VOWEL-ONSET DETECTION: APPROACHES BASED ON VOWEL-STRENGTH MEASUREMENT, COCHLEAR NUCLEUS SIMULATION AND MULTI-LAYER PERCEPTRONS

Reinier W. Kortekaas (1), Georg F. Meyer (2) and Dik J. Hermes (1)

(1) Institute for Perception Research, PO Box 513, 5600-MB Eindhoven, Netherlands.
(2) Dept of Computer Science, Keele University, Keele, Staffs., ST5 5BG, England.

1. INTRODUCTION

In this paper three vowel-onset detection strategies are presented. Vowel-onsets appear to be of importance for both speech production and perception. A previously developed algorithm [11] is compared with detection schemes based on simulation of a particular response type of cochlear nucleus cells and on neural networks for pattern matching. The comparison between detection algorithms gives clues about the relative importance of physical characteristics of the speech signal for vowel-onset detection.

2. VOWEL-ONSET DETECTION

A prominent characteristic of speech signals is the presence of simultaneous frequency and amplitude modulation. Low-rate modulations can be found on a suprasegmental level in the pitch contour for FM and in the syllabic structure in the case of AM. At the syllabic level, modulations at high rates occur for instance in the fast transitions between phonemes. There is growing evidence that these fast transitions are important for phoneme recognition. Especially in the case of a plosive-vowel combination a short portion of the speech signal, typically 20-40 ms, appears to contain sufficient information for determination of the place of articulation of the consonant or the identity of the vowel [12,27]. In general, much perceptually relevant information is present in the speech portions which show substantial spectral change [26,7,20].

Hermes [11] concentrates on the concept of vowel onset defined as the moment at which a listener starts to perceive the vowel in a CV utterance. By using a gating paradigm, a trained phonetician can aurally detect vowel onsets with an accuracy better than 20 ms. These onsets are believed to coincide with both relatively rapid spectral change and an increase in the amount of 'vowelness'. Intonation research has shown that the prominence lent to a syllable is affected by the position of the pitch movement relative to the vowel onset [8,9]. Hermes [11] presents an algorithm for automatic detection of vowel onsets in natural speech.

The algorithm has been applied in a tutorial system for teaching intonation to profoundly deaf children. The performance of the algorithm was judged to be unsatisfactory in that too many vowel onsets were missed, typically in the order of 10%. Modification of the algorithm, by incorporating processing stages which were more psychophysically inspired, could not substantially improve its performance. A brief description of the Hermes [11] algorithm shall be presented in section 3.

This paper describes further investigations of automatic vowel-onset detection by comparing the Hermes [11] algorithm to two alternative detection strategies. Section 4 of this paper describes a detection scheme based on simulated cochlear nucleus responses. A multi-layer perceptron approach is presented in section 5. Comparative performance tests on two large databases containing natural speech are described in sections 6.1 and 6.2.

3. VOWEL-STRENGTH MEASUREMENT

The Hermes [11] algorithm for vowel-onset detection is based on pitch-synchronous measurement of vowel strength. This measure expresses both the degree to which a formant pattern is present in the amplitude spectrum of a pitch period of the speech signal, and its estimated pitch strength. Vowel onsets are associated with the positive maxima of the smoothed derivative of the sequence of vowel-strength measurements within a sentence. The algorithm will be refereed to as VOD in the following. For further details of the algorithm see Hermes [11].

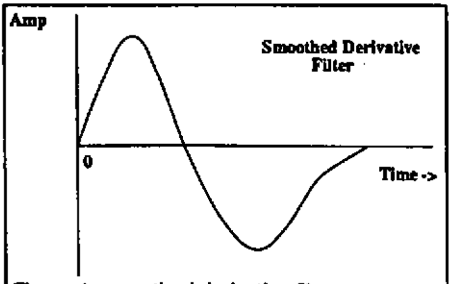

Figure 1: smoothed derivative filter

Figure 1 shows the time signal of the smoothed-derivative filter that is used for convolution with the sequence of vowel-strength measurements. The effective duration of the filter is approximately 100 ms. Both the auditory model and the MLP schemes apply a similar filter for vowel onset detection.

4. SIMULATION OF COCHLEAR NUCLEUS RESPONSES

The initial motivation for investigating simulations of peripheral auditory processes was based on overshoot phenomena observed in the cochlear nerve (e.g. [25,5]). The hypothesis was that rapid changes in the spectro-temporal domain, i.e. vowel onsets, are enhanced in firing rate profiles of cochlear nerve fibres due to short-term adaptation. The dynamic ranges of single nerve fibres however are generally too small to provide differential coding over the whole speech range. The auditory system seems to be provided with a continuum of nerve fibres with different thresholds and dynamic ranges [14,4]. Often a categorisation into a low- and a high-threshold population is made. To code spectral information in terms of discharge rate over the whole range of hearing a combination of both fibre types is necessary.

Such a combination is found in the cochlear nucleus which is the first stage in the central auditory pathway [24]. Stellate cells, which show 'transient chopper' response patterns, receive excitatory input from both high and low threshold cochlear nerve fibres [1]. The characteristic frequencies of these fibres are around 1 bark of the characteristic frequency of the stellate cell [23,24]. The cell receives inhibition from a relatively large receptive field. Spontaneous activity is often absent and dynamic ranges are small [22]. Blackburn and Sachs [1] showed that the spectrum of a synthetic vowel is preserved in the rate profiles of transient choppers over a wide dynamic range (35-75 dB SPL). At the level of the cochlear nucleus, transient chopper responses may provide the most useful information for vowel-onset detection [13].

4.1 The CNet Model

The CNET model comprises a simulation of the auditory periphery and of various response types in the VCN (see [17,18] for details). The peripheral part of the model incorporates:

- 32 channel, recursive 4th order gamma tone filterbank [2,3].
- Filter output scaling for :
    - Hearing threshold adjustment.
    - Dynamic range extension.
- Hair-cell model [16].
- Spike generation on basis of expected firing rates.

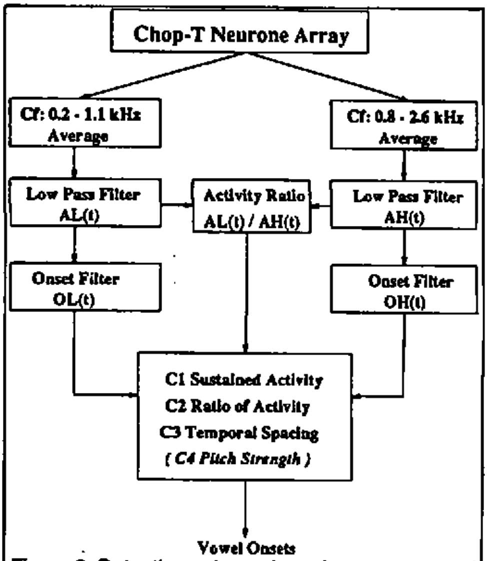

Figure 2: Detection scheme based on responses of simulated transient chopper neurones.

The dynamic range extension is introduced for the simulation two populations of cochlear nerve fibres: one with high spontaneous rates (50sp/sec) and human audiogram thresholds [6] and another with low spontaneous rates (15sp/sec) and thresholds 15 dB above the normal human thresholds.

Simulation of cochlear nucleus responses is based on a point-neurone model, where the membrane potential is controlled by the Goldman-Hodgkin-Katz equation. Instead of generating action potentials, we concentrate on the extracellular potential relative to the firing threshold. The output of the transient chopper simulation is the extracellular potential (re threshold) of an array of 22 neurones. Centre frequencies of those neurones range from 0.2 to 2.6 kHz with 0.5 bark spacing.

4.2 Vowel-Onset Detection Scheme

The vowel-onset detection scheme integrates the activity of the neurone array into two bands, roughly corresponding to the first and second formant region. Band 1 spans the centre-frequency range from 0.2 up to 1.1 kHz, band 2 receives input from 0.8 up to 2.6 kHz. Activity in both bands is averaged over centre frequency and low-pass filtered in time by leaky integration (-3 dB points of the LP filter at ~25 Hz). The signals AL(t) and AH(t), as shown in figure 2, contain the LP filtered activity. Onsets are found by taking the smoothed derivative as described in section 3, but with effective duration ~40 ms, resulting in the signals OL(t) and OH(t). Vowel-onset candidates are found at the positive maxima of OL(t) with synchronous increase of activity in the second band, i.e. OH(t) > 0.

In order to exclude false alarms, vowel-onset candidates should meet the following criteria:

C1 Sustained activity. The activity averaged over 25 ms following the vowel-onset candidate should be at least half the activity at the vowel-onset candidate. In this way, short bursts of activity, which are often found in plosive contexts, are discarded.

C2 Ratio of activity. The ratio of the activity in both bands should be within a lower and an upper bound. This criterion ideally excludes nasals because most of their spectral energy is in the lower band and, mutatis mutandis, ideally discards fricatives.

C3 Temporal spacing.

A If two consecutive onsets are found while the activity in band 1 is continuously increasing, then the first vowel-onset candidate is discarded. In this way the vowel of a CV combination is detected if the consonant could not be excluded by criterion C2.

B. If the activity in band 1 does decrease between the consecutive onsets, the temporal spacing between the onsets should be at least 60 ms.

Figure 3a-e displays the activity and onsets signals for a Dutch sentence from the PM database of section 6.2 (sentence 13: "Eindelijk kwam de trein op gang").

This scheme does not, unlike VOD, include information about the periodicity of the signal. Transient-chopper responses show phase locking up to approximately 400 Hz for pure tone stimulation. Information about voicing of the speech signal could thus in general be derived. A rather ad-hoc solution is to calculate the short-term autocorrelation of the output of each transient-chopper neurone. These autocorrelation functions are summed over all neurones to obtain the summary autocorrelogram (cf. [19]). The amplitude of the peak of the summary autocorrelation is taken to represent the pitch strength (see figure 3e). In this extended scheme, referred to as CN-ACF, a criterion is introduced that takes voicing into account:

C4 For each vowel-onset candidate, the corresponding pitch strength should be above 10 % of the maximum pitch strength observed in the utterance.

All parameter settings given above were found by a trial and error method.

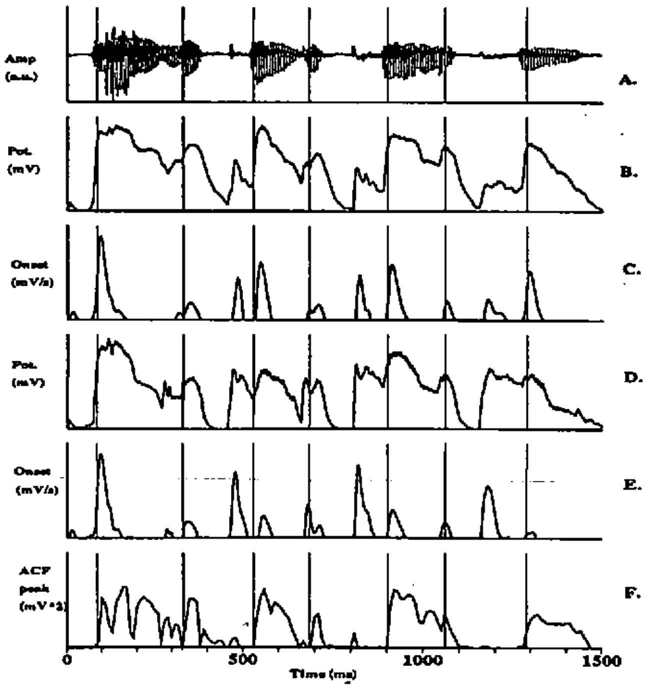

Figure 3: (a) Waveform of PM sentence 13, (b) Activity in band 1 at 55 dB SPL, (c) Smoothed (positive) derivative of band 1, (d)-(e) Activity and smoothed derivative of band 2, and (f) Amplitude of the peak of the summary autocorrelogram. Actual vowel onsets are marked by vertical lines

5. MULTI-LAYER PERCEPTRONS

A possible weak point of the manually adjusted detection scheme is that the decision boundary for the vowel/non-vowel categories may not be optimal. MLPs have been shown to be excellent tools for pattern classification so that they were used as a 'vowel identification stage'. Three sets of experiments were performed:

1. Classification performance was compared using the auditory model and mel scale spectra as input to the neural network to ensure that the auditory model signal presentation is at least comparable to conventional pre-processing techniques.
2. The neural network was trained with mel-scale spectra with and without amplitude information (spectra normalised frame by frame) to evaluate the importance of amplitude cues for the task.
3. The network performance for input with the reduced spectral representation as used in the CN paradigm was compared with performance for input at the full auditory model resolution.

5.1 Network Architecture

The output representation of the network was pre-determined by the chosen postprocessing: one unit, representing the presence/absence of a vowel in its activation. Full and reduced spectral representations were experimented with (26/25 and 2 input units). Performance for a range of 0-10 hidden units was evaluated. The best results were obtained for 2-5 hidden units. If no hidden units are used the network will not learn, while 10 hidden units cause 'over-training': the network performs very well on the training task, but not the test data.

5.2 Network Training

Neural networks are difficult to train with time varying signals because the standard network architectures do not allow for the concept of time. The training paradigm employed was to train the network purely as a pattern matching stage. The neural network was presented with single spectra calculated over 25.6ms long time slices, either from the mel-scale spectra or by the auditory model. Spectral resolution was 26 or 2 bins spanning 0.2-3.3kHz for the auditory model and 25 frequency bins (0-5 kHz) for the mel scale spectra. Training samples were taken starting at the aurally detected vowel onsets and 25.6ms after the reference data. The two spectra were checked visually to exclude erroneous training data caused by early detections or very short vowels. Training data for the 'non-vowel' category was chosen in two passes. Initially a small set of examples of non-vowels and silences chosen. In a second pass examples were added at positions where the network wrongly detected vowel onsets.

The MLP was trained using standard back-propagation with a learning rate of 0.0005 and a momentum term of 0.1, to prevent over-training an error threshold of 0.05 was set. Larger learning rates or faster training algorithms, such as resilient propagation and quickprop, were experimented with but proved unsatisfactory [28].

The networks were trained in steps of 100 training cycles until the summed error in the vowel onset detection performance no longer declined, usually for 400-600 cycles. Performance on the test data, full T sentences, usually deteriorated as training progressed past the optimal point.

5.3 Vowel-Onset Detection

The sentences, processed in 1ms steps, were presented to the network and the output unit activation was recorded. The trace, representing vowel presence was then processed using the onset filter used by both Hermes and in the auditory model. The only difference was that a threshold of 0.6 was used before vowel onsets were detected. An example neural network output for PM sentence 13 is given in figure 4.

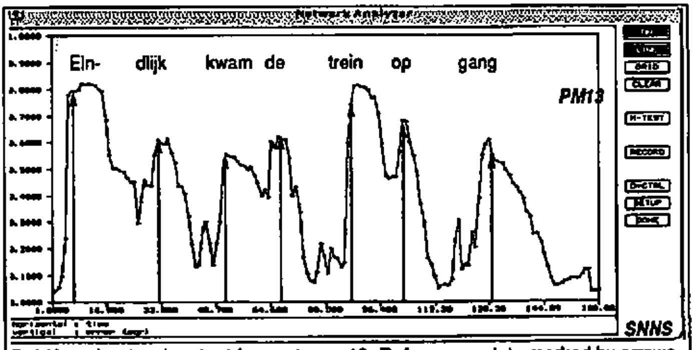

fig4:Neural network output for sentence 13. Reference points marked by arrows

6. COMPARING DETECTION PERFORMANCES

In Hermes [11] a 28 sentence database, referred to as 'T-sentences', was used for the performance test. Both the CN (CN-ACF) and the MLP approach were optimised for this database. In 6.1 we will present the comparative tests for this database and in 6.2 the performances for a new database will be described. In all tests, the automatically detected onsets will be compared to aural detections, also called actual onsets. Detected onsets should be within +/- 50 ms to the actual onsets. Performance will be marked by the missed onsets and false alarms, as percentages of the actual number of onsets. Moreover, accuracy figures indicate the proportion of automatically detected onsets that are within +/-20 ms of the actual onsets.

6.1 The T-Sentence Database

The database contained Dutch read speech from non-professional speakers (14 male / 14 female). The total number of actual onsets was 377. In table 1 the performance of the different approaches is given.

TABLE 1: **T database results**	missed onsets (%)	false alarms (%)	accuracy (%)	Comments
VOD	8	3	91	
CN 35dB SPL	20	5	85	
CN 55dB SPL	10	9	86	
CN 75 dB SPL	10	15	84	
CN ACF 35dB SPL	27	4	79	
CN ACF 55dB SPL	11	6	87	
CN ACF 75dB SPL	10	6	83	
MLP Mel spectra	9 / 14	9 / 18	85 / 81	sent / spec normalised
MLP 35dB train 35	13 / 11	5 / 8	90 / 86	2 / 26 input units
MLP 55dB train 55	10 / 9	9 / 6	88 / 88	trained only on
MLP 75dB train 75	8 / 9	13 / 8	81 / 88	test amplitude
MLP 35dB train all	28 / 17	12 / 10	86 / 90	2 / 26 input units
MLP 55dB train all	19 / 11	21 / 21	76 / 82	trained on all amps
MLP 75dB train all	13 / 9	22 / 18	74 / 81	simultaneously

In the CN (-ACF) approach, the sound pressure level of 35 dB SPL was seen to be too low. Introduction of voicing information in the CN scheme did reduce the number of false alarms substantially. Most of all this was achieved by rejecting vowel-onset candidates for unvoiced plosives and fricatives. In VOD, the main category of missed onsets were schwas, in unaccented syllables. This also holds for the CN (-ACF) results at 55 and 75 dB SPL.

Neural networks are good pattern classification tools, as the training data, particularly for the mel-scale spectra shows. If the signal is normalised, so that each spectrum contains amplitude information, the pure pattern matching strategy performs adequately well, vowel onset detection performance is good (9% missed) but the network introduces too many false alarms (9%). Missed onsets occurred mostly for schwas and high vowels like /i/ and /I/ whereas false alarms were found mostly in /l/, /r/ nasal and long vowel contexts. The performance deteriorates significantly when the amplitude cues are removed from the mel spectra.

Training the MLPs with data from the auditory model proved less successful than training with mel-spectra largely because the auditory model representation changes with amplitude. When the network was trained and tested on one amplitude level only, the performance is comparable to the other approaches, but training on all levels simultaneously does not give satisfactory results. This limitation can easily be avoided by scaling the input to a fixed level or by using one of a number of amplitude specific networks matched to the signal level.

6.2 The PM-Sentence Database

A database was composed by randomly selecting 28 sentences from the Plomp & Mimpen [20] set. This database consisted of 56 Dutch sentences, read by 14 male and 14 female non-professional speakers. The number of actual vowel onsets was 466. None of the detection schemes were optimised for this data.

TABLE 2: PM database results	missed onsets (%)	false alarms (%)	accuracy (%)	Comments
VOD	7	9	90	
CN 35dB SPL	21	10	90	
CN 55dB SPL	9	15	90	
CN 75 dB SPL	11	28	89	
CN ACF 35dB SPL	29	8	92	
CN ACF 55dB SPL	11	10	91	
CN ACF 75dB SPL	11	15	89	
MLP MEL Spectra	7 / 15	10 / 25	94 / 83	signal/spec normalised
MLP 35 train 35	15 / 16	7 / 8	88 / 94	2 / 26 input units
MLP 55 train 55	9 / 8	7 / 14	93 / 90	training on
MLP 75 train 75	7 / 8	11 / 21	91 / 87	test amplitudes only
MLP 35 train all	22 / 24	12 / 12	87 / 94	2 / 26 input units
MLP 55 train all	17 / 11	27 / 29	82 / 84	training on all
MLP 75 train all	11 / 10	47 / 50	74 / 80	amps simultaneously

The number of false alarms is substantially higher for all schemes than the respective number for the T-sentence database. Missed onsets mainly occurred for schwa, /i/ and /ε/ vowels respectively in the VOD and CN schemes. False alarms were mainly found in schwa-like and /r/ contexts for VOD and in schwa-like, /r/ and unvoiced plosive contexts for the CN schemes.

The MLPs performed very well on the previously unseen PM sentences indicating that the networks are extracting useful features from the signal rather than performing a simple pattern matching task. The performance for the mel-spectra and the amplitude dependent auditory model data are very competitive. As expected from the T sentence experiment the performance for the network trained on all signals levels is disappointing.

7. DISCUSSION

In this paper we have assumed that human vowel-onset detection is based on a process of categorical perception of vowels versus non-vowels; more specifically, vowel onsets are contrasted with other onsets. Another hypothesis may state that vowel-onset detection is derived indirectly from a phoneme recognition process. In that case, vowel-onset detection will be governed to a great extent by higher-order processes in speech perception which will be difficult to model.

One may interpret the present comparative tests in terms of the relative importance of signal characteristics like amplitude, periodicity and spectral content. Normalising the input spectra in the MLP scheme resulted in moderate detection performance, indicating that amplitude information plays a role. Both VOD and the CN schemes take amplitude into account. Introduction of information about the periodicity of the speech signal in CN-ACF did reduce the number of false alarms while leaving the missed-onset rate almost unaffected. On the other hand, voicing information is not present in the mel-scaled input spectra in the MLP approach. Nevertheless, the missed-onsets figures obtained with this approach were satisfactory. This may indicate that pitch strength is taken as an secondary source of information only. The rather crude spectral weighing in the CN (-ACF) and MLP schemes gives support to the hypothesis that vowel-onset detection does not rely on detailed spectral analysis.

Vowel-onset detection can also be conceived as an approach for automatic syllabification. In this respect, the question rises what the relationship is between the vowel onset within a syllable and its Perceptual Centre (P-centres; [15]). This topic is addressed by current research.

8. REFERENCES

[1] **CC Blackburn and MB Sachs** (1990) "The representation of the steady-state vowel sound /e/ In the discharge pattems of cat anteroventral cochlear nucleus neurons." J. Neurophysiology 63, 1191-1212.

[2] **E de Boer** (1969) "Reverse correlation II: Initiation of nerve impulses in the inner ear", Proc Kon Ned Acad Wet, 72, 129-151.

[3] **AM Darling** (1991) "Properties and implementation of the gammatone filter: a tutorial" In: Speech, Hearing and Language. Work in progress, 1991, Vol 5. Dept. of Phonetics and Linguistics, University of London.

[4] **B Delgutte** (1987) "Peripheral auditory processing of speech information: implications from a physiological study of intensity discrimination" In: The psychophysics of Speech Production, ed. M.Schouten, Nijhof Dordrecht, The Netherlands, pp333-353.

[5] **JJ Eggermont** (1985)"Peripheral auditory adaptation and fatigue: a model oriented review." Hear Res 18, 57-71.

[6] **RR Fay** (1988) "Hearing in vertebrates: a psychophysics data book" Hill-Fay Associates. Winnetka

[7] **S Furui** (1986) "On the role of spectral transition for speech perception" J. Acoust. Soc. Am. 80, 1016-1025.

[8] **H 't Hart and S Cohen** (1973) "Intonation by rule: a perceptual quest" J. Phon. 1 309-327

[9] **H 't Hart and R Collier** (1975) "Integrating different levels of intonation analysis." J. Phon. 3, 235-255.

[10] **DJ Hermes** (1988) "Measurement of pitch by subharmonic summation" J. Acoust. Soc. Am. 83, 257-264.

[11] **DJ Hermes** (1990) "Vowel-onset detection." J. Acoust.Soc.Am. 87(2), 866-873.

[12] **DJ Kewley-Port** (1983) "Time-varying features as correlates of place of articulation in stop consonants." J Acoust Soc Am 73, 322-335.

[13] **RWL Kortekaas and GF Meyer** (1994) "Vowel-onset detection using models of the auditory periphery and the nucleus cochlearis: physiological background.", IPO report 963.

[14] **MC Liberman** (1978) "Auditory nerve response from cats raised in a low noise chamber." J Acoust Soc Am 63, 442-455.

[15] **S Marcus** (1981) "Acoustic determinants of perceptual center (P-center) location", Perception and Psychophysics, Vol. 30, 247-256.

[16] **R Meddis** (1988) "Simulation of auditory-neural transduction: Further studies" J. Acoust. Soc. Am. 83(3) 1056-1063.

[17] **G Meyer** (1993a) "Models of the neurones in the ventral cochlear nucleus: signal processing and speech recognition." Unpublished PhD thesis, Dept. of Communication and Neuroscience, University of Keele.

[18] **G Meyer** (1993b) "CNet - point neurone simulator" Tech Report TR93-01 Dept. of Computer Science, University of Keele.

[19] **GF Meyer and ID Dewar** (1994) "Pitch extraction in the auditory nerve and cochlear nucleus", *this volume*

[20] **ZB Nossair and SA Zahorian** (1991) "Dynamical spectral features as acoustic correlates for the initial stop consonant" J. Acoust. Soc. Am. 89(6), 2978-2991.

[21] **R Plomp and AM Mimpen** (1979) "Improving the reliability of testing the speech reception threshold for sentences" Audiology 18, 43-52.

[22] **Rhode W.S. and R.E. Kettner** (1987) "Physiological study of neurons in the DCN and PVCN of the unaesthetized CAT." J.Neurophysiology 57, 414-442.

[23] **WS Rhode and DH Smith** (1986) "Encoding timing and intensity in the VCN of cat." J.Neurophysiology 56(2), 287-307.

[24] **WS Rhode and S Greenberg** (1992) "Physiology of the cochlear nuclei" In: The Mammalian Auditory Pathway: Neurophysiology, ed. A.N. Popper and R.R. Fay, Springer Handbook of Auditory Research.

[25] **RL Smith and JJ Zwislocki** (1975) "Short-term adaptation and incremental responses of single auditory-Nerve Fibers." Biological Cybernetics 17, 169-182.

[26] **W Strange JJ Jenkins and TL Johnson** (1983) "Dynamic specification of coarticulated vowels" J. Acoust. Soc. Am. 74, 695-705.

[27] **ML Tekieli and WL Cullinan** (1979) "The perception of temporally segmented vowels and consonant-vowels in syllables." J. of Speech and Hearing Disorders 22, 103-121.

[28] **A Zell et al.** (1993) "Stuttgart neural network simulator V3.2" Institute for Parrallel and Distributed High Performance Systems, Universität Stuttgart, SNNS manual.