gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

PHONETIC AND PROSODIC ANALYSIS OF SPEECH

H. Niemann, E. Nöth
E.G. Schukat–Talamazzini
A. Kießling, R. Kompe
T. Kuhn, S. Rieck

F.–A.–Universität Erlangen–Nürnberg

Vm

Report 34
Oktober 1994

Oktober 1994

H. Niemann, E. Nöth
E.G. Schukat–Talamazzini
A. Kießling, R. Kompe
T. Kuhn, S. Rieck

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Tel.: (09131) 85 - 7774
e-mail: {niemann}@informatik.uni-erlangen.de

Gehört zum Antragsabschnitt: 3.11, 3.12, 6.4

Aus B. Horvat and Z. Kacic, MODERN MODES OF MAN – MACHINE COMMUNICATION, International Workshop, June 1994, Maribor, Slovenia, (pp. 12-1 – 12-12).

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 C 6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

PHONETIC AND PROSODIC ANALYSIS OF SPEECH*

H. Niemann, E. Nöth, E.G. Schukat–Talamazzini, A. Kießling, R. Kompe, T. Kuhn, S. Rieck

Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen-Nürnberg, Martensstr. 3
91058 Erlangen, F.R. of Germany

Abstract: In order to cope with the problems of spontaneous speech (including, for example, hesitations and non-words) it is necessary to extract from the speech signal all information it contains. Modeling of words by segmental units should be supported by suprasegmental units since valuable information is represented in the prosody of an utterance. We present an approach to flexible and efficient modeling of speech by segmental units and describe extraction and use of suprasegmental information.

Keywords: speech recognition, hidden Markov models, prosody,

INTRODUCTION

This paper presents an approach towards statistical modeling and use of segmental and suprasegmental information in a speech signal. We treat the aspects of word recognition and improvement of linguistic analysis by suprasegmental information.

Sect. 1 gives an account of acoustic–phonetic analysis in the ISADORA system for word recognition. It will be demonstrated that it is general enough to also include prosodic information. In Sect. 2 we present the extraction and utilization of prosodic cues. Results are summarized in Sect. 3 and a conclusion and outlook are given in Sect. 4.

1 WORD RECOGNITION

1.1 Introductory Remarks

We use a statistical approach to word recognition based on hidden Markov Models (HMM). Early work in recognition of isolated words and continuous speech is, for example, [3, 2, 12]. Recent work on continuous speech recognition is given, for example, in [6, 11, 13, 18, 20].

During word recognition it is tried to segment the speech signal into words and to classify the words under the premise that they belong to a finite set of known words represented in a lexicon. Problems encountered are, for example, variations among speakers, omission of phonemes and even syllables, background noise, speech pauses which may be filled (e.g. by cough, "uh") or unfilled and sometimes fairly long, erroneous transcription of training sentences, and pronunciation variants caused by dialects.

A speaker utters a sequence $\boldsymbol{w}$ consisting of N words w_i out of a given vocabulary W. Depending on the application and the situation this may be, for example, an utterance to be translated to another language or a question of a user to a speech understanding and

**Acknowledgment:* The work reported here has been partly supported by the German Ministry of Research and Technology (BMFT) in the VERBMOBIL Joint Research Project on Speech Understanding and by the European Community in the ESPRIT Project P218 SUNDIAL. This support is gratefully acknowledged; only the authors are responsible for the content of this paper.

dialog system. Acoustic evidence is defined by a set O of L observation symbols O_l. They may be, for example, a set of phones or in general of 'labels'. Observed is a sequence $\boldsymbol{o}$ of T acoustic units $o_i \in O$. They may be, for example, feature vectors or automatically assigned (soft or hard) labels.

The task of word recognition is to find the correct sequence of words uttered by the user. In general the correct sequence can only be approximated and hence one tries to find the *most probable sequence given the observation.* It is well known from statistical decision theory that this approach minimizes the probability of error and in this sense it is an optimal approach. Word recognition then aims at computing the sequence $\boldsymbol{w}^*$ of words having the property

$$\boldsymbol{w}^* = \arg\max_{\{\boldsymbol{w}\}} \left\{ \frac{P(\boldsymbol{o}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{o})} \right\} . \tag{1}$$

The numerator of this equation contains in the first term the *acoustic evidence* ($P(\boldsymbol{o}|\boldsymbol{w})$) and in the second term the *linguistic constraints* ($P(\boldsymbol{w})$) which are represented in a *language model.* The denominator does not affect the maximization because it is independent of $\boldsymbol{w}$; hence, it is ignored.

Parametric representation of the speech signal is by mel–cepstral coefficients, differential coefficients, and energy.

1.2 The ISADORA Network

The intention behind the development of ISADORA (integrated system for automatic decoding of observation sequences of real–valued arrays) was to have a system for experimenting with different speech recognition strategies in order to optimize recognition performance. It provides cepstral feature extraction, hard or soft vector quantization, discrete, continuous, or semi–continuous HMM, beam–search driven training and Viterbi decoding, and a modeling capability encompassing phonetic units, morphemes, words, syntactic constituents, sentences, finite–state grammars, and vocabularies.

In order to model a speech understanding task an inventory of basic HMM's or *A–nodes* or *atoms* is provided together with a set of *model building operators* which build larger HMM's from the atoms. The result of a model building operation is a new node type. A model of a speech recognition task then is written in terms of a definition language summarized in Figure 1.

The **A–nodes** are the elementary units in the network and consist of a dedicated HMM representing the corresponding speech unit. They do not have any successors. An example is the A–node `'/f/'` providing a model of the fricative phone `'f'`. Speech units corresponding to A-nodes are represented by *linear* hidden Markov models. Linear HMM's are left–to–right and consist of a varying number of states, each state being connected to itself and to its immediate successor. Any series of adjacent states can be *tied*, that is, the probabilistic parameters controlling the state transitions and output distributions are pooled. Because A-node HMM's serve merely as starting points in acoustic modelling, the choice of such a limited topology is acceptable. The present implementation allows four different parametrizations of probability density functions (PDF's) determining the output behaviour of states, that is, the discrete PDF, the soft vector quantization version of discrete PDF, the continuous density HMM, and the semi-continuous models or tied mixtures.

An **S–node** defines the *sequential* concatenation of its successor nodes. This provides the means to build, for example, words or compound words from smaller units. A **P–node** represents a set of elements which are the successors of this node. It is realized by the

node type	node name	successors	remark
A:	/f/	;	model of a speech unit no successors
S:	Bahn	/b/ /a:/ /n/ ;	define a word
S:	Bahnhof	Bahn Hof	define a compound word
P:	DIGIT	null eins zwei ... neun	define digits $0 \ldots 9$
P:	<NP>	John Mary Lassie	define proper nouns
S:	<S>	<NP> loves <NP>	define a simple phrase
D:	zwei	/tsvaI/ /tswo:/ ;	alternative pronunciations
R:	DIGITS	DIGIT ;	define connected digits

Figure 1: Some examples of the different node types

$$\begin{array}{c|c|ccccc}
 & & v_1 & \dots & v_j & \dots & v_M \\ \hline
 & & \multicolumn{5}{c}{\underbrace{u_1 \quad \dots \quad u_j \quad \dots \quad u_M}_{exits}} \\
v_1 & s_1 & t_{11} & \dots & t_{1j} & \dots & t_{1M} \\
\vdots & \vdots & & & \vdots & & \\
v_i & s_i & & & t_{ij} & & \\
\vdots & \vdots & & & \vdots & & \\
v_M & \underbrace{s_M}_{entries} & \multicolumn{5}{c}{\underbrace{t_{M1} \qquad \dots \qquad t_{MM}}_{transitions}}
\end{array}$$

Figure 2: Adjacency matrix; if node v_i is an entry node, $s_i = 1$, if there is a transition from node v_i to node v_j, $t_{ij} = 1$, if node v_j is an exit node, $u_j = 1$

parallel connection of the successor nodes. We use the P–nodes, for example, to define the elements of a lexicon or of a syntactic word category. A **D–node** also provides a *parallel* connection of its successors as is done by the P–node, but it represents *alternatives* of the speech unit. An **R–node** defines the arbitrary *repetition* of its *uniqe* successor node. It is implemented by a feed–back loop. The **F–node** allows one to define a finite–state network by appropriate interconnection of successor nodes. The interconnections are defined in an adjacency matrix as shown in Figure 2.

To summarize, each node of an ISADORA network represents a particular concept in the world of speech units. On the other hand, each node can be mapped to an acoustic model in a standard way. The Markov models belonging to the A-type nodes are given explicitly. The models corresponding to any other type of node are given implicitly and can be constructed recursively from the acoustic models of their successor nodes. It is guaranteed that the construction process will eventually halt when encountering the predefined A-type models.

1.3 Subword Units

Three types of subword unit inventories were investigated in more detail: the context–freezing units, the generalized triphones, and the polyphones.

The first approach towards phonetic modelling is to *freeze* the contextual variations of

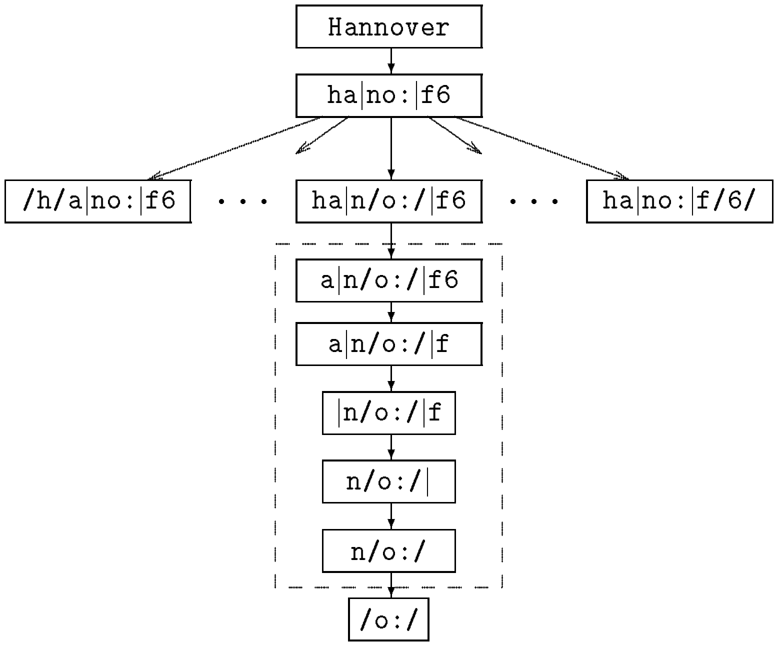

Figure 3: A representation of the word 'Hannover' by polyphones

speech into models for larger-than-phone sized units. These *context-freezing* units (CFU) were described in [21]. For comparison purposes, we also implemented a word modelling scheme based on context-dependent phone-like units, which is well-known as the generalized triphones approach [11].

Finally, the so-called *polyphones* [19] are phone-like units which generalize the well-known concept of *triphone* units. In contrast to a triphone the context information of which is restricted to one phone symbol on the left and one on the right, the core phone of a polyphone may be surrounded by (in principle) arbitrarily large context. Furthermore, the context items may also include suprasegmental markers like syllable, morpheme, or word boundaries. Figure 3 demonstrates how words are represented by polyphones. Whole-word models are included into the polyphone approach, too.

1.4 Decoding

For reasons of space we do not treat the training of HMM parameters. In priciple it is a variant of Baum–Welch training. Two slightly different software packages for word recognition are available. The first one is integrated into the ISADORA environment and exploits the full flexibility offered by it. This means that word recognition is treated as the instantiation of the relevant ISADORA concept and that new recognition tasks may be defined at analysis time, for example, in order to use dialog step dependent vocabularies. The second one is a separate software package. It takes parameters obtained from ISADORA training, it can work with a stochastic bigram grammar, and in view of the task domain it has a spelling mode; it is about a factor of 2.5 faster than the first package. Both packages employ beam search. The second version achieves real–time performance on an HP 735, using 1082 words and a bigram grammar with a perplexity of 111. When using a language model for each category the corresponding lexicon has to be inserted. Finally, to speed up word recognition a beam search algorithm is used. It prunes all states whose score is sufficiently below that of the best scoring state.

2 PROSODY

2.1 Introductory Remarks

Prosody means properties of speech which refer not only to a phone, but to larger units, e.g. a syllable, a phrase, or a whole utterance. Hence, prosodic features are also termed *suprasegmental* features. The main perceptual parameters of prosody are pitch, loudness, duration, and timbre with the acoustical correlates fundamental frequency F_0, energy, length of a phone, and spectral characteristics, respectivley.

The main functions of intonation or prosody in speech and spoken dialogs are to emphasize parts of a word (e.g. 'it is *im*possible to get a cheaper flight'), to mark important words in an utterance (e.g. 'ICASSP 1997 will be in *Munich*'), to delimit (meaningful) parts of an utterance (e.g. 'the father, said the son, is ill' versus 'the father said, the son is ill'), and to differentiate the mood of an utterance (mainly: declarative or question).

Our interest is mainly the determination of sentence mood, focal accent (and thereby focus), and boundaries within an utterance. Such boundaries can be caused by prosodical phrasing or hesitations. It has been shown that prosodic information is an important cue to indicate these properties of an utterance [23, 15].

The most important parameter carrying prosodical information is pitch [1, 15]. The acoustical correlate of pitch is the *fundamental frequency* F_0. Hence, often the fundamental frequency is used as the main or the only acoustic feature for prosodic analysis. This is sufficient, for example, to distinguish interrogative and declarative sentences fairly reliably — a problem which is important in spoken dialog systems. The *fundamental frequency* F_0 of *voiced speech* is determined by the frequency of the oscillating vocal cords. It has to be estimated from the recorded speech signal; an algorithm for doing this is described in Sect. 2.2 below. The fundamental frequency is undefinded for *unvoiced speech*. Problems are caused by irregularities or laryngealizations which may cause, for example, a doubling of the pitch period for about one or two periods.

Early work on prosody is reported, for example, in [7, 10, 17] and recent work is found, for example, in [5, 16, 15, 24, 25].

2.2 Fundamental Frequency

Several algorithms have been proposed for the determination of fundamental frequency, see for example [8]. A main drawback of them is that F_0 is computed only locally without taking into account information about other portions of the utterance. Therefore, these methods usually work well in regular portions of speech, but often fail in irregular portions. We developed a new algorithm for the determination of fundamental frequency contours of speech signals. It is robust even when it encounters irregular portions of speech, and performs well with telephone quality speech.

An overview of the algorithm is given in Figure 4. It was described in detail in [9]. The algorithm is based on the well–known observation that the frequency of the absolute maximum of the short–time spectrum of a voiced speech frame is a harmonic of the fundamental frequency; hence, this frequency divided by the fundamental frequency is an integer. The problem then is to find the correct integer divisor of the frequency of the absolute maximum. This problem is solved here by determining several *candidate values* of the fundamental frequency and to select the (hopefully) correct ones by *dynamic programming* (DP). It is assumed that changes in fundamental frequency between two voiced frames usually are small. One target value per voiced region is estimated to guide the DP search. The DP algorithm searches for the path minimizing the weighted sum of the difference between consecutive candidates plus the distances of the candidates to

1. Preprocessing

Partition the digitized speech signal f_j into *frames* r_k of fixed size. The frames are numbered consecutively by the index $k \in \{0, 1, \ldots, K-1\}$. For each *voiced frame* a value of F_0 is to be determined.

For each frame make a *voiced/unvoiced decision*; Adjacent voiced frames are grouped to a *voiced region* V_l. Each voiced region is defined by an index tupel (l_b, l_e) which gives the frame number of the beginning and end frame, respectively, of V_l. Between two consecutive voiced regions there is at least one unvoiced frame.

Compute the *energy* E_k per frame by the sum of squared amplitude values. A target value (see below) of the fundamental frequency will be computed at a local maximum of E_k within a voiced region.

2. Short–time spectrum

Perform *low–pass filtering* of the speech signal with cut–off frequency of 1100 Hz. Perform a downsampling of the speech signal at a ratio of 1:7 (16 kHz sampling frequency).

Define the sample values in an *analysis window* s_k, corresponding to a frame r_k, by the sequence of sample values in the three frames r_{k-1}, r_k, r_{k+1}.

For each analysis window of a voiced frame compute the absolute value of the *short–time spectrum* S_ν, $\nu = 0, 1, \ldots, 127$.

The *expected interval of fundamental frequency values* is assumed to be $S_{0,min} = 55$ Hz, $S_{0,max} =$ 550 Hz.

3. Target values $F'_{0,l}$

FOR each voiced region V_l (defined by the index tupel (l_b, l_e) and containing frames r_k with speech energy E_k) DO:

- determine a frame r_κ for which a target value of the fundamental frequency is computed as follows:

IF	$l_e - l_b + 1 \leq 5$
THEN	$\kappa = (l_b + l_e + 1)/2$
ELSE	select κ such that $E_\kappa = \max_{k \in [l_b+2, l_e+2]}\{E_k\}$

- Determine for this frame r_κ the fundamental frequency by two independent algorithms. This value of the fundamental frequency is the *target value* $F'_{0,l}$ of the voiced region V_l.

4. Fundamental frequency candidates

FOR all voiced regions V_l, $l = 1, \ldots, L$ DO:

- compute an *average target value* $F'_0 = \frac{1}{L}\sum_{l=1}^{L} F'_{0,l}$.
- FOR all frames r_k, $k \in [l_b, l_e]$, in voiced region V_l DO:
  - determine the maximal value S_{max} and the frequency ξ_{max} of this value in the short–time spectrum S_ν; set the integer divisor $n = \xi_{max}/F'_0$
  - five fundamental frequency candidates $F'_{k,l,a}$ of frame number k in voiced region number l are defined by $F'_{k,l,a} = \{\frac{\xi_{max}}{n+a-3}, \ a = 1, \ldots, 5\}$; a candidate is undefined if $n + a - 3 \leq 0$

5. Fundamental frequency contour

FOR each voiced region V_l DO:

- compute the *optimal path* in the matrix of fundamental frequency candidates by dynamic programming
- the fundamental frequency contour is computed by tracing back the optimal path

Figure 4: The steps for computing the fundamental frequency

a local target value. The path obtained this way is considered to be the fundamental frequency contour.

2.3 Sentence Mood

Our work in speech understanding and dialog treats the task domain of enquiries about intercity train connections. The evaluation of typical dialogs showed that in many cases the user repeats a departure or arrival time stated by the information officer. Repeating

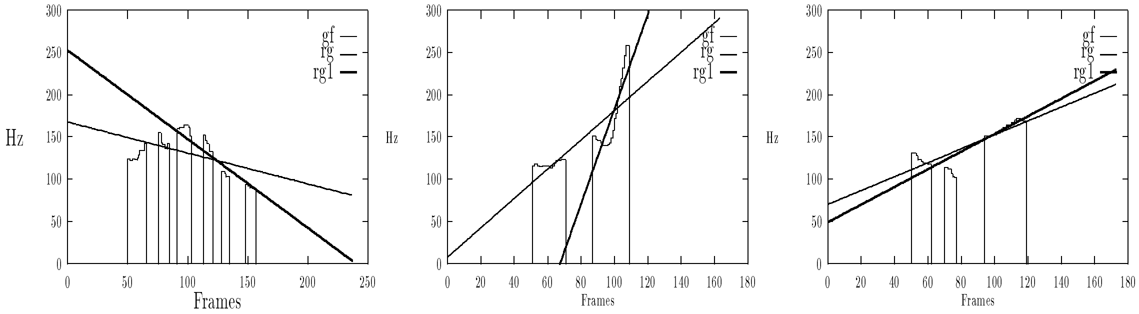

Figure 5: Fundamental frequency (gf) of, from left to right, *a declarative, an interrogative, and a continuation rise type sentence*; regression line of the whole utterance (rg); and regression line of the final voiced region (rg1)

the time may consist of the time *in isolation*, for example, '10.25' or '10.25 o'clock' (in German: "10 Uhr 25"), or it may consist of the time plus some other remarks, for example, 'yes, 10.25!' or 'was it 10.25?'. If the time is uttered in isolation, only the intonation indicates whether the user confirms (declarative sentence mood) or asks for some type of confirmation by the officer (interrogative mood). If the time is repeated together with some other remarks, those can usually be used to determine the mood of the utterance; in this case prosody could help to make the decision more reliable, but it is not the only cue.

To get a quantitative idea of the importance of prosodic discrimination, we considered 107 information dialogs about train connections recorded in three different German cities. In 92 of them the user asked for departure or arrival times. They contained 215 utterances where the users articulated 227 clock times. In 99 cases, or in 46% out of the 227 clock times, only intonation was relevant for the discrimination of the utterance mood. This means that an average dialog contains two to three clock times ($227/92 = 2.46$); on average, one clock time per dialog ($99/92 = 1.07$) or every second of the total clock times ($99/227 = 0.43$) can only be understood correctly with the use of intonation.

The evaluation of the dialogs showed that in addition to the obvious utterance moods 'declarative' and 'interrogative' there is a third one, the 'continuation rise' type indicating from the side of the user that (s)he is still listening or taking notes, but understood everything. The following Figure 5 shows the fundamental frequency of a declarative utterance, a question, and a continuation rise type. For the reasons given above the utterances are clock times in isolation. Their length is about 2 sec. In addition the regression line fitting the fundamental frequency of the whole utterance and the regression line fitting only the last voiced region are given.

From the many heuristically conceivable features we selected after some preliminary experiments the slope of the regression line (of the whole utterance), the difference between the offset and the value of the regression line at the offset, the slope of the regression line of the last voiced region, and the difference between the offset and the value of the regression of the last voiced region at the offset.

A Bayes classifier with class–conditional normal probability density functions, full covariance matrices, and equal a priori probabilities was used in the classification experiments.

2.4 Prosodic Phrase Boundaries

Another forthcoming use of prosodic information will be the classification of phrase boundaries. Based on the consistency of perception experiments with about 30 persons we distinguish four types of boundaries:

B3 type boundary: those are boundaries between main and subordinate clause, between an elliptic clause and a clause, or at a particle coordinating two clauses.

B2 type boundaries: those are boundaries between constituents, or at coordinating particles between constituents.

B1 type boundaries: those are boundaries that belong to normal constituent boundaries B2, but which are most likely not to be marked prosodically because they are either close to a B3 boundary or to the beginning/end of the utterance.

B0 type boundary: every word boundary which is not B1, B2 or B3 is of type B0.

A set of prosodic features was computed at every word boundary which was obtained from HMM word recognition. The features are the length of a pause, the normalized and unnormalized duration of the syllable and syllable nucleus prior to the boundary, the mean and the standard deviation of the duration of the phoneme class of the syllable nucleus, the energy and the position relative to the boundary of the frame having the maximum energy within the two syllables to the left and the right of the boundary, the average energy of the two syllables to the left and right of the boundary, the linear regression coefficients of the F_0 contour computed over two and four syllables to the left and to the right of the boundary, onset, minimum, maximum, and offset F_0 and their positions in time relative to the boundary, computed over two syllables to the left and right of the boundary.

3 RESULTS

Word Recognition

Tests of the above system were performed with the so called 'ERBA' sample of speech. The training sample consists of about 11 hours of speech. It was produced from 31 female and 48 male speakers who read 100 different utterances with 949 different words. Speaker–independent recognition tests were performed with a test sample of 27 minutes of speech from 1 female and 3 male speakers not contained in the training sample. The test sample contained 162 different words. Recognition tests were performed with a lexicon of 1081 words. Partly a bigram language model with perplexity $PPX = 111$ and 95 categories was used.

In Figure 6 detailed experimental results on different subword units are given. With adequate settings the word accuracy is $WA = 92.5\%$ and sentence accuracy $SA = 64\%$ (speaker–independent).

Fundamental Frequency

Fundamental frequency extraction was tested on two different German speech databases (called databases A and B, which both were recorded at the *Institut für Phonetik* at the Ludwig-Maximilian Universität, München). They contained minimal sentence pairs, that is, pairs where mood and focus of the second sentence was determined by the first (context) sentence, and mood and focus of the second sentence could only be discriminated by intonation. This design of the sentences resulted in high variations of F_0 thus making them interesting for testing the algorithm. For both databases the average difference between the minimal and the maximal fundamental frequency within an utterance was about 120 Hz. These values were computed on the automatically determined and hand-corrected F_0 contours. Database A consisted of 195 utterances from 7 speakers (4 male, 3 female). Database B consisted of 357 utterances from the speakers of database A except one male speaker. With the algorithm for voiced/unvoiced decision mentioned above in database A 333 sec of speech were classified as voiced, in database B 469 sec were classified

cutting	units	number of HMM	number of PDF	PPX162 WA	PPX162 SA	PPX1081 WA	PPX1081 SA	PPX111 WA	PPX111 SA
right	MONO	101	185	87.1	42.8	79.4	18.4	88.2	51.3
	BI	546	1243	89.6	49.0	82.1	28.0	91.9	61.0
	TRI	1257	2872	91.6	54.0	82.4	28.0	92.1	60.5
	PENTA	2087	6704	91.6	55.5	83.4	32.8	91.7	60.5
	POLY	2385	5464	91.7	56.0	83.9	34.5	91.9	59.8
	POLY+SYL	2801	6439	91.7	55.8	84.3	35.3	91.9	61.3
left	BI	560	1272	90.9	55.0	82.9	32.8	92.2	64.5
	TRI	1246	2863	90.9	54.3	84.3	34.8	92.1	62.5
	PENTA	1925	4398	91.7	56.8	85.0	36.2	92.2	63.3
	POLY	2612	7813	91.5	56.0	85.2	37.5	92.5	63.5
	POLY+SYL	2790	6412	91.8	56.0	85.4	37.3	92.5	64.8
right+words	MONO	308	2622	86.6	42.3	78.1	28.8	86.3	46.8
	BI	753	3680	91.1	54.8	85.2	38.8	91.9	62.8
	TRI	1464	5309	92.0	58.3	83.7	33.3	92.4	63.3
	PENTA	2087	6704	92.1	58.5	85.1	36.0	92.4	63.5
	POLY	2574	7683	92.1	58.5	85.1	37.6	92.5	64.3
	POLY+SYL	2991	8674	92.2	57.5	85.9	40.0	92.3	63.5
left+words	BI	767	3709	90.4	54.8	85.6	39.5	92.1	64.0
	TRI	1453	5300	91.4	57.3	85.2	38.3	92.4	64.0
	PENTA	2130	6819	91.7	56.8	85.5	40.3	92.5	63.8
	POLY	2612	7813	91.8	58.0	85.6	40.5	92.5	64.0
	POLY+SYL	2978	8628	92.0	58.3	85.9	39.3	92.5	63.8

Figure 6: Word accuracy WA and sentence accuracy SA for different inventories of subword units

as voiced. Parameters and thresholds of the algorithm have been manually adjusted using database A. Database B was only used for a final test.

Within our speech system F_0 contours will be used for determining the sentence mood and focus of utterances as well as for phrase boundary detection. For these tasks it is important to have a reliable fundamental frequency contour where the values do not have to be very accurate. Hence we consider here only the so called coarse error rate [8]. A *coarse error* occurs if the automatically determined F_0 value and the reference value differ by more than 30 Hz. The coarse error rates for frames and sentences are given in Figure 7. The error rates were determined by comparing the automatically computed F_0 contours manually with contours produced by a mechanical pitch detector. If necessary an exact reference value was determined from the signal and with perception tests. The fact that the performance of our algorithm on database B is better than on the 'training' database A is due to the greater number of laryngealizations in database A.

Sentence Mood

The experiments reported here are based on a sample of 30 declarative, 30 interrogative, and 30 continuation rise utterances spoken by four speakers giving a total of 360 utterances.

At first a perception test was performed with 2 listeners in order to determine how reliable a human listener can distinguish the three types of utterance. On average, 95.4% of the declarative, 93.8% of the interrogative, and 85.8% of the continuation rise type

	coarse error, frame				coarse error, sentence			
database	DP	DP_s	$AMDF_s$	$Seneff_s$	DP	DP_s	$AMDF_s$	$Seneff_s$
A	1.7	1.6	1.9	1.7	12.3	8.9	30.3	17.9
B	0.6	0.6	1.9	1.3	8.1	6.4	41.9	27.9

Figure 7: Percentage of frames and sentences with coarse errors (difference more than 30 Hz)

utterances were classified correctly by the two human listeners.

It became evident from the above experiments that in some cases the sentence mood was not produced correctly. In addition, the determination of the fundamental frequency F_0 was wrong in some cases. For the classification experiments with the Bayes classifier these erroneous utterances were eliminated from the sample. In total, 322 or 89.4% out of 360 sentences were used in the classification experiments.

Among others, experiments of the type 'leave–one-(speaker)–out' were performed. The average recognition rate when successively leaving out the four speakers is 86.5% if $n = 1$ and 87.5% if $n = 2$.

Prosodic Phrase Boundaries

A Gaussian classifier was trained on 6900 sentences with 74,000 word boundaries from the ERBA sample and tested on 2100 sentences with 22,000 word boundaries from the same sample. In first experiments we achieved an average recognition rate of 67%. Better results can be obtained with a polynomial classifier as reported in [4]. The automatic labeling of potential phrase boundaries is not yet integrated into our system. It should enhance linguistic analysis.

4 CONCLUSION AND OUTLOOK

The paper described a powerful environment for acoustic–phonetic word modeling and decoding of speech. A general phonetic context, determined automatically during training by the frequency of occurrence of phonetic contexts is determined. Speaker–independent recognition results were presented. Besides the acoustic–phonetic (segmental) information the speech signal contains prosodic (suprasegmental) information which can be used, for example, to determine the sentence mood and phrase boundaries. An algorithm for the estimation of fundamental frequency was described and results of its performance and its application to determination of sentence mood and phrase boundaries were presented. The integration into a speech understanding and dialog system is described, for example, in [14].

Presently, phonetic and prosodic (segmental and suprasegmental) information are evaluated and used in two seperate channels. We plan to investigate the joint exploitation in order to make full use of the information contained in the speech signal.

REFERENCES

[1] H. Altmann, A. Batliner, and W. Oppenrieder. *Zur Intonation von Modus und Fokus im Deutschen.* Max–Niemeyer–Verlag, Tübingen, 1989.

[2] L.R. Bahl, J.K. Baker, P.S. Cohen, A.G. Cole, F. Jelinek, B.L. Lewis, and R.L. Mercer. Automatic recognition of continuously spoken sentences from a finite state grammar. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 418–421, Tulsa, 1978.

[3] J.K. Baker. The DRAGON system — an overview. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 23:24–29, 1975.

[4] A. Batliner, R. Kompe, A. Kiessling, E. Nöth, H. Niemann, and U. Kilian. The prosodic marking of accents and phrase boundaries: Expectations and results. In *Proc. NATO ASI, this volume*. Springer, 1993.

[5] D. Bolinger. *Intonation and its Use: Melody in Grammar and Discourse*. Edward Arnold, London, 1989.

[6] Y. Chow, M. Dunham, O. Kimball, M. Krasner, G.F. Kubala, J. Makhoul, P. Price, S. Roucos, and R. Schwartz. BYBLOS: The BBN continuous speech recognition system. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 89–92, 1987.

[7] B. Downing. *Syntactic Sturcture and Phonological Phrasing in English*. PhD thesis, Univeristy of Texas, Austin, 1970.

[8] W. Hess. *Algorithms and Devices for Pitch Determination of Speech Signals*. Springer, Berlin, 1983.

[9] A. Kiessling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP–based determination of f0 contour from speech signals. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages II–17 – II–20, San Francisco, CA, 1992.

[10] W.A. Lea, M.F. Medress, and T.E. Skinner. A prosodically guided speech understanding strategy. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 23:30–38, 1975.

[11] K.-F. Lee and S. Mahajan. Corrective and reinforcement learning for speaker-independent continuous speech recognition. In *Proc. European Conf. on Speech Communication and Technology*, pages 490–493, Paris, 1989.

[12] B. Lowerre and D.R. Reddy. The HARPY speech understanding system. In W.A. Lea, editor, *Trends in Speech Recognition*, pages 340–360. Prentice Hall, Englewood Cliffs, NJ, 1980.

[13] H. Ney, R. Haeb-Umbach, B.-H. Tran, and M. Oerder. Improvements in beam search for 10000–word continuous speech recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages I–9 – I–12, Minneapolis, MN, 1992.

[14] M. Mast, F. Kummert, U. Ehrlich, G. Fink, T. Kuhn, H. Niemann, G. Sagerer. A Speech Understanding System With a Homogeneous Linguistic Knowledge Base. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16:179-194, 1994

[15] E. Nöth. *Prosodische Information in der Sprachverarbeitung, Berechnung und Anwendung*. PhD thesis, Technische Fakultät, Universität Erlangen-Nürnberg, Erlangen, 1990.

[16] E. Nöth, H. Niemann, and S. Schmölz. Prosodic features in German speech: Stress assignment by man and machine. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 101–106. Springer, Berlin, 1988.

[17] J.B. Pierrehumbert. *The Phonology and Phonetics of English Intonation.* PhD thesis, Massachusetts Institute of Technology, Cambridge; distributed by the Indiana University Linguistics Club, 1980.

[18] L.R. Rabiner. Mathematical foundations of hidden Markov models. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 183–205. Springer, Berlin, 1988.

[19] E.G. Schukat-Talamazzini, M. Bielecki, H. Niemann, T. Kuhn, and S. Rieck. A non–metrical space search algorithm for fast gaussian vector quantization. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages II–688 – II–691, Minneapolis, MN, 1993.

[20] E.G. Schukat-Talamazzini and H. Niemann. Das ISADORA system — ein akustisch phonetisches netzwerk zur automatischen spracherkennung. In *Proc. 13. DAGM–Symposium*, pages 251–258, Berlin, 1991. Springer.

[21] E.G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Acoustic modeling of subword units in the ISADORA speech recognizer. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 577–580, San Francisco, CA, 1992.

[22] M. Steedman. Grammar, intonation and discourse information. In G. Görz, editor, *KONVENS 92*, pages 21–28, Berlin, 1992. Springer.

[23] J. Vaissiere. The use of prosodic parameters in automatic speech recognition. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 71–100. Springer, Berlin, 1988.

[24] A. Waibel. *Prosody and speech recognition.* PhD thesis, Carnegie-Mellon Univ. Pittsburgh, USA, 1986.

[25] M.Q. Wang and J. Hirschberg. Automatic classification of intonational phrase boundaries. *Computer Speech & Language*, 6:175–196, 1992.