specification and incremental abstraction

Incremental specification is a sequential process of providing information gain. It is sometimes called synthesis. Another interesting process is incremental abstraction in the reversed direction, also referred to as analysis. Starting from a full specification of an object, we want to extract a sequence of useful features which represent meaningful underspecifications of that object, see Fig. 3.3.

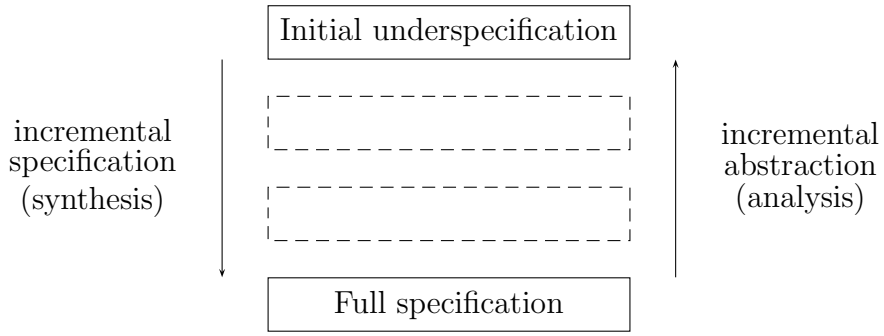

Abbildung 3.3: Incremental specification and incremental abstraction

We believe that both processes are closely related and can not be treated separately. The goal of this subpackage is to study their interactions, both in general and with respect to the particular examples in AP1.1 and AP1.2. We will attempt to provide formal rigor to concepts like "analysis-by-synthesis". This will greatly improve our understanding of the incremental specification process.

AP1.5: Static and dynamic incremental specification

By static incremental specification we understand such a process for a fixed full specification. It does not change in time and with respect to other contexts. In contrast, a dynamic incremental specification is necessary if the full specification itself is changing. A simple example is when we continuously collect information about an object and the information amount is growing in time.

In speech processing, frame based processing is usually considered as static. How can we represent a fixed speech frame by a few distinctive features and how can we improve this underspecification by adding more details if necessary (cf. AP2)? When we consider speech as a continuous stream of information, the specification process becomes dynamic and adaptive since the speech characteristics, both linguistic and speaker specific, are time varying. Interestingly, the two terms static and dynamic are relative. Inside a static process like a speech frame, a dynamic incremental specification is possible (but probably not meaningful). On the other hand, a dynamic process at a global time scale often consists of many static processes at a local time scale. Thus the terms "static" and "dynamic" depend strongly on the time scale and is closely related to the issue of local and global context. In signal processing, we use the words "estimation" and "tracking" to reflect the static and dynamic situation.

Clearly, we have to deal with different incremental specifications in both cases. But do they share with some common approaches? What is their relationship and interaction?

How does the innovation approach look like in a static and dynamic context?

AP1.6: Context dependent incremental specification

Another important aspect in incremental specification of speech is its dependence on context. As we already mentioned, the process of incremental specification is not unique. It strongly depends on what we want to do with the specification. Even if the purpose of a specification is given, there still exist many different ways. An illustrative example is the speech compression discussed in AP1.2. For the same purpose of data reduction in speech communication, two different principles of speech compression have been applied resulting in different data rates and speech qualities. The purpose of incremental specification in context is to foster perception. The broader and the clearer the context information becomes, the easier it is to recover perceptual models underlying speech sounds and prosodic units.

Context has another meaning in the sense of neighborhood information. Semantic information, for example, simplifies the analysis and synthesis at the acoustic, phonetic, and prosodic level and helps to resolve ambiguities. Speech recognition is such an example. We believe that semantic information also plays an important role for the speaker specific features, see AP3 for more details.

In this subpackage we will study the impact of context to the speech representation. One key question will be how to represent context information. In particular, we are looking for ways how to incorporate the impact into the general computational procedure of incremental specification. A close cooperation with projects A2 and D5 will enhance the results of this study.

AP2: Incremental specification of perceptually critical regions of speech

This work package develops perceptually critical regions of speech and studies their contextual aspects. The perceptually critical regions of speech are thought as extensions to landmarks and pivots. Pivots are an abstract idea with currently no implementations available. Landmark detectors exist and will be improved.

AP2.1: Implementation of landmark detectors

Our implementation of landmark detectors will be completed. Not all types of landmarks defined so far were studied in our preparatory works. The completed implementation will be used to gain baseline results on the labelled speech data already available.

The phonation parameter estimators from our preparatory works will be merged into the processing framework.

AP2.2 Study of these landmark detectors on phonemes and syllables

The task of information extraction from a given speech signal in a given phonetic and prosodic context will be studied now. If no context is assumed, the situation is similar to a free running speech recognizer, and the information extracted are phoneme and word hypotheses together with hypotheses about word and phrase stress. If, e.g. the speaker is

known, models specialized to him will be used. These will usually yield hypotheses with increased confidence measure.

The studies will start with standard labelled speech corpora for German and English and are easily extensible to speech and context material results from other projects of this SFB including A1, A2. Cooperation with B, C and D projects may be established on the basis of the study of a particular articulatory or phonatory phenomenon but are not planned in detail now.

AP2.3 Generalization of the landmark detector

The different landmark detection schemes will now be unified and extended to become a statistical model of time-frequency patterns. Using the speech corpora, the observations will train context dependent models that collect the patterns that occur at and in the vicinity of the different landmark types. By taking into consideration the phonetic and prosodic context, conditional models will be trained. Besides the positive part the model will be constructed to contain an negative part for the observable structures of other landmarks that do not occur with the actual one.

The baseline results should be reproduced. In addition, the improvements of conditional models will be documented. If, e.g. the correct speaker specific model is used, the landmarks may be detected either with less error or under stronger disturbances.

The generalized detector will work on both the time-frequency structures, prosodic, and voice quality parameters as mentioned in 2.1. This allows for the study of simultaneous effects in phonation and articulation of word and phrase stress.

AP2.4 Apply this generalized detector to critical regions

The linguistic information called distinctive features is transmitted over the acoustic channel and we focussed on the structures localized by landmark detectors. It is not unlikely that during the definition of the landmark detectors we restricted our focus too much to small regions. The generalized detector is not restricted to predefined regions and may be trained to find its relevant regions by supervised learning. By presenting as many audio examples as possible for distinctive features we investigate the ability of our detector to build more general acoustic models for this feature. We intend to use the name *critical regions* to the resulting time-frequency-parameter models.

AP2.5 Structured critical regions model

Having connected the notion of linguistic information transmitted over articulatory and phonatory channels of the acoustic production-perception path, we want to study the information increment under different phonetic and prosodic contexts. The critical regions model is made conditional by the context. The resulting collection of context dependent models is the structured critical regions model.

AP2.6 Dependency of the structured critical regions model on word stress, phrase stress, and interpersonal context

The structured critical regions model will be used to study joint articulatory and phonatory effects of stress and interpersonal context systematically. The known main correlate of word stress is duration, but changes in voice quality are not unreasonable because the perceived intensity increase may be produced via this channel. Phrase stress is already linked to pitch contours, hence voice quality might change systematically too. Finally voice quality was found to be significantly linked to the role as instruction giver versus instruction follower (Classen et al., 1998) which might perhaps best be studied by a new set of map task experiments or related dialog recordings.

AP3.1: Speaker specific features and voice quality parameters

Researching paralinguistic features of speech is challenging in several aspects. First, the spectrum of these features is very broad. Even in the research community, it is not always agreed upon what the different categories are and which aspects they include. Traunmüller (2000), for example, classified speaker specific features into the following categories: organic (age, gender, pathology, ...), expressive (emotion, attitude, adaption to environment, ...), and perspectival (place, orientation, channel, ...). Other people used other classifications (Schötz, 2002). Second, even for an agreed category like emotion, there is no general agreement on how to define different emotions (anger, happiness, sadness, fear, ...) precisely. In contrast to speech recognition where there is a text reference, there is no exact reference for the speaker specific features. Third, even human have difficulties to recognize speaker specific features correctly. Finally, there is a continuous debate as how to seek measurable acoustic correlates of speaker specific features.

In this work package, we focus on these speaker specific features as far as they can be coded by the voice quality parameters (Stevens & Hanson, 1998; Wokurek, 2003; Lugger & Yang, 2006): open quotient gradient (OQG), glottal opening gradient (GOG), skewness gradient (SKG), rate of closure gradient (RCG), and incompleteness of closure (IC). They are closely related to the glottal excitation and describe well the speaker characteristics. As mentioned in section 3.3.2, we demonstrated the correlation of these glottal parameters to different types of speaker specific features like gender, health, word stress, emotion, and voice quality. But the incremental specification of these features by taking the context into account has not been addressed. It is the goal of this work package.

We first review the concept of voice quality parameters. Note that they can be both estimated from the time domain and frequency domain. Then we study which speaker specific features can be well represented by the voice quality parameters.

AP3.2: Context dependent changes of speaker specific features based on the MapTask database

We will start with our present MapTask database (Classen, 2000, 2005) to study the changes of the voice quality parameters in context. This database consists of 16 speakers and 982 utterances. It is annotated for context already and is phonetically and prosodically labelled throughout. In this database, one speaker is the instruction giver and the other one is the instruction follower. It is to be expected that the voice quality parameters of

the instruction giver would differ from those of the instruction follower. We will study how does the role of the speaker in the dialog influence the acoustic voice quality parameters and how can we model this dependence.

AP3.3: Creation of a project specific database

Since the context in the MapTask database is limited to the role of the speaker in a dialog only, we plan to record more acoustic signals with predefined contexts and create a project specific database. A discussion with other projects of this SFB (e.g. project A1) will help us to determine which kind of context information we will additionally take into account.

AP3.4: Context dependent changes of speaker specific features based on the new database

Based on the new database, we will again study the changes of the voice quality parameters in context as in AP3.2.

3.5 Stellung innerhalb des Sonderforschungsbereichs

3.5.1 Stellung zum Gesamtkonzept des SFBs

This project makes three contributions to the whole SFB:

- How to derive highly underspecified speech representations from the acoustic wave
- How to develop a mathematically sound framework for incremental specification to include phonetic and phonological context incrementally
- How to apply statistical approaches to numerical data to model underspecification and specification processes in general

While the first contribution targets immediately at the acoustical and phonological level of incremental specification in context, the second focus of the project should provide a general computational environment in the A area of the SFB. We hope and speculate that some of the results could also be useful for incremental specification in context at even higher linguistic levels in area B and C. These cooperations will be sought and established after our framework of incremental specification is developed in the phonetic and phonological context. This might be project B1: The formation and interpretation of derived nominals (Alexiadou, von Heusinger) and C1: The syntax of nominal modification and its interaction with nominal structure (Alexiadou).

The application of our mathematical theory of incremental specification of speech will be context dependent. In A3 this is mainly phonetic and prosodic context. The other A-projects allow the test of the theory in the context of exemplars A2 as well as prosody A1. The exemplar based models of A2 are very likely to require complex boundary regions between exemplar regions similar to the complex regions of acoustic events and features. Prosody introduces a symbolic context that is expressed in acoustic events on a wider time scale compared to speech sounds.

3.5.2 Interaktion mit anderen Teilprojekten

From the point of view of content our project has strongest connections to the projects of the area A of the SFB. A3 will provide a formal basis for incremental specification in speech. Speech phenomena are studied in Projects A1 (prosody) and A2 (exemplar theory). A1 (Incremental specification of Focus and Giveness in a Discourse Context) will develop a corpus of context dependent prosodic events. The corpus data will be annotated with both semantic and prosodic labels (pitch-accents, focus, giveness, availability etc). The incremental specification model developed in our project will be tested on this corpus. A2 (Exemplar-based speech representation) will specify the interactions between under-specified and fully specified forms of representation from linguistic phonetic point of view. A3 will incorporate these theoretical insights into the computational model of incremental speech specification/abstraction.

From the formal and methodological point of view our project has the strongest connections to the projects of the area D of the SFB. A3 will apply methods from mathematics, information science and statistical signal processing to develop a computational model of specification and abstraction processes. Most of the projects of the area D have a goal of developing a computational model and many use methods of mathematics and information science. D5 (Biased learning for syntactic disambiguation) uses statistical approaches like neighborhood density (kNN) to model linguistic data. D5 attempts to develop a formal model of similarity. The methods and goals are formally similar to the innovation approach scheduled in our project. A fruitful discussion of the merits of these two formally related approaches to modelling numerical and symbolic aspects of language is expected. D4 (Modular Lexicalization of Probabilistic Context-Free Grammars) uses statistical models trained on large annotated corpora (tree-banks) for parsing, and mathematical methods (like entropy) to solve data sparsity problems. Similar formal mechanisms are employed in the incremental specification of speech model. The interaction between the two projects is expected to lead to a higher degree of methodological precision. D3 (Inferenzen auf unterspezifizierten Strukturen zur Informationsextraktion) combines symbolic and statistical methods for information extraction from a specified corpus. Particularly important in this project is the interaction between linguistic and speaker specific types of information. In our project a similar problem of interaction between linguistic phonetic and speaker specific information is encountered. The modelling of such interactions of information structures with different granularity of specification in both areas is expected from the interaction of the projects.

It is also expected that some projects of the B and C area which work with corpora might test the model of incremental specification on their symbolic data once this model has been formally specified. This applies in particular to the project B1 (The formation and interpretation of derived nominals) and C1 (The syntax of nominal modification and its interaction with nominal structure.

3.6 Abgrenzung gegenüber anderen geförderten Projekten der Teilprojektleiter

The following tables summary the other projects of the project leader and describe their differentiation to this SFB project.

Titel	Hochauflösende Radargruppenantenne für verbesserte Umfelderfassung in aktiven Fahrerassistenzsystemen
Projekleiter	B. Yang
Zuwendungsgeber	BMBF
Abgrenzung	Not related to the proposed SFB project

Titel	Robuste Verfahren zur Schätzung der Stimmqualität mit Mikrofongruppen
Projekleiter	B. Yang and W. Wokurek
Zuwendungsgeber	Land Baden-Württemberg
Abgrenzung	This project focuses on the estimation of voice quality parameters from microphone array speech signals and their applications. It consists of three parts: preprocessing of the microphone array signals in order to reduce the acoustic disturbances to the speech (noise, reverberation, ...); estimation of the voice quality parameters; classification. This is an application driven project focused on detections. It is not related to the proposed SFB project.

Titel	Brain mechanisms subserving the generation and implementation of "phonetic plans" during speech production: Clinical and functional-imaging investigations.
Projekleiter	G. Dogil and H Ackermann and W. Ziegler
Zuwendungsgeber	DFG (DO 536/5-1)
Abgrenzung	Not related to the proposed SFB project

Titel	SmartWeb
Projekleiter	G. Dogil
Zuwendungsgeber	BMBF
Abgrenzung	Not related to the proposed SFB project

3.7 Ergänzungsausstattung für das Teilprojekt

	2006/2			2007			
PM	Verg.- Gr.	Anz.	Betrag EUR	Verg.- Gr.	Anz.	Betrag EUR	Verg.- Gr.
	BAT IIa			BAT IIa			BAT IIa
	StHi			StHi			StHi
	zus.:			zus.:			zus.:

	Kostenkategorie oder Kennziffer	Betrag EUR	Kostenkategorie oder Kennziffer	Betrag EUR	Kostenkat
SM	Kleingeräte		Kleingeräte		Kleingerät
	Verbrauchsmaterial		Verbrauchsmaterial		Verbrauch
	Reisen		Reisen		Reisen
	Sonstiges		Sonstiges		Sonstiges
	zus.:		zus.:		

	Investitionsmittel insges.	Investitionsmittel insges.	
IM	Keine	Keine	

3.7.1 Personal im Teilprojekt

LSS: Lehrstuhl für Systemtheorie und Signalverarbeitung
IMS: Institut für maschinelle Sprachverarbeitung

Grundausstattung

	Name, akad. Grad, Dienststellung	engeres Fach des Mitarbeiters	Institut der Hochschule oder der außeruniv. Einrichtung	Beantragte Förderperiode: Mitarbeit im Teilprojekt in Std./Woche (beratend: B) B	2006/2	2007	2008	2009	2010/1	Vergütungsgruppe
3.7.1.1 wissenschaftl. Mitarbeiter (einschl. Hilfskräfte)	1. Yang, Bin, Prof. Dr.-Ing.	Signalverarbeitung	LSS		5	5	5	5	5	C4
	2. Dogil, Grzegorz, Prof. Dr.,	Phonetik	IMS		5	5	5	5	5	C4
	3. Wokurek, Wolfgang, Dr.	Phonetik	IMS		5	5	5	5	5	BAT Ib

Ergänzungsausstattung

	Name, akad. Grad, Dienststellung	engeres Fach des Mitarbeiters	Institut der Hochschule oder der außeruniv. Einrichtung	Beantragte Förderperiode: Mitarbeit im Teilprojekt in Std./Woche (beratend: B) B	2006/2	2007	2008	2009	2010/1	Vergütungsgruppe
3.7.1.3 wissenschaftl. Mitarbeiter (einschl. Hilfskräfte)	1. N.N. (X)	Signalverarbeitung	LSS		41	41	41	41	41	BAT IIa
	2. N.N. (X)	Phonetik	IMS		41	41	41	41	41	BAT IIa
	3. N.N. (X)	Signalverarbeitung	LSS		0	15	20	20	20	Std. Hiwi
	4. N.N. (X)	Phonetik	IMS		10	15	20	20	20	Std. Hiwi

Aufgabenbeschreibung von Mitarbeitern der Ergänzungsausstattung für die beantragte Förderperiode

zu 1: N.N. Bearbeitung der Arbeitspakete AP1 und AP3
zu 2: N.N. Bearbeitung des Arbeitspakets AP2
zu 3: N.N. stud. Hilfskräfte für Arbeitspakete AP1 und AP3
zu 4: N.N. stud. Hilfskräfte für Arbeitspaket AP2

Begründung für die vollen BAT-IIa-Stellen: Die Mitarbeiter, die mit den Aufgaben im geplanten Projekt betraut werden sollen, müssen im schwierigen Schnittbereich zwischen experimenteller Phonetik und statistischer Signalverarbeitung eigenverantwortlich forschen. Da alle Stelleninhaber Absolventen von Elektrotechnik und Informationstechnik mit Schwerpunkt auf statistische Signalverarbeitung oder auf diesem Gebiet sehr gut vorgebildete Computerlinguisten sein sollen, sehen die Antragsteller keine Möglichkeit, Kandidaten für halbe Stellen zu gewinnen. Solchen Kandidaten werden ausnahmsweise ganze Stellen für Forschung und Promotion angeboten.

Begründung für die studentischen Hilfskräfte: a) Hilfe bei der Durchführung von Sprachaufnahmen mit Probanden b) Erstellung von Software für Messungen und Auswertungen c) Auswertung von Meßdaten d) Hilfe bei der Durchführung von Computersimulationen

3.7.2 Aufgliederung und Begründung der Sachmittel (nach Haushaltsjahren)

	2006/2	2007	2008	2009	2010/1
Für Sächliche Verwaltungsausgaben stehen als **Grundausstattung** voraussichtlich zur Verfügung:	5500	1000	1000	1000	500
Für Sächliche Verwaltungsausgaben werden als **Ergänzungsausstattung** beantragt (entspricht den Gesamtsummen „Sächliche Verwaltungsausgaben“ in Übersicht 3.8):	0	0	0	0	0

(Alle Angaben in EUR)

Kleingeräte

Bezeichnung	2006/2	2007	2008	2009	2010/1
1. Rothenberg-Maske	9573 EUR				
2. Laryngograph		5118 EUR			

zu 1 und 2: Both devices are from Glottal Enterprises, Inc., USA. They will be both used in the recordings of AP2.1 and AP2.2 as well in AP2.7 to get separate recordings of the oral and nasal sound pressure and volume velocity and to the electroglottogramm.

Reisen

Bezeichnung	2006/2	2007	2008	2009	2010/1
1.	0	3000	3000	3000	3000

zu 1: je eine internationale Konferenzreise vom LSS und IMS im Jahr zu 1500 EUR

Sonstiges

Bezeichnung	2006/2	2007	2008	2009	2010/1
1. Honorar für Sprecher in AP2		500	500	500	500
2. Honorar für Sprecher in AP3.4			200		

zu 1: Honorar für Sprachaufzeichnungen in AP2 (10 Personen mit je 5 Stunden zum Stundenlohn von 10EUR)

zu 2: Honorar für Sprachaufzeichnungen in AP3.4 zur Untersuchung der sprecherabhängigen Merkmale (4 Personen mit je 5 Stunden zum Stundenlohn vin 10EUR).

3.7.3 Investitionen (Geräte über 10.000,- EUR brutto und Fahrzeuge)

Keine

Literaturverzeichnis

Archangeli, D., 1988. Aspects of underspecification theory. *Phonology* **5**, 183–207.

Batliner, A., et al., 2000. *The Recognition of Emotion.* In: Wahlster, W. (ed.), *Verbmobil: Foundations of Speech-to-Speech Translations.* Springer, 122–130.

Byrd, D., 1996. A phase window framework for articulatory timing. *Phonology* **13**, 139–169.

Carson-Berndsen, J., 2000. Finite-state models, event logics and statistics in speech recognition. *Phil. Trans. R. Soc. Lond. A* **358**, 1255–1266.

Classen, K., 2000. *MapTask - Eine Version für das Deutsche, AIMS 6, 65-83.*

Classen, K., Dec. 2005. *Prosodische und dysprosodische Variation linguistischer und paralinguistischer Funktionen im Spontansprachlichen Dialog.* Ph.D. dissertation, IMS, University Stuttgart.

Classen, K., Dogil, G., Jessen, M., Marasek, K., Wokurek, W., 1998. *Stimmqualität und Wortbetonung im Deutschen.* In: Linguistische Berichte 174. 202–245.

Classen, K., Wokurek, W., 2003. *Voice quality and dialogue structure.* In: Proceedings of the 15th International Congress of Phonetic Sciences. 2169–2172.

Clements, G. N., 1987. *Phonological feature representation and the description of intrusive stops.* In: Bosch, A., Need, B., Schiller, E. (eds.), *CLS 23: Parasession on autosegmental and metrical phonoilogy.* Chicago: CLS, 29–50.

Cohn, A., 1990. Phonetic and phonological rules of nazalization. *UCLA Working Papers in Phonetics* **76**.

Cohn, A., 1993. Nasalisation in English. *Phonology* **10**, 43–81.

Cover, T. M., Thomas, J. A., 1991. *Elements of Information Theory.* Wiley.

Cowie, R., et al., 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* **18**, 32–81.

Dogil, G., 1986. Phonological pivot parsing. *Proceedings of the 11th coference on Computational linguistics (COLING) Bonn*, 615–617.

Dogil, G., 1988. *The Pivot model of speech parsing.* Verlag der Österreichischen Akademie der Wissenschaften.

Dogil, G., Möbius, B., 2001. Toward a perception based model of the production of prosody. *Journal of the Acoustical Society of America* **110** (5, Pt. 2), 2737.

Eisner, J., 2000. *Directional constraint evaluation in optimality theory.* In: Proc. of 18th conf. of Computational Linguistics. 257–263.

Fach, M., Wokurek, W., 1995. Pitch accent classification of fundamental frequency contours by hidden markov models. *Proceedings of the EUROSPEECH'95, Madrid* **3**, 2047–2050.

Goldinger, S., 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology, Learning, Memory and Cognition* **22**, 1166–1183.

Goldsmith, J., 1976. *Autosegmental phonology*. Ph.D. dissertation, MIT.

Golub, G. H., Van Loan, C. F., 1983. *Matrix Computations*. Johns Hopkins Press, Baltimore.

Guenther, F. H., Hampson, M., Johnson, D., 1998. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review* **105**, 611–633.

Haderlein, T., Nöth, E., 2003. *The EMBASSI Speech Corpus*. University Erlangen-Nürnberg.

Haykin, S., 1995. *Adaptive Filter Theorey*, 3rd Edition. Prentice-Hall, Englewood Cliffs.

Huber, R., 2002. *Prosodisch-linguistische Klassifikation von Emotionen*. Logos Verlag, Berlin.

Kailath, T., 1970. The innovations approach to detection and estimation theory. *Proc. IEEE* **58**, 680–695.

Kailath, T., Sayed, A. H., Hassibi, B., 2000. *Linear Estimation*. Prentice-Hall, Englewood Cliffs.

Karttunen, L., 1998. *The proper treatment of optimality in computational phonology*. In: Proc. of FSMLP 98. 1–12.

Kavčić, A., Yang, B., 1996a. Adaptive rank estimation for spherical subspace trackers. *IEEE Trans. Signal Processing* **44**, 1573–1579.

Kavčić, A., Yang, B., 1996b. Subspace tracking with adaptive threshold rank estimation. *Special Issue on Array Optimization and Adaptive Tracking Algorithms, Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology* **14**, 75–91.

Kay, S. M., 1993. *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*. Prentice-Hall, Englewood Cliffs.

Keating, P. A., 1988. Underspecification in phonetics. *Phonology* **5**, 275–292.

Kiparsky, P., 1985. Some consequences of lexical phonology. *Phonology Yearbook* **2**, 85–138.

Lahiri, A., Marslen-Wilson, W., 1991. The mental representation of lexical form: A phonological approach. *Cognition* **38**, 245–294.

Lahiri, A., Reetz, H., 2002. *Underspecified recogntion*. In: Gussenhoven, C., Werner, N. (eds.), *Laboratory Phonology*. Vol. 7. Mouton, 637–675.

Lee, C. M., Narayanan, S. S., 2005. Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech and Audio Processing* **13**, 293–303.

Liu, S. A., 1996. Landmark detection for distinctive feature-based speech recognition. *JASA* **100**, 3417–3430.

Lugger, M., Yang, B., 2006. *Klassifikation verschiedener Sprechergruppen mit Hilfe von Stimmqualitätsparametern*. In: 7. ITG-Fachtagung Sprach-Kommunikation.

Lugger, M., Yang, B., Wokurek, W., 2006. *Robust estimation of voice quality parameters ander real world disturbances*. In: Proc. IEEE ICASSP.

Markel, J. D., Gray, Jr., A. H., 1976. *Linear Prediction of Speech*. Springer-Verlag, New York.

McCarthy, J., Taub, A., 1992. Review of "the special status of coronals: Internal and external evidence", by Carole Paradis and Jean-Francois Prunet. *Phonology* **9**, 363–370.

Petrushin, V. A., 2000. *Emotion Recognition in Speech Signal: Experimental Study, Development, and Application*. In: Proc. Int. Conf. on Spoken Language Processing.

Pierrehumbert, J., 2001. *Exemplar dynamics: Word frequency, lenition and contrast*. In: Bybee, J., Hopper, P. (eds.), *Frequency and the Emergence of Linguistic Structure*. Benjamins, 137–157.

Pierrehumbert, J. B., Beckman, M. E., 1988. *Japanese tone structure*. MIT Press.

Pützer, M., Wokurek, W., 2006. Multiparametrische stimmprofildifferenzierung zu männlichen und weiblichen normalstimmen auf der grundlage akustischer analysen. *Laryngol Rhino Otol*, im Druck.

Reetz, H., 1999. Converting speech signals to phonological features. *ICPhS*, 1733–1736.

Ringen, C., 1975. *Vowel harmony: Theoretical implications*. Ph.D. dissertation, Indiana University.

Scharf, L. L., 1991. *The SVD and Reduced-Rank Signal Processing*. In: Vaccaro, R. J. (ed.), *SVD and Signal Processing II: Algorithms, Analysis and Applications*. Elsevier Science Publishers B. V., North-Holland, 3–31.

Scheuing, J., Yang, B., 2006. *Disambiguation of TDOA Estimates in Multi-Path Multi-Source Environments (DATEMM)*. In: Proc. IEEE ICASSP.

Schötz, S., 2002. *Linguistic and paralinguistic phonetic variation in speaker recognition and text-to-speech synthesis*. www.ling.lu.se/persons/suzi.

Shannon, C. E., 1948. A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423,623–656.

Spanias, A. S., 1994. Speech coding: A tutorial review. *Proc. IEEE* **82**, 1541–1582.

Steriade, D., 1987. Redundant values. *CLS* **23** (2), 339–362.

Stevens, K. N., 1989. On the quantal nature of speech. *Journal of Phonetics* **17**, 3–45.

Stevens, K. N., 1998. *Acoustic phonetics*. MIT Press,, Cambridge, MA.

Stevens, K. N., 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *JASA* **111**, 1872–1891.

Stevens, K. N., 2005. *Features in speech perception and lexical access*. In: Pisoni, D. B., Remez, R. E. (eds.), *The Handbook of Speech Perception*. Blackwell, Oxford, UK, 125–155.

Stevens, K. N., Hanson, H. M., 1998. *Classification of glottal vibration from acoustic measurements*. In: Fujimura, O., Hirano, M. (eds.), *Vocal Fold Physiology*. Hiltop University Press, Cambridge MA, 147–170.

Stevens, K. N., Keyser, S. J., Kawasaki, H., 1986. *Toward a phonetic and phonological investigation of redundant features*. In: Perkell, J., Klatt, D. H. (eds.), *Symposium on invariance and variability of speech processes*. Lawrence Erlbaum, Hillsdale, NJ, Ch. 20, 426–463.

Traunmüller, H., 2000. *Evidence for Demodulation in Speech Perception*. In: Proc. Int. Conf. on Spoken Language Processing. 790–793.

Trubetzkoy, N., 1939. *Grundzüge der Phonologie*. Travaux du Cercle Linguistique de Prague VII.

Vary, P., Heute, U., Hess, W., 1988. *Digitale Sprachsignalverarbeitung*. B. G. Teubner, Stuttgart.

Vollmer, K., 1997. *Koartikulation und glottale Transparenz*. Doctoral dissertation, University of Stuttgart.

Wokurek, W., 1995. A noise-robust subspace-based sound-class detector. *Proceedings of the International Conference on Phonetic Sciences ICPhS'95, Stockholm* **4**, 320–323.

Wokurek, W., 1997. Time-frequency analysis of the glottal opening. *International Conference on Acoustics, Speech, and Signal Processing ICASSP'97, Munich* **2**, 1435–1438.

Wokurek, W., 1999. Corpus based evaluation of entropy rate speech segmentation. *Proceedings of the International Conference on Phonetic Sciences ICPhS'99, San Francisco* **2**, 1217–1220.

Wokurek, W., 2003. *Automated Corpus Based Spectral Measurement of Voice Quality Parameters.* In: Proceedings of the 15th International Congress of Phonetic Sciences. 2173–2176.

Wokurek, W., Hlawatsch, F., Kubin, G., 1987. Wigner distribution analysis of speech signals. *Int. Conf. on Digital Signal Processing, Florence*, 294–298.

Wokurek, W., Pützer, M., 2003. Automated corpus based spectral measurement of voice quality parameters. *Proceedings of the 15th International Congress of Phonetic Sciences (Barcelona)*, 2173–2176.

Yang, B., 1991. *Rotationen verwendende Parallelalgorithmen für hochkomplexe Signalverarbeitung und systolische Rechenfelder.* VDI-Verlag, Düsseldorf.

Yang, B., 1995a. An extension of the PASTd algorithm to both rank and subspace tracking. *IEEE Signal Processing Letters* **2**, 179–182.

Yang, B., 1995b. Projection approximation subspace tracking. *IEEE Trans. Signal Processing* **43**, 95–107.

Yang, B., 1995c. A QR multichannel least squares lattice algorithm for adaptive nonlinear filtering. *AEÜ International Journal of Electronics and Communications* **49**, 171–182.

Yang, B., 1996. Asymptotic convergence analysis of projection approximation subspace tracking algorithms. *Signal Processing, Special Issue on Subspace Methods, Part I: Array Signal Processing and Subspace Computations, eds. M. Viberg and P. Stoica* **50**, 123–136.

Yang, B., Böhme, J. F., 1992. Rotation based RLS algorithms: Unified derivations, numerical properties and parallel implementations. *IEEE Trans. Signal Processing* **40**, 1151–1167.

Yang, B., Scheuing, J., 2005. *Cramer-Rao Bound and Optimum Sensor Array for Source Localization From Time Differences of Arrival.* In: Proc. IEEE ICASSP. Vol. 4. 961–964.

Yang, B., Scheuing, J., 2006. *A Theoretical Analysis of 2D Sensor Arrays For TDOA Based Localization.* In: Proc. IEEE ICASSP.

Zsiga, E. C., 1997. Features, gestures, and Igbo vowels: An approach to the phonology-phonetics interface. *Language* **73**, 227–274.