Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Acoustic-phonetic recognition of continuous speech by artificia neural ne

Elenius, K. O. E. and Takacs, G.

journal: STL-QPSR
volume: 31
number: 2-3
year: 1990
pages: 001-044

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

ACOUSTIC-PHONETIC RECOGNITION OF CONTINUOUS SPEECH BY ARTIFICIAL NEURAL NETWORKS

*Kjell Elenius and György Takács**

Abstract

This paper describes an artificial neural network that recognizes phonemes in continuous speech, based on error back-propagation training. The recognition is performed by two connected nets. First, a coarse feature network is trained to recognize seven quasi-phonetic features from 10 ms spectral frames of a Bark-scaled filter bank having 16 filters in the range 200 to 5000 Hz. The features need not be binary but may take any values between 0 and 1. The feature net is shown to be relatively insensitive to changing the speaker or even the language. The outputs of the feature net as well as the spectral outputs of the filter bank are used as input to the second net, the phone net, which recognizes phonemes. A seven frames wide symmetric window of the feature net output is used to include the context of the frame being classified. Some provision is taken to make this information resistant to changes in speed of articulation by adding hidden nodes that have some extent of overlap in their input from the feature net. The outputs of the phone are also used as inputs to a segmentation network. Fifty sentences of one speaker were used for training the different nets, and ten more were used for testing. The features of the coarse net were recognized with 80% to 95% accuracy. Correct phones were recognized with 64% accuracy and in 82% of the cases, the correct phone was among the best three candidates. The segmentation net was compared to a human segmentation, and 82% of the segments were placed at the correct segment or at a displacement of +/- 10 ms; 18% of the segments were deleted and 54% were inserted.

INTRODUCTION

This paper reports on experiments concerning phoneme recognition in continuous speech. A recognition system has been developed and evaluated while one of the authors was on a seven month visit as a guest researcher at the Department of Speech Communication and Music Acoustics.

Speech recognition based on phoneme like units is a long-term research objective. It is attractive since it is inherently free from vocabulary limitations. This is of special importance in highly inflected languages, e.g., Hungarian. Dozens of different forms of the same root word may occur making speech recognition based on word size units encounter severe practical limitations.

The system reported in this paper mixes some well established techniques with some new, recently published elements from the area of neural networks to a novel combination in order to optimize a speaker dependent phoneme recognition system.

It is generally accepted that speech recognition cannot be solved purely on the acoustic-phonetic level. There is, however, no reason to transfer acoustically-phonetically solvable tasks to higher processing levels. A main objective of this project is to recognize phoneme

*Guest researcher from the Hungarian Post Office, Research Institution, Budapest, from October 1989 to April 1990. Names in alphabetic order.

like elements on this bottom level as well as is possible. The recognition system includes both automatic segmentation and automatic classification of speech.

The output of the acoustic-phonetic level of the system is a string of phoneme candidates. Each phonetic event is related to one or several candidates in parallel. The data format may be extended by including information on durations and probabilities of the phonetic candidates if necessary. The output has no rigid sequential structure why simultaneous and overlapping events in the speech production process easily can be described.

Usually the first step in the acoustic processing of speech is a spectral analysis using 10 - 20 ms time frames. A considerable part of the phonemes have an inherently transient character making their recognition on the basis of single frames very uncertain. In spite of this a very innovative neural network based strategy has offered good recognition results based on this description (Kohonen, 1988). However, most systems use longer windows for phoneme recognition. A window having a duration of several speech frames slides across the preprocessed speech material in single frame steps. The minimal number of frames in the window is determined by the slowest phoneme transition. Transitions to phoneme targets are known to carry essential information as well as transitions from the targets. Thus the neighbourhood has two sides: the preceding and the succeeding. This supports the use of a symmetrical window to handle coarticulation effects of neighbouring phonemes.

The most traditional solution to windowing is using a section of a spectrogram-like representation (Waibel, Hanazawa, Hinton, Shikano & Lang, 1989). This representation has a high redundancy. Since automatic speech recognition sooner or later seems to be limited by the amount of data or the speed of computation, data compression has a value in its own. A widely used compression method is the description of speech frames by cepstral coefficient vectors instead of spectral vectors (Krause & Hackbarth, 1989; Elenius & Blomberg, 1982). This method yields a compression rate of approximately 2. In our system a completely different method has been introduced, partly for the purpose of data compression. It represents each 10 ms frame by some basic articulation related features that are calculated from the speech spectrum. These "coarse phonetic features" describe the manner and the place of articulation. Another purpose of this feature representation is to divide the phoneme classification task into two subtasks, which are managed separately. In the final phoneme classification of a single speech frame the values of these features within a 15 frame symmetrical window are used together with the spectral vector of the current, central, frame. Using this representation the compression rate is approximately 5 when compared to the spectrographic representation above.

The recognition of phoneme-like units is traditionally done in two sequential steps: segmentation and classification. However, an exact segmentation frequently needs information available only after the classification. Many of the papers reporting excellent recognition performance only deal with manually segmented speech samples. Thus a very problematic part of the processing is not included. Komori et al. have made experiments with automatic segmentation and they report 9.2 % errors on segmentation (Komori, Hatazaki; Tanaka, Kawabata & Shikano, 1989). We have tried to build a system that does not depend on manual segmentation during the recognition phase.

The system has a hierarchical structure and the underlying idea is to separate the system into a language independent and a language dependent part. This will facilitate the adaptation to a new language. The phoneme recognition system is based on two hierarchically connected but separately trained neural networks. It builds on well functioning elements of the most frequently used neural net structures. We may to some extent leave it to the network to learn details that we have problems in modelling explicitly. Still, we need a vast amount of general ideas, theoretical knowledge and experimental results to be able to construct good network

structures and training procedures. The structure of a neural network can be used as a carrier of our background knowledge. Some important aspects are:

- what is the task of the neural network and how is it evaluated?
- how well is the speech represented by the input data?
- how does the network structure match the task?
- which kind of training material is used and how is it utilized to train the network?
- how can a complete task be divided into subtasks for the independent training of subnets?

The principles and basics of neural networks are not discussed in this paper. This information can be found in, e.g., Rumelhart & McClelland (1986), McClelland & Rumelhart (1988), Kohonen (1984), Kohonen (1988), Lippmann (1987), Lippmann (1988), Mariani (1989), Niles (1989), Treleaven (1989).

1. FEATURES OF THE PROCESSED SPEECH MATERIALS

The first experiments were conducted using a Swedish speech database. Later, a Hungarian database was created and processed. Most tests were made by the recognition system adapted to a single speaker but the performance of the system to other speakers than those used for training were tested as well. Three different speech materials were used in the experiments.

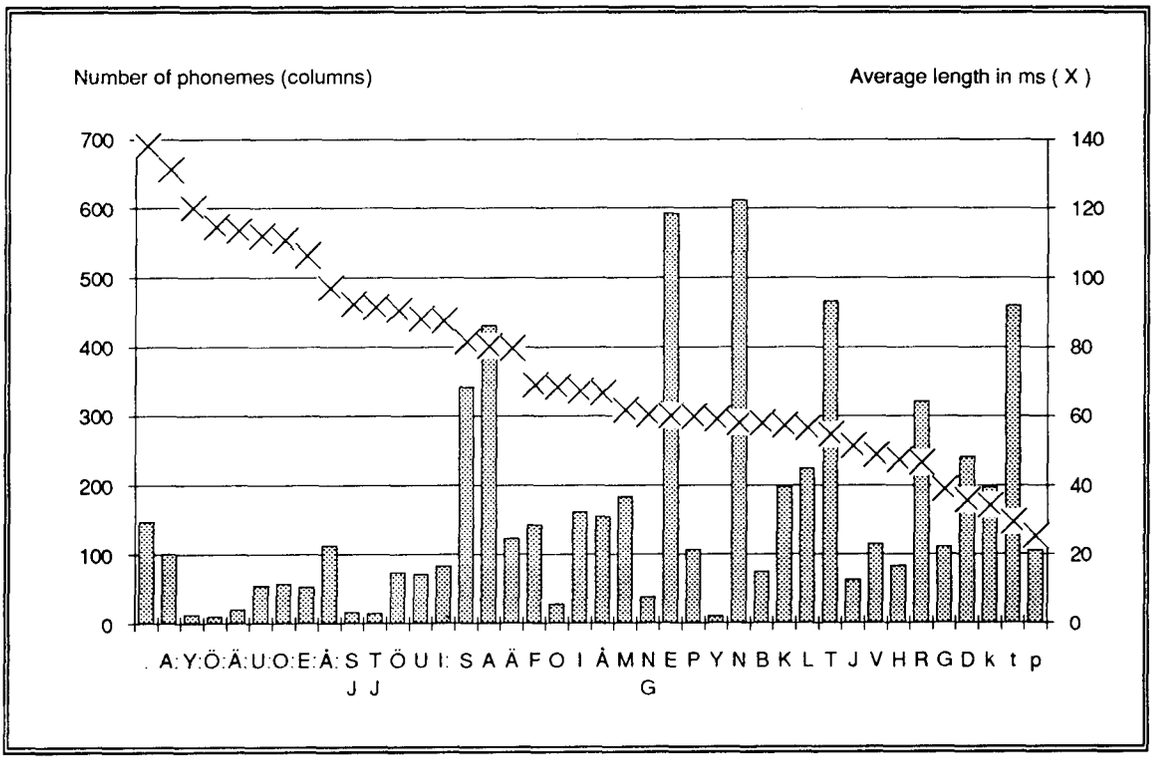

Fig. 1. Phoneme distribution in the Swedish INTRED-material (columns). The average length of phonemes is denoted by X-s. Phonemes are sorted by this length. Technical phonetic alphabet defined in Table IV. The dot-symbol (".") to the far left indicates pauses.

Sixty Swedish sentences constitute the first speech material. They are part of a speech material that has been used in several other studies and will be referred to as the INTRED-material (Hunnicutt, 1987; Nord, 1988). The material consists of 150 sentences read in a natural way by a trained male speaker having a central Swedish dialect. The reading speed was rather fast, 13.1 phonemes/s. The speech signal is sampled at 16 kHz using a 6.3 kHz low-pass filter. The sentences are labelled by a human phonetic expert using both visual and audi-

tory information (Nord, 1988). Fig. 1 shows the natural and uneven distribution of phonemes in this material. Multiplying a phoneme's frequency by its average length gives a measure of the total time conveyed by the phoneme, which is related to the amount of training samples available for the phoneme, when training an artificial neural network by (a subset of) this material.

Table I shows some sentences of the INTRED-material in ordinary Swedish orthography together with a phonetic transcription using a technical phonetic alphabet defined in Table IV. The labelling is basically phonemic and does not show allophonic variations, unless they obviously are the result of some higher level rules, e.g., retroflex variants of the consonants *n*, *d* and *l* and the Swedish vowels *ä* and *ö* have special labels. These variants have not been regarded as separate units in our experiments, i.e., we only used one *d*-phoneme.

The second speech material consists of 10 similar Swedish sentences pronounced by another male speaker. This material will be referred to as the JONSSON-material.

The third speech material was created especially for these recognition experiments. A native Hungarian male speaker read the text in a natural way. The average speed was 12.6 phonemes/s. This material was recorded and analyzed using the same technique as described for the INTRED-material, and will be referred to as the MAMO-material. In these experiments, the Hungarian phonemes were represented by a set of 49 elements. The details can be found in Section 3. The labelling was done on phonetic criteria rather than phonemic, since the analysis of recognition errors in the Swedish material indicated that the phonemic labelling was the origin of some errors. Table II contains some sentences of the MAMO-text in Hungarian orthography and also in a Hungarian technical phonetic alphabet.

Table I. The first 8 sentences of the INTRED-material using two different notations.

Text using Swedish orthography	*Text using a technical phonetic alphabet*
På utflykten grillade barnen glatt korven de fått med hemifrån.	Pp'Å:+"U:Tt#FL'YKkTtEN GR"IL'ADE B'A:2NEN GL'ATt KkORVEN D'E+ F'ÅTt+ M'EH"EMI#FR'Å:N.
Slutligen lagade mannen omsorgsfullt bilen som hade gått sönder.	SL"U:TtL'IEN L"A:G'ADE M'ANNEN "ÅMSÅRS#F'ULTt B'I:LEN S'ÅM+ "AD'E+G'ÅTtS'ÖNDÄ4R.
Det är som att sätta vagnen framför hästen.	D'E+ E+ S'ÅM+ 'Å+ S"ÄTt'A V'AGNEN FR'AM#FÖR+ H'ÄSTtEN.
Vid kontrollen stängde polisen snabbt av gatan med hjälp av stängsel.	V'I:+ KkÅNTtR'ÅLEN STt"ÄNGD'E PpÅL'I:SEN SN'APpTt 'A:V G"A:Tt'AN MÉ+ J'ÄLPp 'A:V+ STt'ÄNGSEL.
En svala gör ingen sommar.	'EN SV"A:L'A J'Ö3R+ "ING'EN+ S"ÅM'AR.
Det är svårt att hitta vagnen där han sitter.	D'E+ 'E+ SF'Å:2Tt 'Å+ H"ITt'A V'ANGEN D'Ä3R+ 'AN+ S'ITtÄ4R.
Efter lunchen lagade snickarna omsorgsfullt golvet i det gamla huset.	'EFTtE+ L'UNSJEN L"A:G'ADE SN"IKk'A2NA "ÅMSÅRS#F'ULTt G'ÅLVETt 'I:+D'E+ G"AML'A+ H'U:SETt.
Var sak har två sidor.	V'A: 2S'A:Kk H'A:+ 2TtF'Å: S"I:D'ER.

Table II. The first 8 sentences of the MAMO-material using two different notations.

Text using HUNGARIAN orthography	*Text using a technical phonetic alphabet*
Az az ember, aki könyvét összefirkálja az egyéb aljasságra is képes.	AZAZEMBbER.AKkIKkWNYVE:TWSZEFIR KkA:LJA.AZ.EGYgyE:BbALJASA:GRAISKk E:PpES.
Az átviteltechnikai osztályon két gépírónö fölmondott.	AZA:TtVITtELTtEHNIKkAIOSZTtA:JON.Kk E:TtGgE:PpI:RO:NW:FWLMONDdOTt.
Hiába vársz levelet töle, már nem szeret igazán.	HIA:BbAVA:RSZLEVELETtW:LE.MA:RNE MSZERETtIGgAZA:N.
A Dombóvár és Kaposvár között épülö autópályán kiskatonák nem dolgoztak.	ADdOMBbOVA:RE:SKkAPpOSVA:RKkWZ WTtE:PpYLW:.AUTtO:PpA:JA:NKkISKkATt ONA:KkNEMDdOLGgOZTtAKk.
Misi és Márti sok szeretettel üdvözöl mindenkit.	MISI.E:SMA:RTtI.SOKkSZERETtETtELYDd VWZWLMINDdENKkITt.
Drága nagymamád kedvében kell járnod minden nap.	DdRA:GgANAGYgyMAMA:TtKkEDdVE:BbE NKkELJA:RNODdvMINDdENAPp.
Merre szorít a cipö? Talán itt erre?	MERESZORI:TtACcIPpW:TtALA:NITt.ERE.
Sok volt nekem ez az ebéd, így a fele otmaradt.	SOKkVOLTtNEKkEMEZAZEBbEDdv.IGYgy. AFELEOTt.MARATt.

2. THE BASIC STRUCTURE OF THE COMPLETE SYSTEM

In this section, we will give an overview of the proposed system. Details regarding functions and designs will be discussed in Section 3. Continuously spoken sentences constitute the input to the system. The output of the acoustic-phonetic processing is a string of phoneme candidates and may be used as input to a language processing stage. The output of a complete system would be ordinary, written text. The connections between the basic units and the different representations used are shown in Fig. 2.

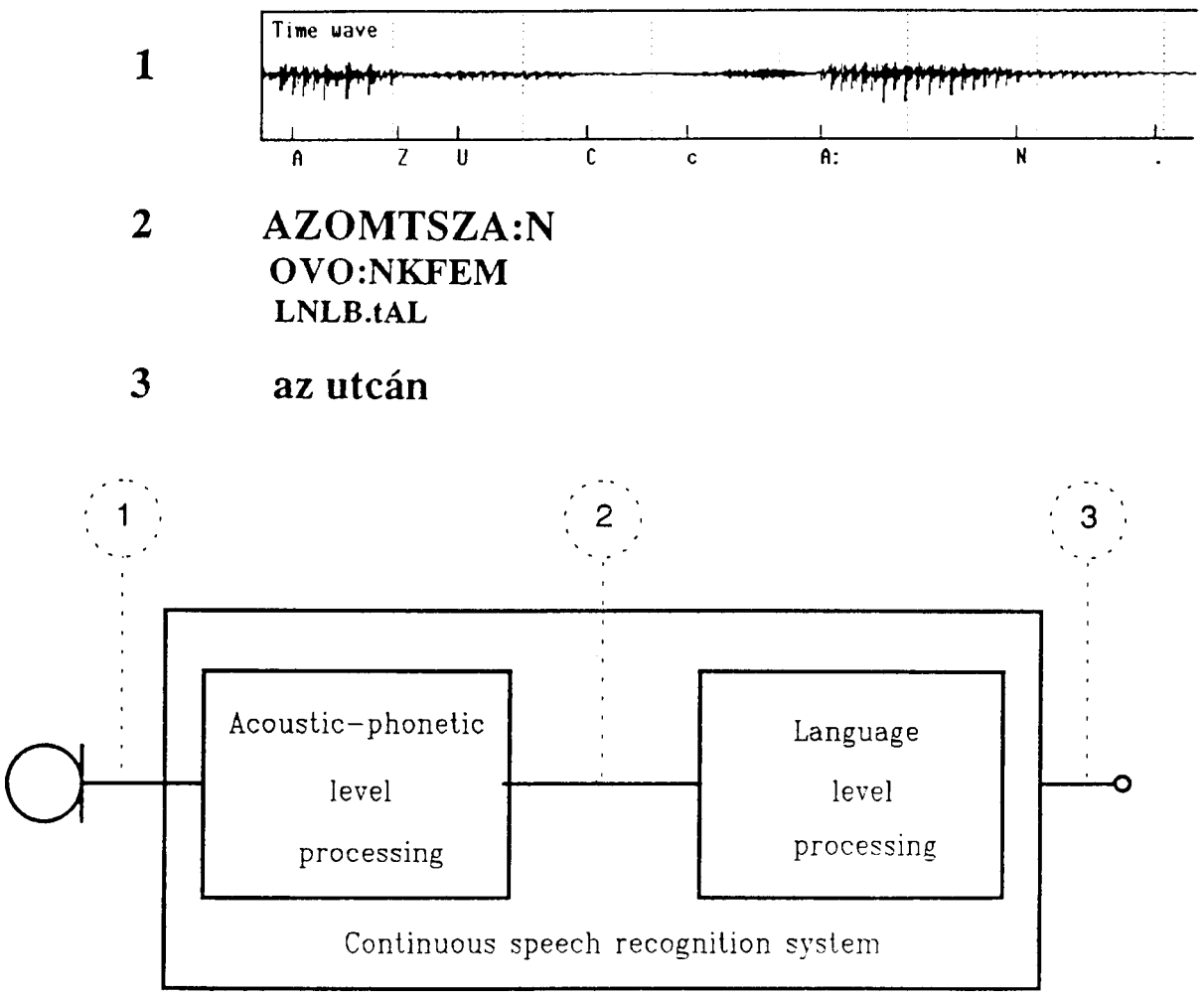

Fig. 2. *The basic structure of the recognition system and the different representations used. The input is the speech wave form (representation 1). The output of the acoustic-phonetic unit is a string of phoneme candidates (representation 2). For each segment, the probability of the phoneme candidates is indicated by the character size, large for the first candidates and smaller for the second and third candidates. This level does not indicate any word boundaries. The output of the language unit is ordinary, written text. The language processing unit is not included in this study.*

The acoustic processing does not make use of any higher level information. All the processing related to syntax and semantics (e.g., lexicon, grammar) is included in the second box in Fig. 2. This paper only deals with the acoustic-phonetic processing in the first box. It consists of 8 basic units as depicted in Fig 3.

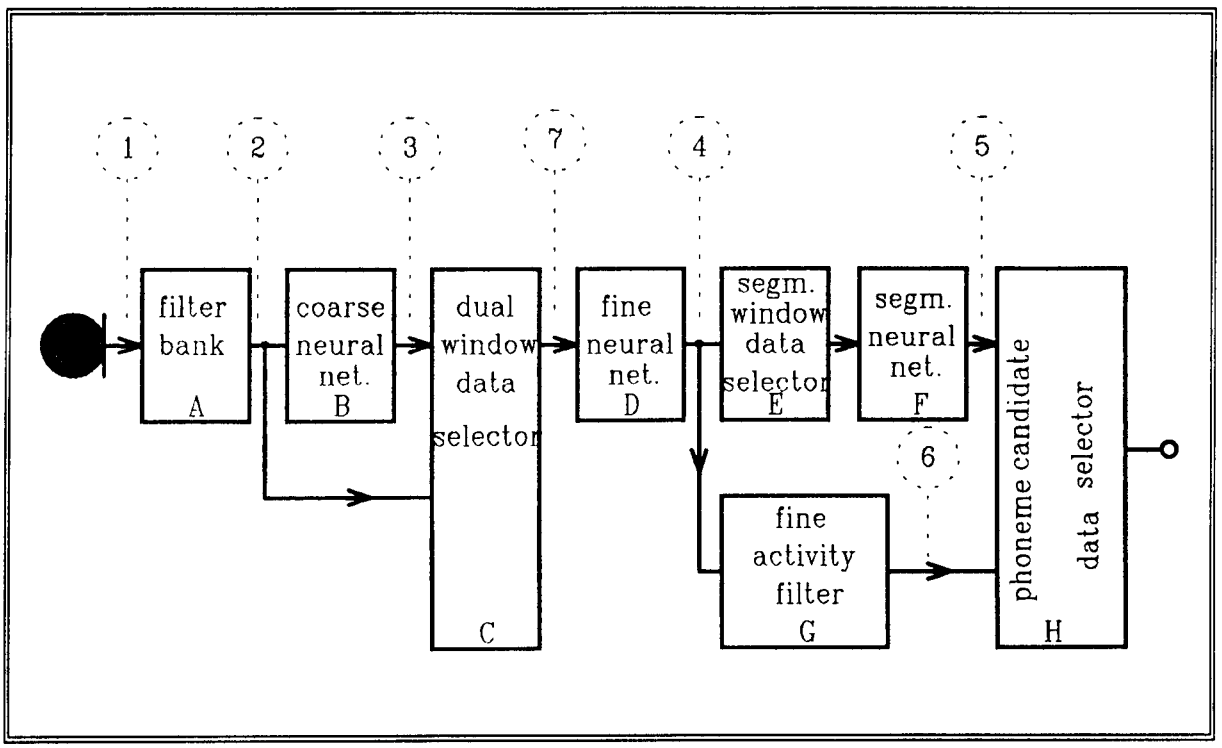

Fig. 3. The elements of the acoustic-phonetic recognition unit. The representations at different points of the system are shown in Fig. 4 and Fig. 5.

The input to the system is the speech wave form, see number 1 in Fig. 4, where the horizontal axis of each representation has the same time scale and the vertical lines indicate the manually marked phoneme boundaries. The smoothed output signal of the filter bank is connected to the inputs of the feature classification neural network, and the sampling interval is 10 ms. Compare representation 2 in Fig. 4, where each column describes one filter frame with 16 filter outputs. Low-frequency filters are at the bottom, and the size of each square is proportional to the filter output magnitude in dB. The filter cut-off frequencies are listed in Table III below.

The output of the feature net is a seven element vector describing manner and place of articulation, (compare Fig. 6 below). This output is shown by representation 3 in Fig. 4, where the sizes of the squares are proportional to the output activity of the feature net. The order of the features is the same as in Fig. 6, with the uppermost feature corresponding to feature number 1, voiceness.

The more detailed, phone classification neural net has a dual input window, which includes the spectral vector describing the actual moment of the speech and the output of the coarse feature net describing the speech context by seven frames centred at the spectral frame, see Fig. 5. The dual window slides past the speech material in a frame-by-frame fashion and the weighting of the inputs is automatically formed during the training of the phone network.

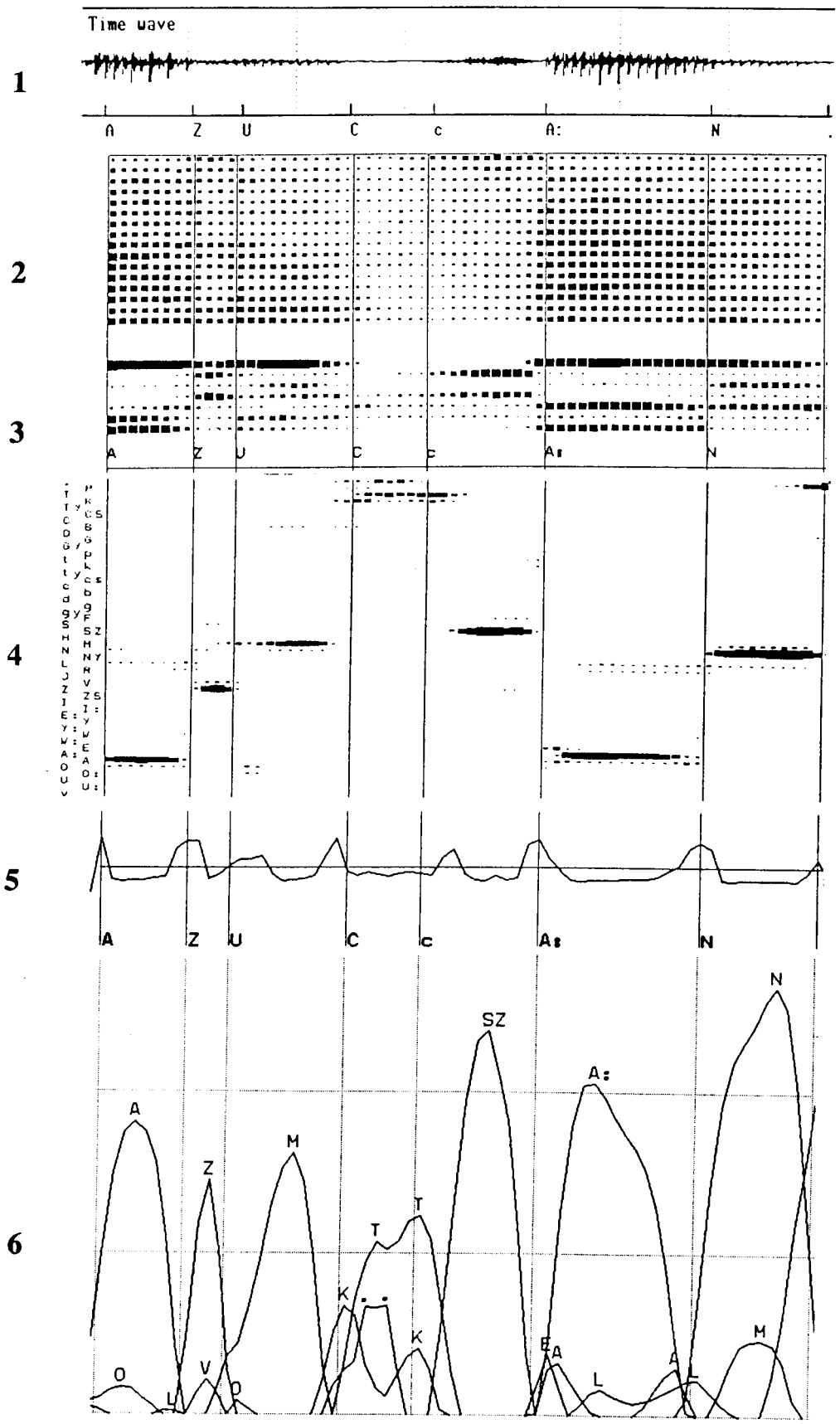

Fig. 4. Different representations in the acoustic-phonetic recognition system.

Ideally, only the output node of the phone net associated with the actually processed phone should have a high activation, while all the others should remain at the base level. An example of a real output of the net is shown by representation 4 in Fig. 4. The sizes of the rec tangles are proportional to the output activations of the phone network. It may be seen that only a few output nodes have high activation simultaneously. The most active node indicates the first phoneme candidate of the network, the next one corresponds to the second candidate and so forth.

Automatic segmentation of speech is performed by a segmentation network. The activation levels of only the first candidates of the phone net are used as input and the input is taken over a 15 frame wide window (150 ms). The single output of the network is expected to be zero except for an activation peak at the very first frame of each phoneme only (compare Fig. 12 below). An example is shown in Fig. 4, representation 5. Phoneme labels for the detected segments are based upon an evaluation of the smoothed output activations of the phone net. This completes the transition from time the time domain into the event domain, and the output can be seen at the bottom of Fig. 4.

Only the units D and H in Fig. 3 contain language specific elements. In the units E and G, the size of the data memory needed depends on the size of phoneme set, but the function and the other parameters are independent of language. Changing language requires a reshaping of the network structure of unit D to the new phoneme set and retraining the unit with a new speech material. The phoneme table in unit H needs to be altered as well.

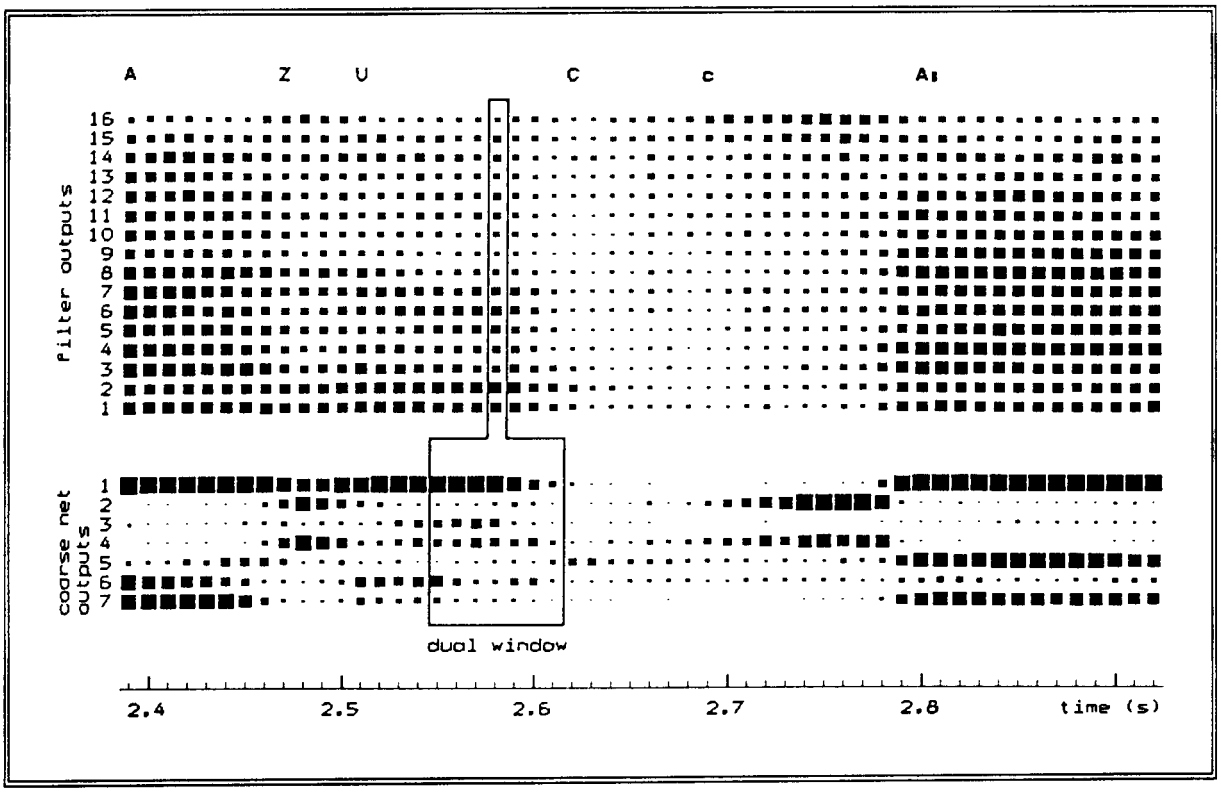

Fig. 5. *The dual input window of the phone network. It includes one section of the filter bank output and the coarse phonetic feature parameters in seven frames.*

3. STRUCTURE, PRINCIPLES AND LEARNING OF DIFFERENT NETWORK ELEMENTS

3.1 Filter bank

The filter bank simulation program is based on an FFT-procedure (Blomberg, 1989) using a Hamming window. The FFT calculates 1024 spectrum lines, and their energy is merged within the filter bands. The mel-scaled filter cutoff frequencies are listed in Table III.

Table III. Characteristic frequencies of the filter bank

Filter number	*Low frequency cutoff (Hz)*	*High frequency cutoff (Hz)*
1	188	294
2	294	406
3	406	527
4	527	661
5	661	809
6	809	977
7	977	1167
8	1167	1384
9	1384	1633
10	1633	1920
11	1920	2251
12	2251	2634
13	2634	3077
14	3077	3592
15	3592	4190
16	4190	4884

3.2 The coarse feature classification neural network

This processing unit has multiple tasks. The most important one is to transfer spectral amplitudes into phonetically related features describing manner and place of articulation. A key-problem is to select an appropriate set of coarse features. In spite of the existence of some different phonetic feature sets (Jacobson, Fant, Halle, 1963; Chomsky & Halle, 1968; Fant, 1973; Singh, 1976), a new feature set was constructed according to the following principles:

- the feature parameters should be detectable within a single spectral frame
- the feature set should be language and speaker independent
- the feature set should describe the manner and place of articulation
- the feature set should be usable for different phonetic classes
- the feature target values should be readable from an existing speech data base
- the total number of features should be small
- the features should conform to traditional phonetic feature sets in the steady state phase of the phonemes
- the feature set may be non-distinctive for phonemes, i.e., some phonemes may be identical in all features.

A fundamental problem is that the classical phonetic features are interpreted only at the phoneme level and we need a description for 10 ms frames. This means that direct use of features is difficult, but there are still good reasons to use something related to them. At a phoneme border, two adjacent spectral frames may be very similar although they are labelled as different phonemes having different phonetic features. From an acoustical point of view,

the difference may often be larger between the central frame and the boundary frames of a phoneme than between two adjacent frames on opposite sides of a phoneme boundary. The feature set is intended to capture some of these intra-phoneme variations, but the labels of speech data does not contain this information explicitly. However, the learning capacity of a neural network can be used for this purpose. The target values of the features are binary in the training phase, but by exposing the net to thousands of speech frames having varying acoustic representations for the same feature, we expect continuous feature values to develop that will convey also subtle details about each feature. This especially holds for the place of articulation related features.

As a result of a compromise among the principles listed above, a seven-element feature set was constructed based on manner and place of articulation related features. The manner related features are: voiceness, noiseness, nasalness and vowelness while the place related features are: frontness, centralness and backness, since these three can easily be assigned to both vowels and consonants. Since we expect the features to take continuous values, we use the suffix "*-ness*" in analogy with the term *loudness* for perceived sound level. Vowels have three features set positive: voiceness, vowelness and one place related feature. Consonants are described by one place feature and the relevant manner features (vowelness is of course set to minus). Tables IV and V show feature target values for Swedish and Hungarian phonemes, respectively. The features are not completely independent, and the targets are not able to discriminate between all the phonemes. However, this set is intended as a practical compromise.

Table IV. Coarse phonetic features for the Swedish phoneme set.

Swedish technical alphabet	*IPA*	*voice-ness*	*noise-ness*	*nasal-ness*	*front-ness*	*central-ness*	*backness*	*vowel-ness*
.	(pause)	-	-	-	-	-	-	-
P	p (occ.)	-	-	-	-	-	-	-
T	t (occ.)	-	-	-	-	-	-	-
K	k (occ.)	-	-	-	-	-	-	-
p	p (burst)	-	+	-	+	-	-	-
t	t (burst	-	+	-	-	+	-	-
k	k (burst)	-	+	-	-	-	+	-
B	b	+	-	-	+	-	-	-
D	d	+	-	-	-	+	-	-
G	g	+	-	-	-	-	+	-
M	m	+	-	+	-	-	-	-
N	n	+	-	+	-	-	-	-
NG	ŋ	+	-	+	-	-	-	-
R	r	+	-	-	-	+	-	-
L	l	+	-	-	-	+	-	-
V	v	+	+	-	+	-	-	-
J	j	+	-	-	-	+	-	-
F	f	-	+	-	+	-	-	-
S	s	-	+	-	+	-	-	-
SJ	ʃ	-	+	-	-	+	-	-
TJ	ç	-	+	-	-	-	+	-
H	h	-	+	-	-	-	+	-
I	ɪ	+	-	-	+	-	-	+
E	e	+	-	-	-	+	-	+
Ä	ɛ	+	-	-	+	-	-	+
Y	ʏ	+	-	-	+	-	-	+
Ö	ø	+	-	-	-	+	-	+
U	ɵ	+	-	-	-	+	-	+
O	ɷ	+	-	-	-	-	+	+
Å	ɔ	+	-	-	-	-	+	+
A	a	+	-	-	-	+	-	+
I:	i:	+	-	-	+	-	-	+
E:	e:	+	-	-	+	-	-	+
Ä:	ɛ:	+	-	-	+	-	-	+
Y:	y:	+	-	-	+	-	-	+
Ö:	ø:	+	-	-	+	-	-	+
U:	ɯ:	+	-	-	+	-	-	+
O:	ɷ:	+	-	-	-	-	+	+
Å:	o:	+	-	-	-	-	+	+
A:	ɑ:	+	-	-	-	-	+	+

Table V. Coarse phonetic features for the Hungarian phoneme set

	voiceness	*noiseness*	*nasalness*	*frontness*	*centralness*	*backness*	*vowelness*
. [pause]	-	-	-	-	-	-	-
B [b, occ.]	+	+	-	+	-	-	-
b [b, bst]	+	+	-	+	-	-	-
C [ts, occ]	-	-	-	+	-	-	-
c [ts, bst]	-	+	-	+	-	-	-
CS [tʃ, occ]	-	-	-	-	+	-	-
cs [tʃ, bst]	-	+	-	-	+	-	-
D [d, occ]	+	-	-	-	+	-	-
d [d, bst]	+	+	-	-	+	-	-
F [f]	-	+	-	+	-	-	-
G [g, occ]	+	-	-	-	-	+	-
g [g, bst]	+	+	-	-	-	+	-
GY [J;- ,occ]	+	-	-	-	-	+	-
gy [J;- ,bst]	+	+	-	-	-	+	-
H [h]	-	+	-	-	-	+	-
J [j]	+	-	-	-	+	-	-
K [k,occ]	-	-	-	-	-	+	-
k [k,bst]	-	+	-	-	-	+	-
L [l]	+	-	-	-	+	-	-
M [m]	+	-	+	+	-	-	-
N [n]	+	-	+	-	+	-	-
NY [j;n]	+	-	-	-	-	+	-
P [p,occ]	-	-	-	+	-	-	-
p [p,bst]	-	+	-	+	-	-	-
R [r]	+	-	-	-	+	-	-
S [ʃ]	-	+	-	-	+	-	-
SZ [s]	-	+	-	+	-	-	-
T [t,occ]	-	-	-	-	+	-	-
t [t,bst]	-	+	-	-	+	-	-
TY [c,occ]	-	-	-	-	+	-	-
ty [c,bst]	-	+	-	-	+	-	-
V [v]	+	+	-	+	-	-	-
Z [z]	+	+	-	+	-	-	-
ZS [ʒ]	+	+	-	-	+	-	-
A [ )]	+	-	-	-	-	+	+
A: [a:]	+	-	-	-	+	-	+
E [ɛ]	+	-	-	-	+	-	+
E: [e:]	+	-	-	+	-	-	+
I [i]	+	-	-	+	-	-	+
I: [i:]	+	-	-	+	-	-	+
O [o]	+	-	-	-	-	+	+
O: [o:]	+	-	-	-	-	+	+
W [Φ]	+	-	-	-	+	-	+
W: [Φ:]	+	-	-	-	+	-	+
U [u]	+	-	-	-	-	+	+
U: [u:]	+	-	-	-	-	+	+
Y [y]	+	-	-	+	-	-	+
Y: [y:]	+	-	-	+	-	-	+
v	+	-	-	-	+	-	+

A multi-layer perceptron net was trained using error back-propagation to recognize the coarse phonetic features from the spectral outputs of the filter bank. The network structure can be seen in Fig. 6. The net has 36 nodes (13 in the hidden layer) and 299 connections.

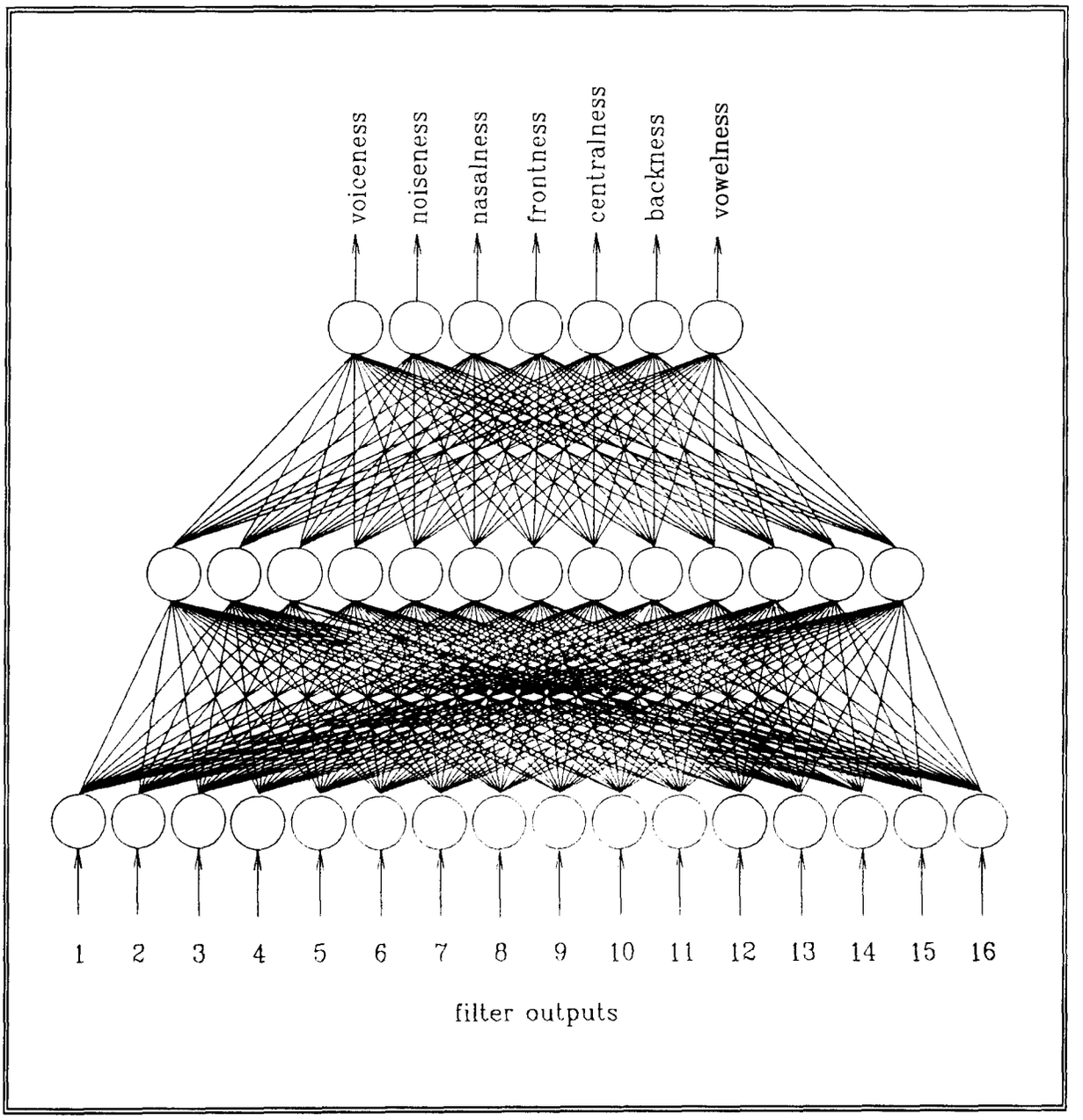

Fig. 6. The structure of the coarse feature network.

The filter amplitudes were scaled in the interval 0.0 - 1.0 to unify them with the dynamic range of the system. The feature target values were set to 0.1 or 0.9 for the plus and minus values in Tables IV and V, since this resulted in better convergence for the training than using 0.0 and 1.0 as targets. The sizes of the pattern files used are summarized in Table VI.

Table VI. Volume data for the different speech materials used.

Speech material (sentences)	*Purpose*	*Number of frames*	*Number of phon-emes*	*Input data file size (MByte)*			
				Coarse feature network	*Phone network*	*Base-line network*	*Segment-ation network*
INTRED 1-50	training	15653	2044	1.993	9.843	4.665	0.701
INTRED 51-60	testing	2422	354	0.307	1.566	0.722	0.230
JONSSON 1-10	training	2327	309	0.295	1.508		0.221
MAMO 1-50	training	21680	2796	2.748	14.921		2.072
MAMO 51-60	testing	4363	548	0.553	3.022		0.417

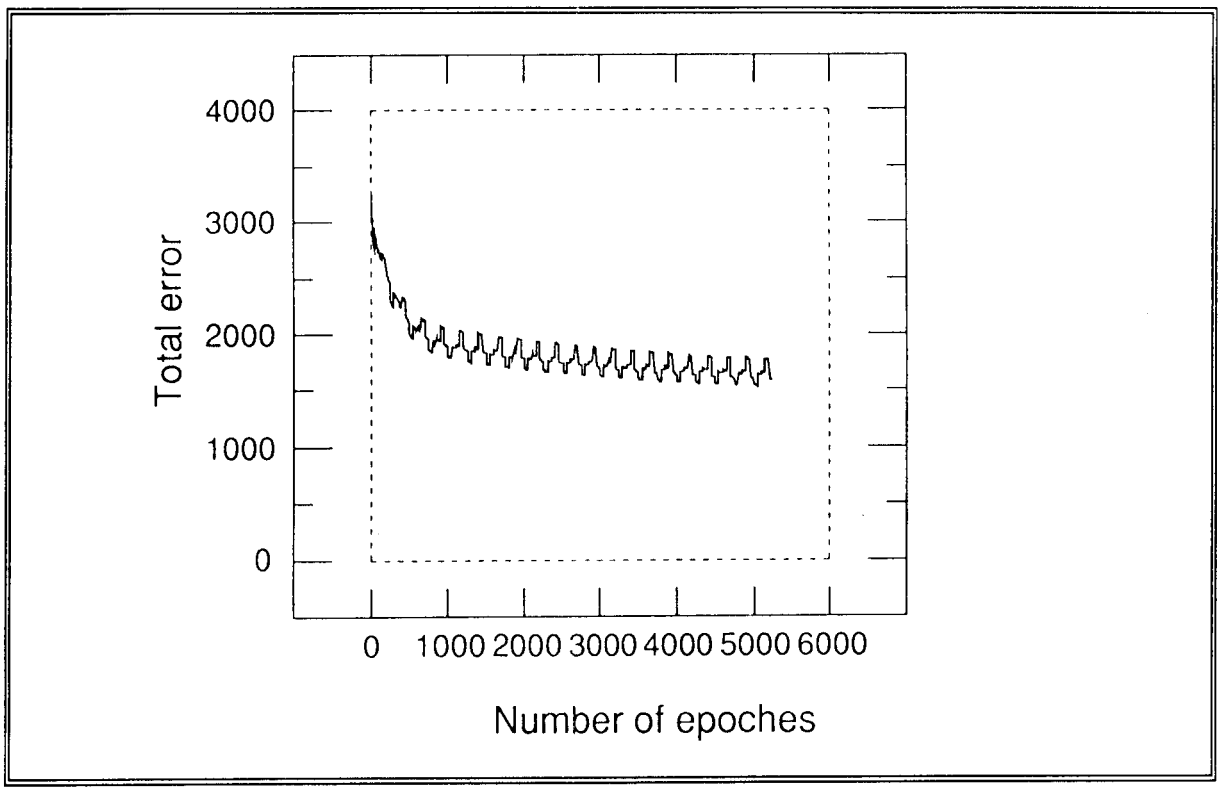

Fig. 7. The total error as a function of the number of epoches during the training of the coarse feature network. The fluctuations are caused by periodically changing the training material between five different training partitions. One epoch in this figure means one presentation of all the patterns in one partition only and the error is also related to one partition at a time.

The fastest learning without getting disturbing oscillations in the output error was achieved when using a learning rate of 0.001 and a momentum term of 0.9. The input frames were chosen at random and the weights were updated after each presentation. The recognition results for the test set stabilized after a couple of thousand epoches. The total error as a function of the number of epoches is shown in Fig. 7. The training material was subdivided into five partitions that each could fit into the working memory of the computer. This speeded up the training by substantially reducing disk accesses. Each training fragment was used for a period of fifty epoches after which it was substituted by the next one. The fluctuation in the total error in Fig. 7 is due to this. One epoch normally means one presentation of each pattern in the training set, but since the training material was divided into five partitions one epoch in this Figure only means that one fifth of the training patterns has been presented to the net.

3.3 The phone network

The phone network receives two types of input using a dual window (see Fig. 5). They are different both in content and in function. The first input is a seven frame window covering the coarse phonetic features. It contains some redundant information, since during steady states adjacent speech frames have similar feature values. Data compression can also be made by reducing the number of features. It is reasonable to compress these data before the final phoneme decision in order to reduce the network complexity and to decrease the number of weights to be trained. The compression is combined with a sort of time warping function to compensate for tempo changes in phoneme pronunciations. Both the compression and the time compensation function are realized by a hidden layer with a special connection structure, as discussed below. The second input to the phone net is the filter bank spectrum of the frame being processed, and the net should mix the two kinds of input data appropriately.

The number of output nodes in the network (the size of the phoneme set) is also the result of a compromise. The result can be seen in Table IV for the Swedish material and in Table V for the Hungarian material. A particular feature of both sets is to treat the occlusion and burst phase of stops and affricates as two different segments, where both segments have a specific label. Both phoneme sets have a special class for silent intervals marked by ".". These segments occur mostly at the end of the sentences. The vowels are represented by short and long pairs. The "v" symbol marks the neutral vowel.

According to the discussed requirements, the phone network consists of four layers: one input layer, one compression hidden layer, one "mixing" hidden layer, and one output layer. The network topology can be seen in Fig. 8. As a result of separating the compression and mixing functions, the connection structure is quite complex. Each column of four nodes in the compression layer is connected to a group of three successive columns (frames) of the seven feature nodes, which are the output of the coarse feature net. There is a one frame overlap in the coarse feature columns connected to the compression node columns. This means that all three compression columns will cover seven coarse feature columns. The rationale for this is that this will make them more insensitive to tempo changes in the speech.

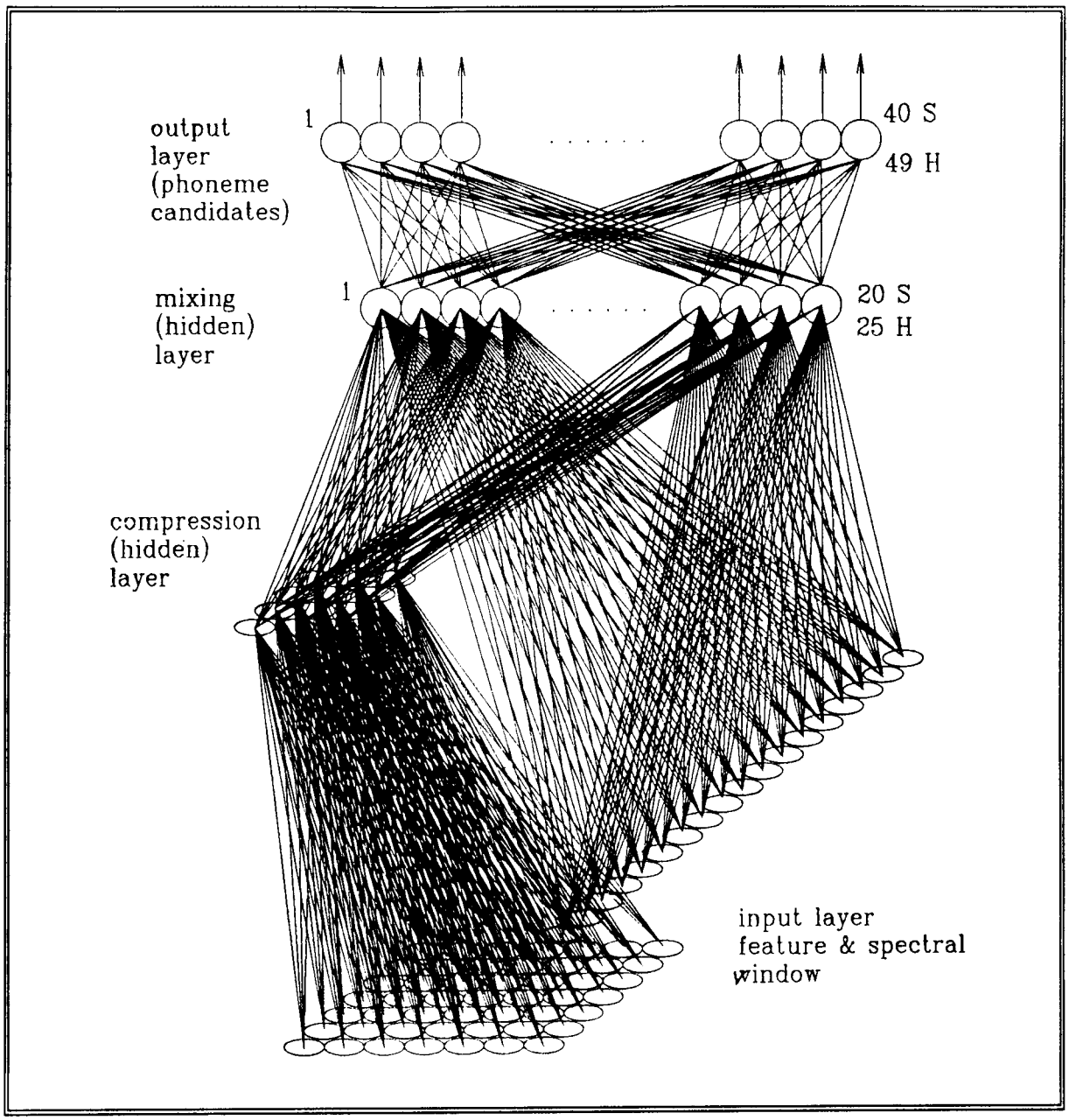

Fig. 8. *The structure of the phone classification network. The Swedish network has 20 nodes in the mixing layer and 40 output nodes. The Hungarian network has 25 and 49 nodes, respectively.*

The number of output nodes is determined by the phoneme inventory of the language being processed. The Swedish phone network has 40 output nodes and the Hungarian has 49. The number of nodes in the mixing layer is proportional to the size of the phoneme set. The mixing layer of the Swedish network has 20 nodes, the Hungarian 25. There are no differences in the lower level layers of the phone networks. The Swedish net has a total of 137 nodes and 1612 connections and the Hungarian has 151 nodes and 2177 connections. Fig. 8 shows only the first four and the last four nodes of the mixing and output layers. Some characteristics of the data files used for training and testing the phone network are summarized in Table VI.

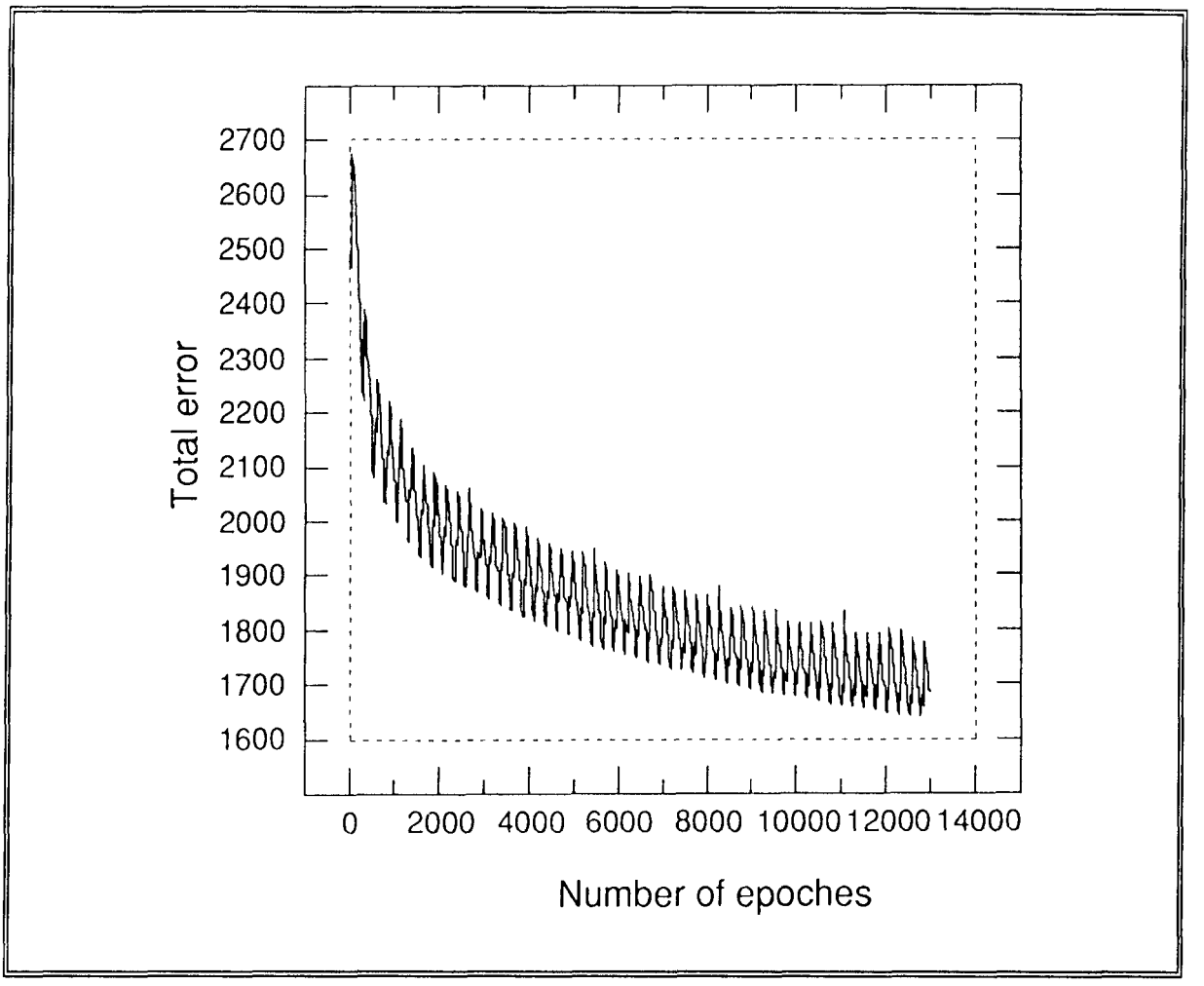

Fig. 9. *The total error of the phone net as a function of the number of training epoches for the MAMO material. The fluctuations are caused by periodically changing the training material between five different training partitions. One epoch in this figure means one presentation of all the patterns in one partition only and the error is also related to one partition at a time.*

The fastest learning of the phone net was reached at a learning rate of 0.001 and a momentum term of 0.1. The weights were updated after each pattern. The training material was divided into the same five partitions used for the phone net in order to speed up the training time (still the training time was 100 hours CPU-time for the INTRED material and 250 hours for the MAMO material on an Apollo DN10000). As can be seen in Fig. 9, the total error has not reached a stable minimum value – the training process was interrupted due to running time limitations. On the other hand a low output error may be an indication of over-training, why it´s hard to know from this figure only whether the training time was too short. The oscillations are caused by the subdivision of the training set, as noted above. The number of "full" epoches for the training set is one fifth of the values in the Figure.

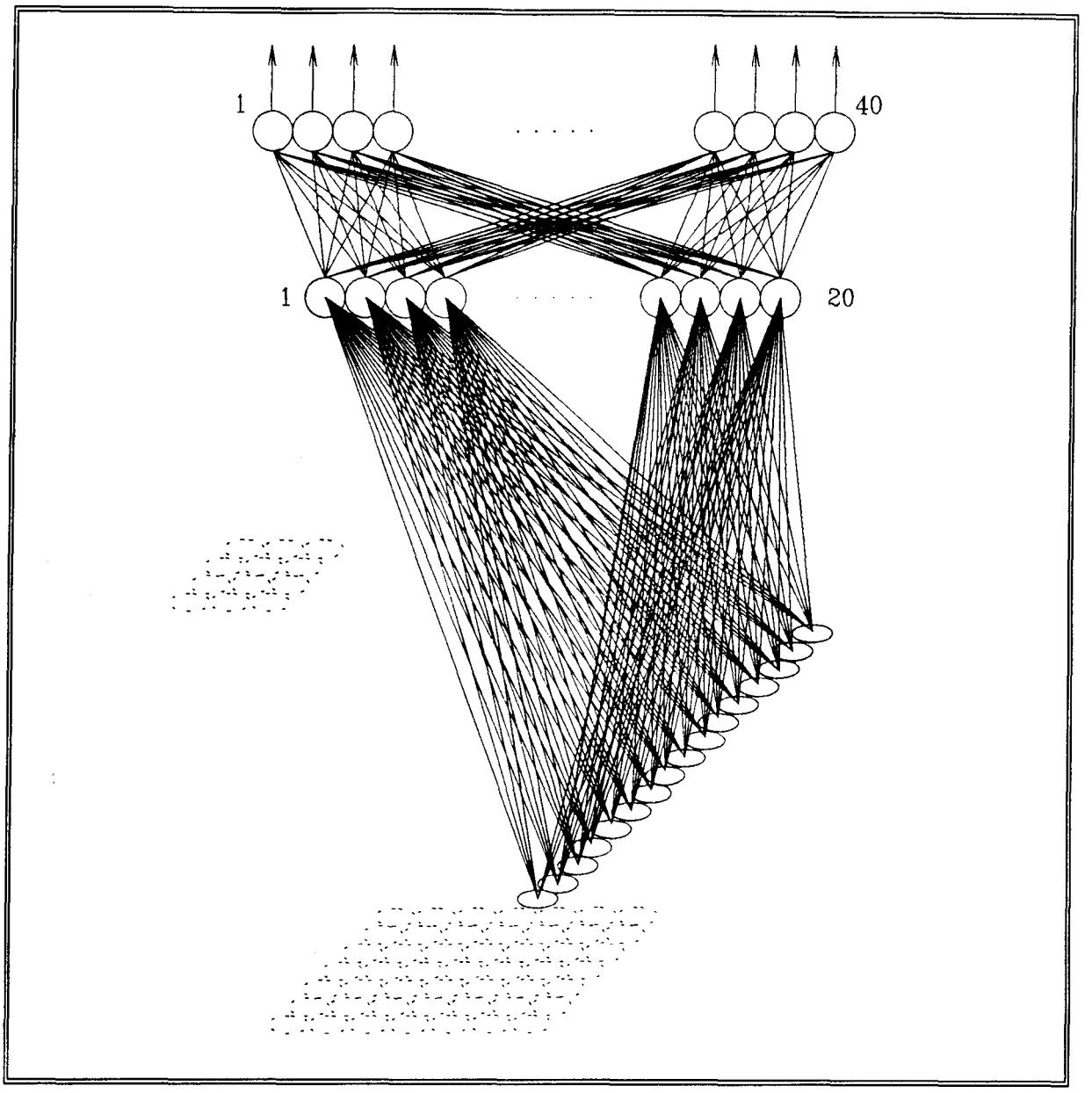

Fig. 10. The base-line reference network. The compression layer and the feature window of the phone net have been removed.

A base-line reference network was constructed in order to test the effect of the dual input window in the phone net. The input layer of the reference net is identical to the corresponding part of the phone network. The compression layer and the feature window have been removed, but the hidden layer is exactly the same as the "mixing layer" in the phone network. The output layer and the connections to it are also identical to the Swedish phone network. The topology of the net can be seen in Fig. 10.

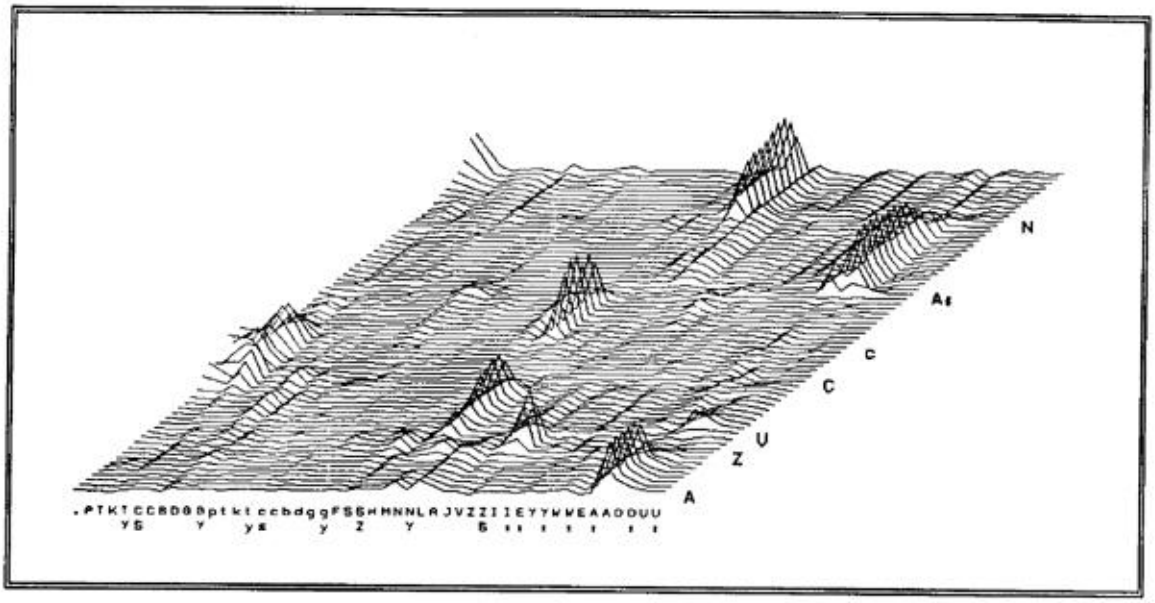

Fig. 11. The output activations of the phone network represented as a surface in a three-dimensional space: time, phoneme set, and output activations. The manually decided phoneme labels are displayed along the time axis. Each ridge refers to one phoneme.

3.4 Segmentation neural network

The output activity of the phone classification network is shown as a surface in a three-dimensional space in Fig. 11. The three axes are: the time, the phoneme set and the output activations. The target phoneme labels are marked along the time axis. The activation surface is typically flat with some "mountain ridges" running parallel to the time axes. Each ridge represents one phoneme. Multiple ridges in parallel indicates the existence of simultaneous phoneme candidates for that segment. The ridges are generally well separated in time and this implies that they might be used as a basis for segmentation. But it is not directly evident how to utilize them. The task to establish the decision criteria was again solved by a multi-layer network. The input parameter to the segmentation network is the amplitude of the phoneme output having the highest activation within each frame inside the input window that is 15 frames long (150 ms). The window is symmetric with seven frames on each side of the actual frame. The length of the window is selected to normally include at least one phoneme border. The segmentation network has a single output trained to show high activity (0.9) for the first frame of each phoneme and low activation (0.1) for all other frames – it is set up to "fire" at phoneme borders. The structure of the net is shown in Fig. 12. Table VI shows some data on the segmentation file sizes. The training was done with a learning rate of 0.005, a momentum term of 0.1, and the weights were updated after each pattern.

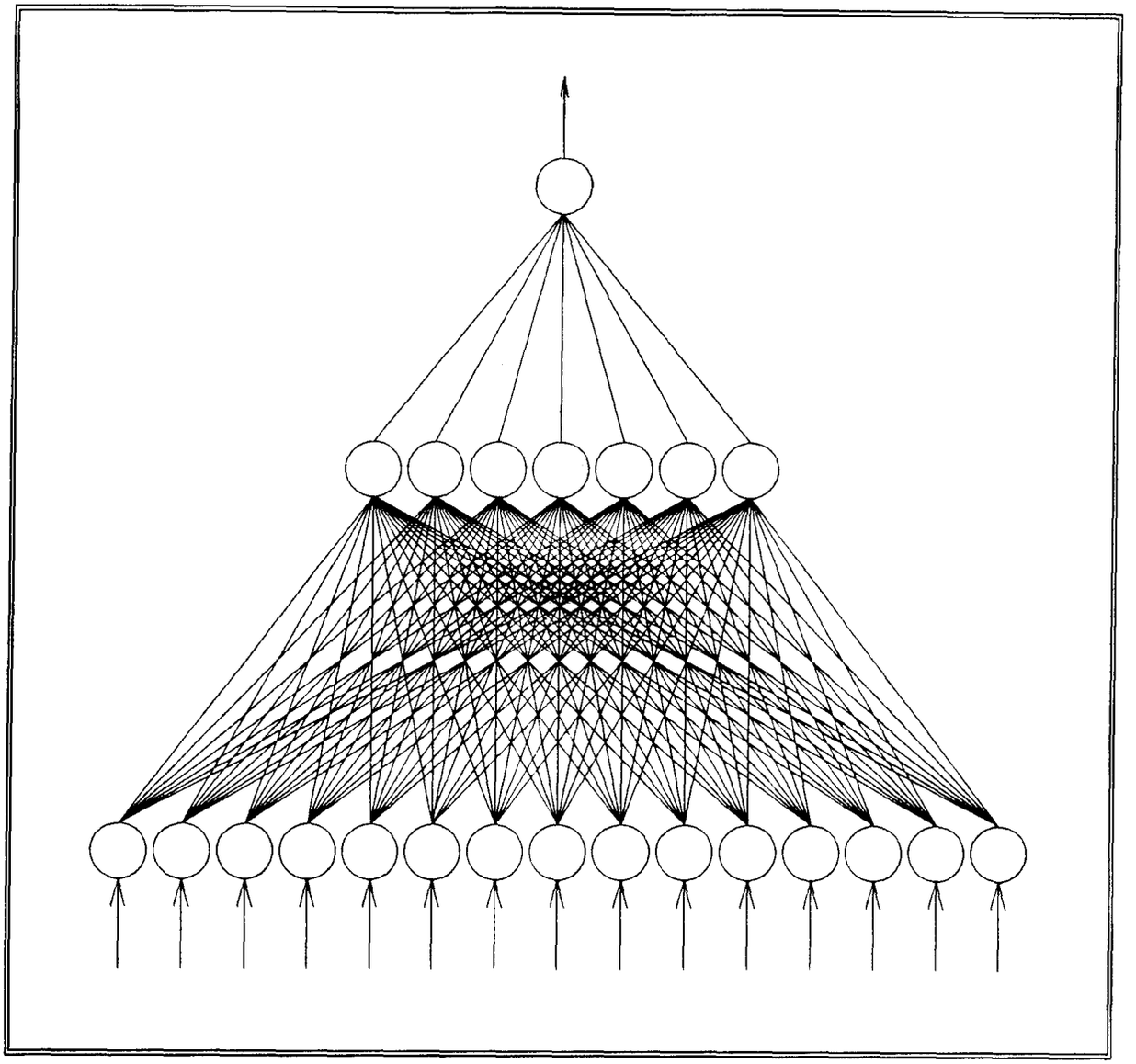

Fig. 12. The structure of the segmentation network with 15 input nodes, 7 hidden nodes and 1 output node.

3.5 Phone activation filter

It seems evident that the phoneme candidates for a segment should be selected among the candidates having the highest output activations. However, this will result in a lot of recognition errors due to spurious, short activation peaks. According to our experience, smoothing the activations by using a simple mean value filter gives significantly better recognition performance. The filter parameters were calculated empirically to produce optimal recognition. The first, second, and third phoneme candidates for each recognized segment were selected according to the level of their smoothed activation peaks. These phonemes constitute the output of the complete acoustic-phonetic level processing.

4. RESULTS

The performance statistics reported below are based upon ten independent test sentences both for Swedish and Hungarian, compare Table VI. The speech material used for training was never used for testing the performance of the different networks. However, sometimes we give some results on the training set to indicate the generalization performance of the training.

4.1 Recognition of coarse phonetic features

A typical output of the coarse feature network may be seen in Fig. 13, where the output activations of the seven coarse features are plotted as a function of time. The segmentation and labelling was performed by a human phonetic expert. The straight horizontal lines indicate a 0.5 activation level. We notice steep transitions for the voiceness, noiseness and nasalness features as well as parts of the vowelness feature. These features almost behave like binary signals. The (nonlinear) network has become sensitive to the gradual filter bank variations and is usually able to follow the discrete target values quite reliably. Some exceptions to this can be found in Fig. 13, where the phoneme "Z" shows an uncertainty in the voiceness feature, maybe due to strong frication, and in the phoneme "R", where the vowelness is somewhat ambiguous, indicating its character of a semivowel. Most of the time the feature transitions are in good synchrony with the manually set boundaries. Features related to the place of articulation generally vary at a slower rate and these parameters frequently demonstrate intermediate values in spite of being binary during the training. The continuous distribution of these parameters agree with our original intentions, since the articulation features covered by these parameters have a continuous character, and we expected the net to learn this by being exposed to numerous different varieties of these features during the network training. Some slow transitions for these parameters can be seen in Fig. 13, e.g., in the "AN"-sequence on the right hand side.

Table VII. Performance of the coarse phonetic feature network on the frame level when evaluating the feature activations as binary signals.

feature	*percent correct feature recognition*	
	INTRED	*MAMO*
voiceness	93.1	93.3
noiseness	91.0	92.9
nasalness	95.4	93.1
frontness	81.7	88.4
centralness	83.2	80.8
backness	88.7	88.2
vowelness	88.2	88.0
all features correct	76.9	80.0

The performance of the coarse feature network was tested by using three different methods. In the most evident evaluation, a binary signal was formed by the output activations using a comparison level of 0.5 – if the signal exceeded this threshold the corresponding feature was set to one and otherwise it was set to zero. These features were then compared to the features derived from the manual phoneme labels for each of the tested frames. This binary evaluation of the frame level recognition rate is summarized in Table VII. The network has a closely similar performance for both the INTRED- and the MAMO-material. Many features are correctly recognized for more than 90% of the frames and all features are recognized

above an 80% level. The manner of articulation related features perform better than the place of articulation features. The last row gives the results when all the features of a frame are correctly recognized. Most errors occur at phoneme transitions. The results show that in a substantial majority of the frames, the selected discrete phoneme features are detectable by the neural network at the frame level. Testing on the training set results in a 78.8% score for frames having all features correct, which is just 2% higher than the results for the test set, indicating a good generalization for the coarse feature network.

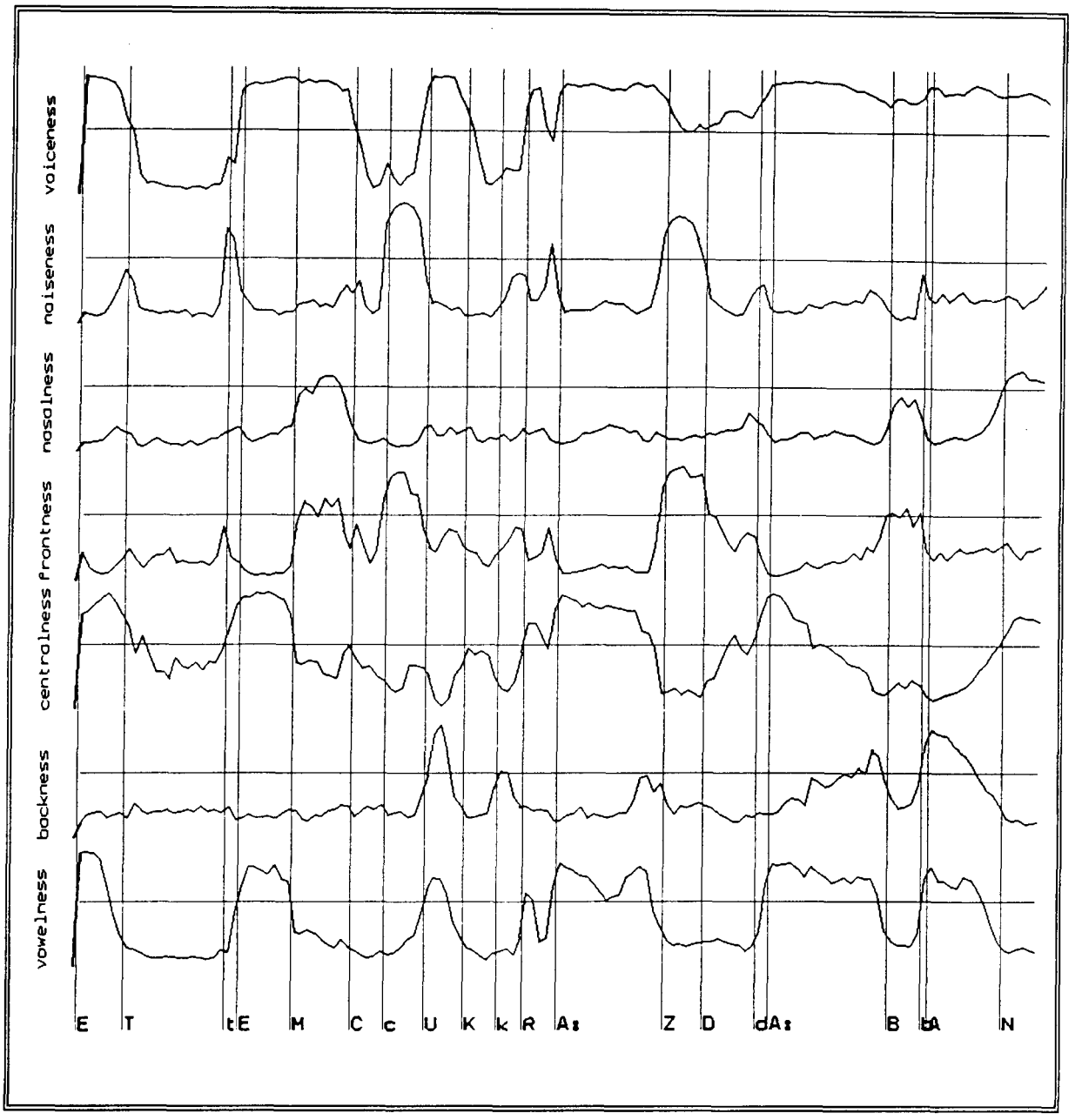

Fig. 13. Output activations of the coarse feature network as a function of time. Manually labelled segments. Horizontal lines indicate a 0.5 activation level for each feature.

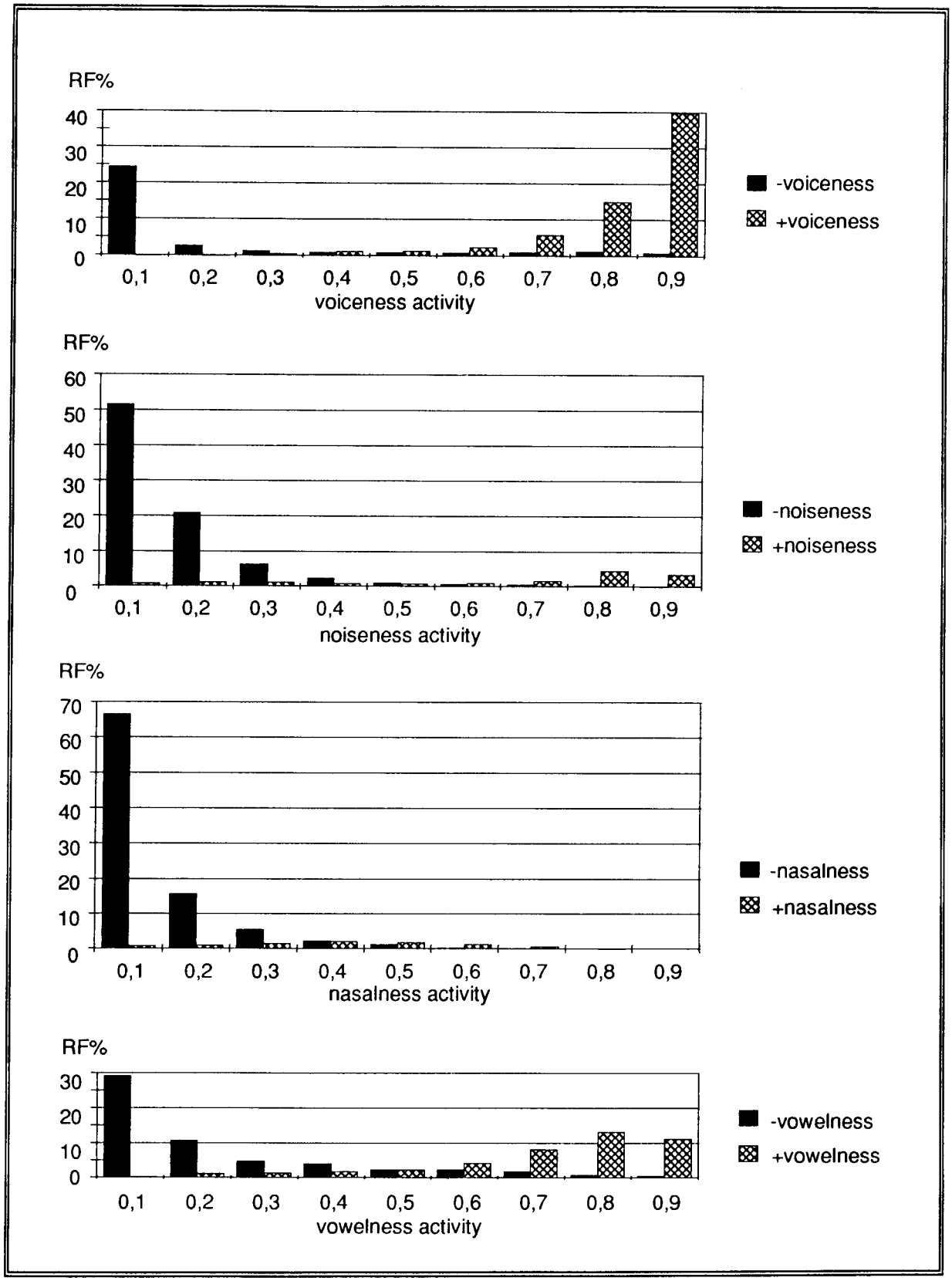

Fig. 14. Distribution of the manner of articulation related coarse feature activation levels in relative frequencies for the MAMO-material. Evaluation based on (all) frames in the test set. RF% means relative frequency in percent. The legends: +voiceness, +noiseness, +nasalness and +vowelness denote frames trained with a 0.9 target value of the respective feature while complementary frames (e.g., -voiceness) used a 0,1 target value of that feature.

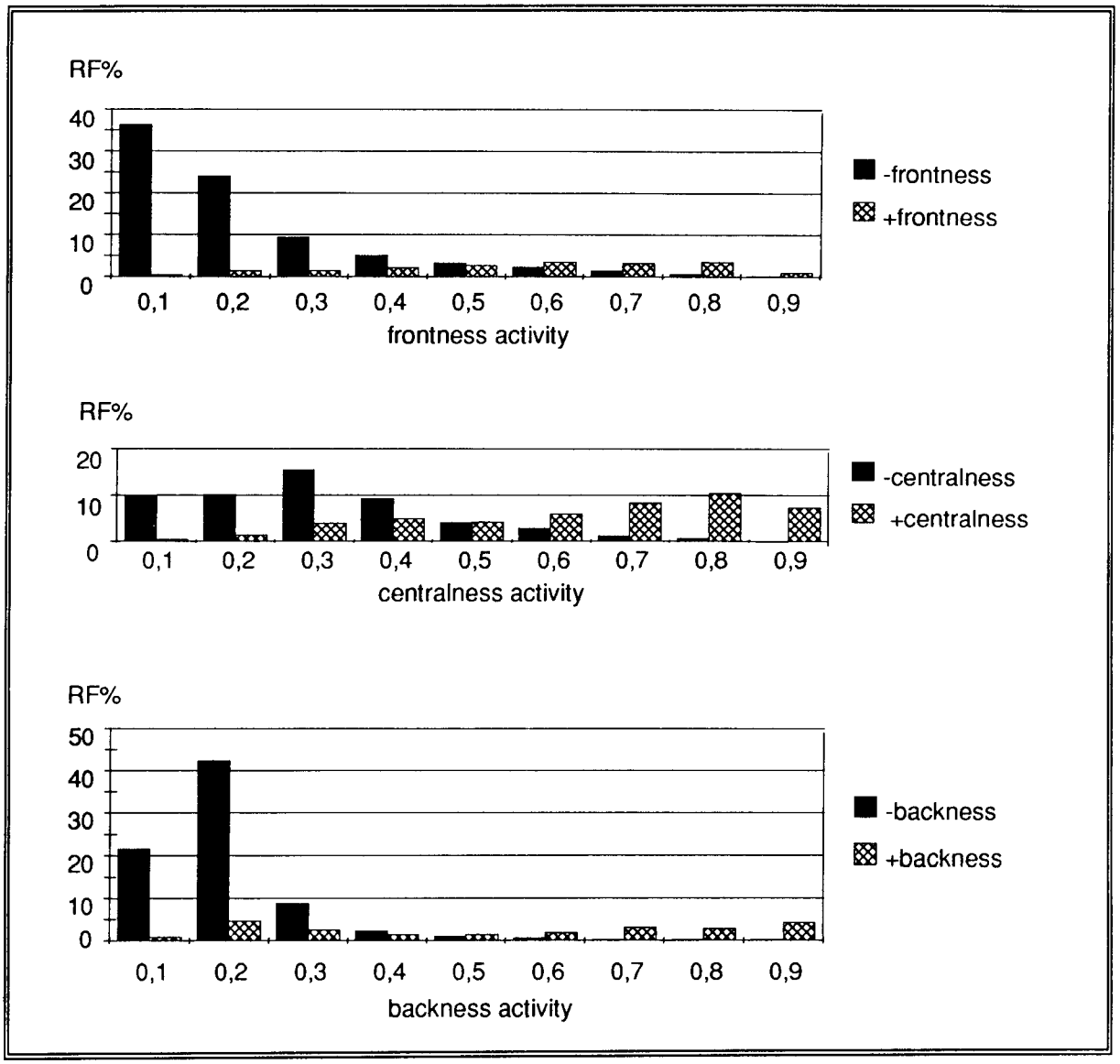

Fig. 15. Distribution of the place of articulation related coarse feature activations levels in relative frequencies for the MAMO-material. Evaluation based on (all) frames in the test set. RF% means relative frequency in percent. The legends: +frontness, +centralness and +backness denote frames trained with a 0.9 target value of the respective feature while complementary frames (e.g., -backness) used a 0,1 target value of that feature.

The second evaluation of the coarse feature net concerns the distribution of the network output activations. Results for all of features in the MAMO-material may be seen in Figs. 14 and 15. The target values for the output activations were set to 0.1 or 0.9, since this resulted in better convergence for the training. The diagrams show distribution statistics of complementary features, i.e., +voiceness and –voiceness in relative frequencies, i.e., each bar shows the percentage of the frames having the indicated activation level in relation to the total number of frames in the test material. This means that the sum of the bars are 100% when adding them for each feature and its complement. Fig. 14 shows that the distribution statistics of complementary features are well separated according to their target values during training. There is a very small overlap along the activation scale for the voiceness, noiseness and nasalness features, and the envelopes of the relative frequencies cross each other very close to a 0.5 activation level. This fact indicates that using a threshold of 0.5 when producing the above-mentioned results was correct. Another consequence of the polarized distribution is

that a nonlinear quantizer can describe the coarse features using only a 2 - 4 bit resolution. This fact is important when regarding hardware implementations of the net. The place of articulation parameters in Fig. 15 are not as well separated and this is especially true for the centralness feature. This is in accordance with their more continuous distribution.

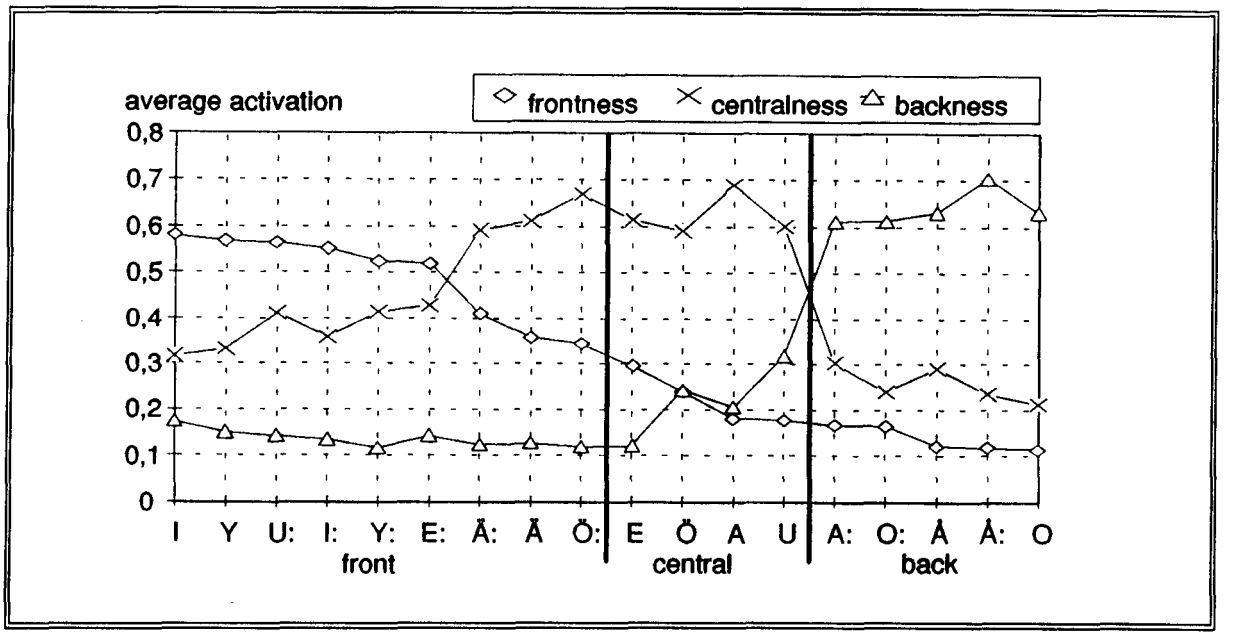

Fig. 16. Average output activations of the frontness, centralness and backness features based on the frames of the INTRED-test-material. All vowels have been trained with a 0.9 target value for either frontness, centralness or backness, while the target value for the other two place related features has been 0.1.

The third assessment of the feature net is related to parametric evaluations of the articulation place related features. The average activation levels of the frontness, centralness and backness features for the most frequent vowels may be seen in Fig. 16 for Swedish and Fig. 17 for Hungarian. During the training process all vowels were assigned to one of the three classes: front, central, or back. Thus, the vowels in each of these groups have the same target features. In spite of these binary and non-distinctive target values, the output activations of the feature net separate the vowels quite well using the continuous-valued place features. This indicates that on the basis of the thousands of phonemes used for training, the feature network develops a "sense" for the amount of frontness, centralness and backness in each vowel, which is expressed by the values of the scaled activations. The vowels have their maximum activation according to their targets during training, except the Swedish Ä-, Ä:- and Ö:-vowels, that are classified as more central than front by the network. Both Ä-vowels are probably influenced by the spectrally very similar and about five times more frequent E-vowel, that is labelled as central. There are only minor differences in the first two formants. As for the Ö:-vowel, it may have been affected by the similar short Ö, that is more frequent and labelled as central. One could. of course, argue that these vowels should be labelled as central rather than front, considering their spectral appearance, and this would be an interesting thing to do. The front-back distinction is, however, taken from the articulatory domain.

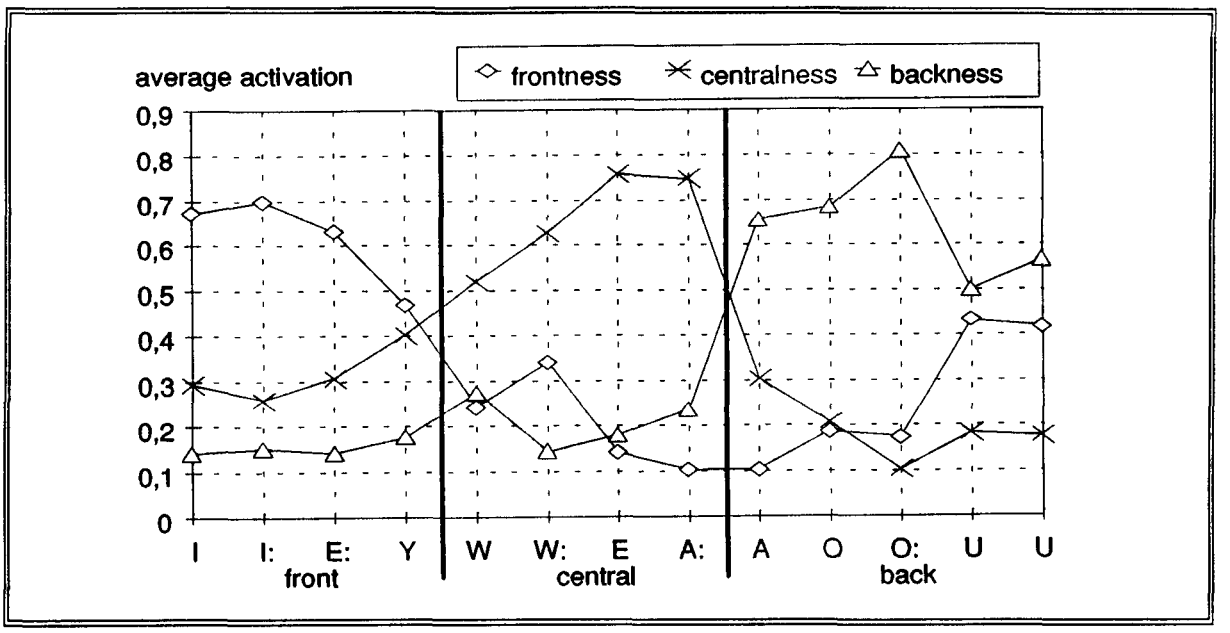

Fig. 17. *Average output activations of the frontness, centralness and backness features based on the frames of the MAMO material. All vowels have been trained with a 0.9 target value for either frontness, centralness or backness, while the target value for the other two place related features has been 0.1.*

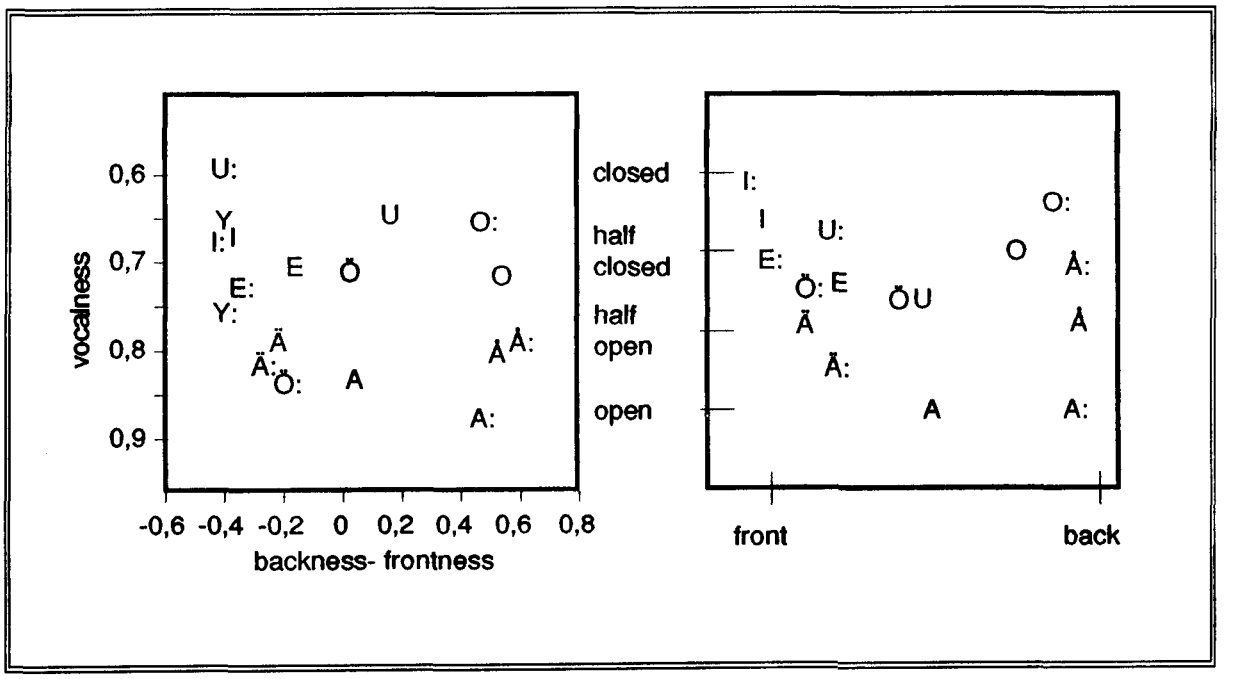

Fig. 18. *The coarse features related to place of articulation compared to a standard phonetic description of the Swedish vowels (Table IV).*

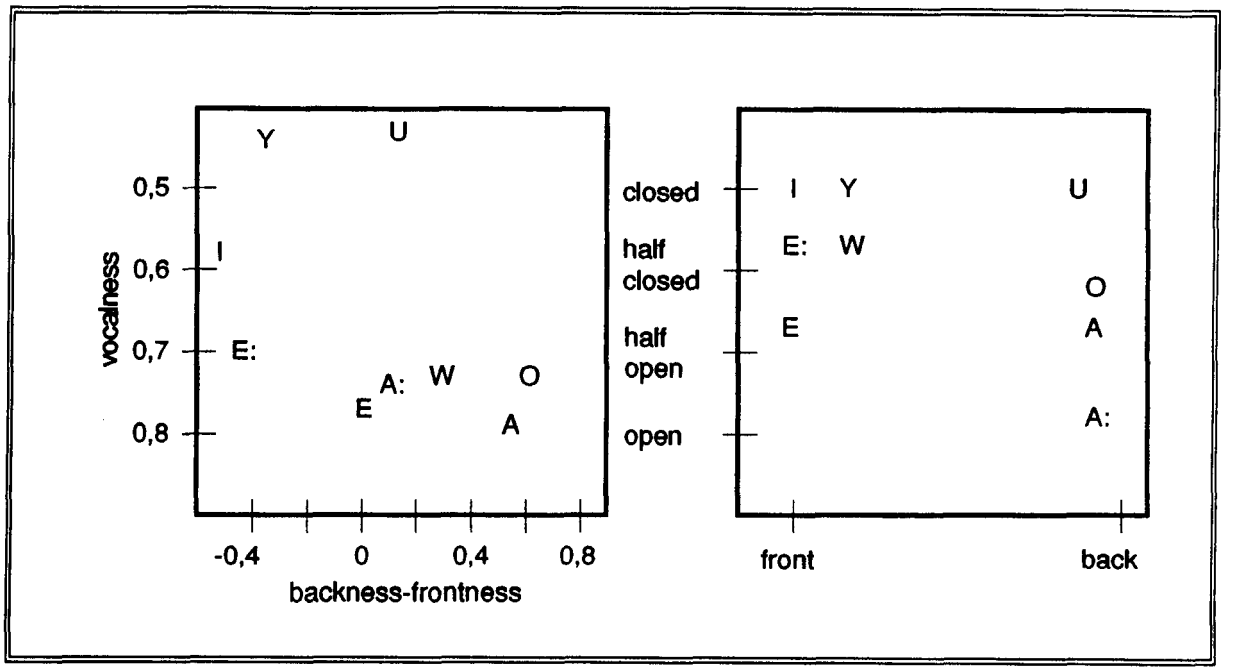

Fig. 19. The coarse features related to place of articulation compared to a standard phonetic description of the Hungarian vowels.

The relation between a classical phonetic representation of vowels and some coarse feature parameters is shown in Figs. 18 and 19 (Elert, 1989). The network representation of the vowels has a clear resemblance to the phonetic classification into the dimensions high/low and front/back. The agreement in the front/back-dimension is rather natural considering that most vowels have been assigned one of these features during training. The high/low-dimension is, however, not explicitly used in the training. Still, it seems that the net modulates the vowelness feature to do almost the same distinctions. This feature is the only one common to vowels and no other sounds. Values used are all above 0.5, which means that the intended use of this parameter as a vowel label also is fulfilled.

4.2 Phoneme classification on the frame level

The phone network outputs an ordered list of phoneme candidates for each speech frame according to the level of their output activations – the best candidate has the highest activation. It should be noted that this decision is based on a relatively wide window (150 ms). Each frame has a target label selected by a phonetic expert making it easy to compare the assigned target phoneme to the network candidates. After the training process, 54.2% of the frames in the INTRED-test-material and 54.7% of the frames in the MAMO-material were recognized correctly. Testing on the training set for the INTRED-material resulted in 58.6% recognition, indicating a relatively good generalization also for the phone net but also suggesting that the performance might be improved somewhat by continued training. It should be noted that all frames are included in these results – also transitional parts of the speech, where the acoustic character of the phonemes is changed due to coarticulation. The nature and the origin of the errors are treated in the next section.

There is a close relation between the frequency of occurrence of each phoneme and its recognition rate. This is obviously a training effect and the relation is shown in Fig. 20, which is based on the MAMO-material. The correct classification rate always exceeds 50% if the phoneme is represented by more than 200 frames in the test material, or approximately 5% of all frames. Phonemes never recognized correctly are represented by less than 2% (80 frames) in this material and together they represent 15% of the frames. The same relations hold for

the training material as well, keeping in mind that it is five times larger. For the INTRED-material the corresponding values are: phones never recognized correctly all have less than 1.5% of the frames and together they represent less than 10% of all frames, phones represented by more than 4% of all frames have a recognition score of more than 50%. The lower values for Swedish are not surprising, since there are 9 phones less to discriminate between. These results may indicate that the training material should be chosen to have a more even phoneme distribution. This would certainly give a better performance for less frequent phonemes. However, it would be at the expense of an increased error rate for the more frequent phonemes, and we cannot be sure about the total effect.

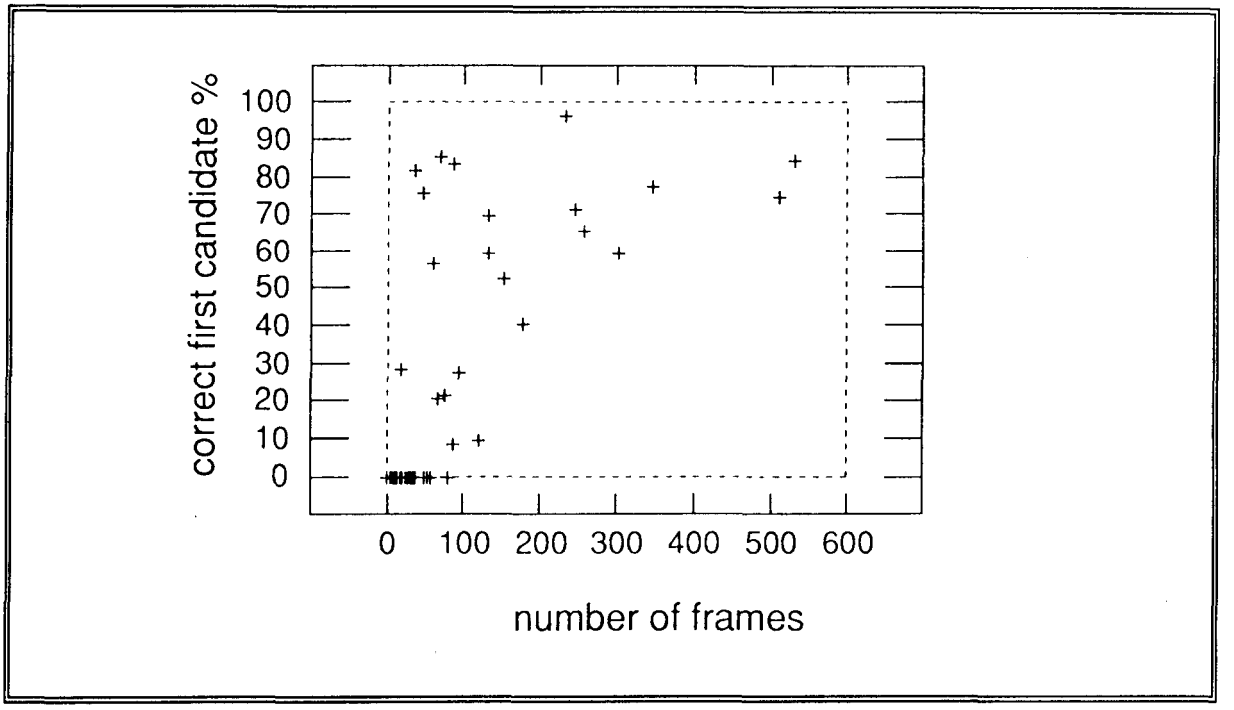

Fig. 20. Relation between the occurrence of frames in the test set and the recognition rate for the Hungarian phonemes.

The training of the phone net was probably not completed when it was stopped due to time limitations. Some typical patterns can be seen in Fig. 21. The most frequent phoneme frames in the Hungarian material (E,A) have a very good recognition rate in the very beginning of the training process. However, not only the correct phonemes but almost all less frequent phonemes are classified as belonging to some of the most common ones in the beginning. As the net starts to differentiate among more phonemes, the performance for the most common phonemes goes down, while the net becomes more and more sensitive to the low frequency phonemes. Their recognition rate starts to increase at different later phases of the training procedure and the increase has a different speed for different phonemes. The net result is that the overall recognition rate is monotonously increasing during the training time, though the rate is rather small in the end. It would, of course, be interesting to continue the training process, but it is very time consuming on our current facilities.

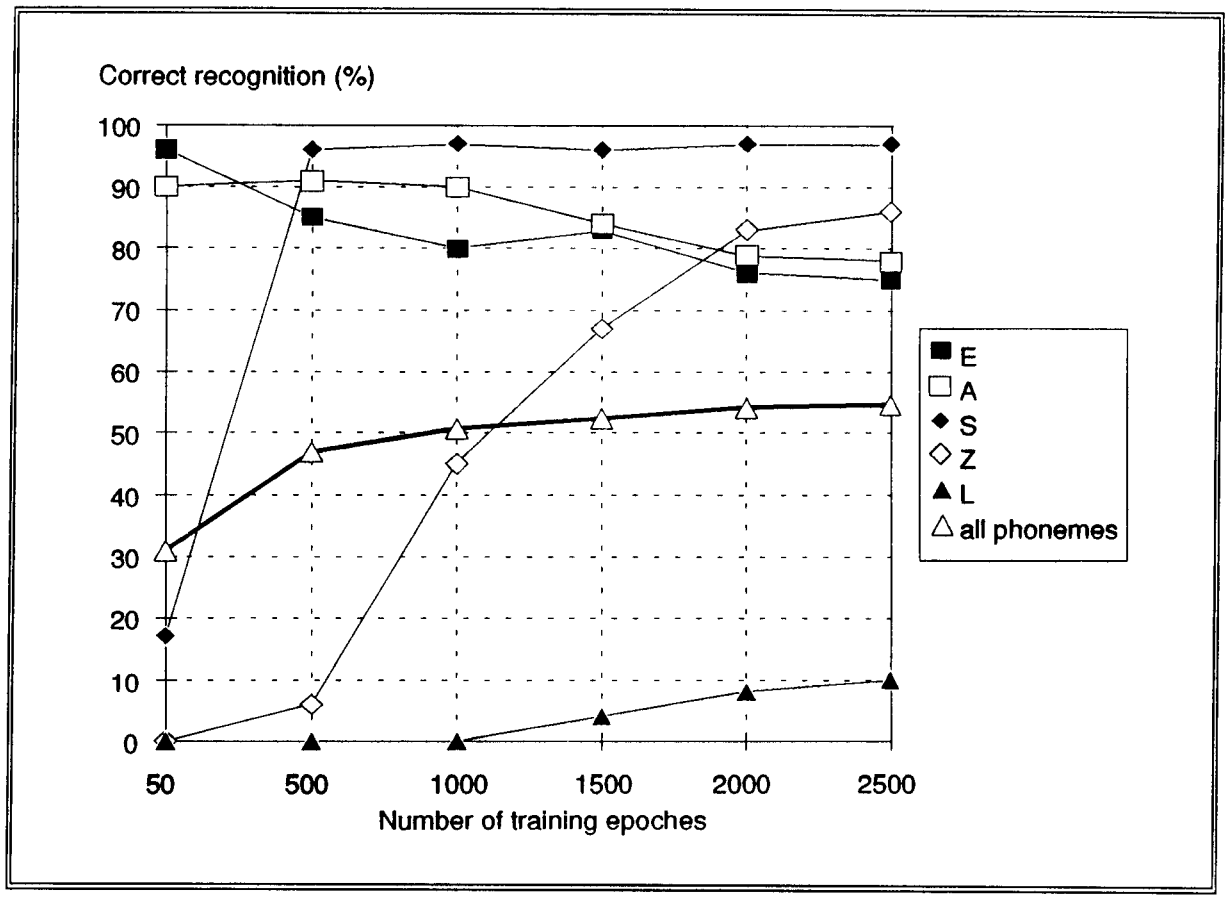

Fig. 21. *The evolution of the phoneme recognition performance as a function of the training time for the Hungarian material. Each epoch means one presentation of all the training material.*

Phoneme substitutions are summarized by the confusion matrices in Figs. 22 and 23. The area of the black squares is proportional to the corresponding number of occurrences. The sum of the values in each row is normalized and the sum over each row is a completely filled square. All of the most frequent confusions are caused by acoustic similarities.

In the confusion matrix for the Swedish INTRED-material in Fig. 22, the occlusion phase of the voiceless stops [P, T, K] form a separated group as well as the burst phases [p, t, k]. The very frequent t-phone has a recognition score of 84% for the occlusion phase and 69% for the burst phase. The voiced stops show no clear grouping and are mixed with many other sounds but the nasals are clearly grouped together – the least frequent NG is substituted by the other two, M and N, which perform relatively well, with a 59% and 73% recognition score. The laterals R and L form another group and the voiced fricative J is mostly confused with the rather close vowel E. The voiceless fricatives, F and S, are recognized with 67% and 83%, while the voiceless fricatives SJ and TJ have too few frames in the test material.

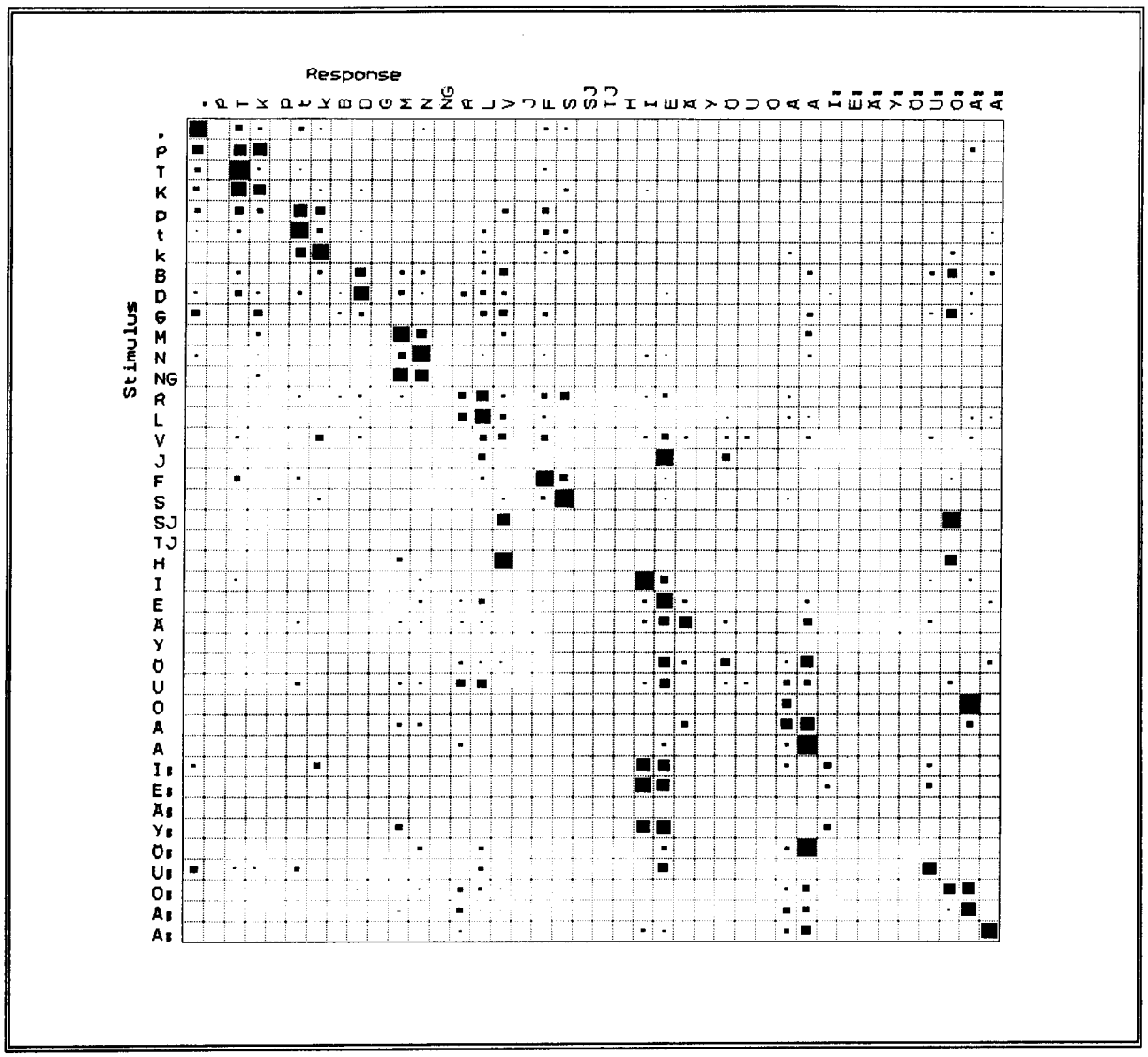

Fig. 22. Phoneme substitutions on the frame level based on the Swedish INTRED-material. The values are normalized for each row. The area of the black squares is proportional to the number of occurrences in each position..

Among the vowels we notice two groups containing the front vowels: I, E, Ä, and the back vowels: Å and A. The long front vowels I:, E: and Y: are recognized as their much more frequent "relatives", I and E. Vowels with the best recognition rates are I-75%, E-61%, A-82% and A:-62%. By ignoring length differences in vowel recognition, e.g. not counting an E: recognized as E as an error, the overall recognition rate would improve by about 2%.

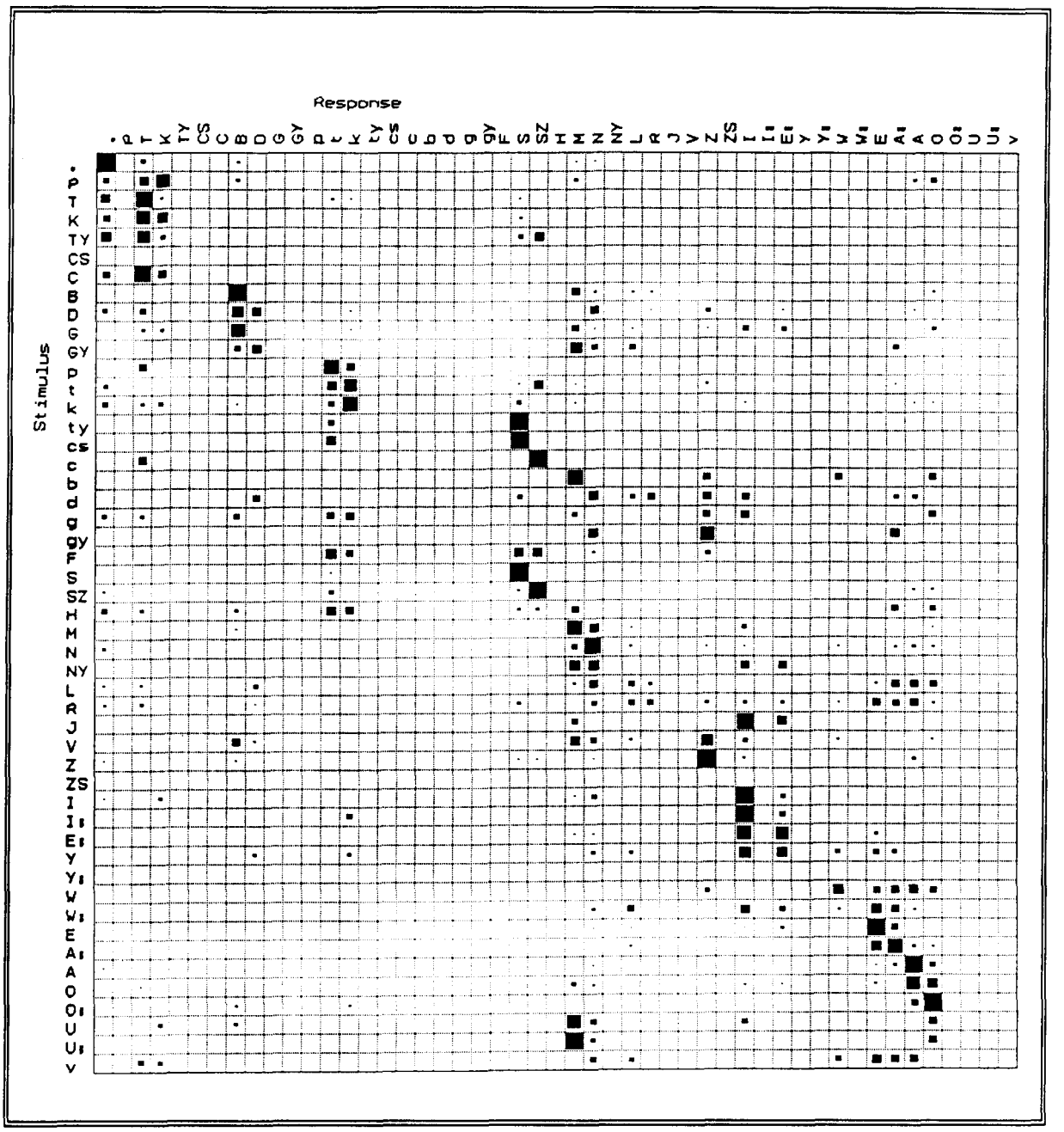

Fig. 23. Phoneme substitutions on the frame level based on the Hungarian MAMO-material. The values are normalized for each row. The area of the black squares is proportional to the number of occurrences in each position.

In the Hungarian MAMO-material in Fig. 23 we notice that the occlusion phases of stops and affricates (P, T, K, C, TY, CS) were often confused. This is an effect of the labelling methodology, in which occlusions are labelled according to the phoneme they appear in. However, one could argue that they all should be labelled by the same symbol, since they only represent a silent interval. Still many of these occlusions were labelled correctly, since the context of each frame is included in the phone net input. Other typical confusions are the substitution of burst noise with fricative noise having a similar place of articulation (c-sz, cs-s, ty-s). These confusions are biased in the fricative direction, since there are more fricative noise frames in the training material making the network more sensitive to them. The long-short confusions (e.g. I:-I) are also based on acoustic similarities. An interesting case is the U-M substitution. It shows how acoustically similar the almost closed vowel and the orally

completely closed, nasalized sonorant are. The most frequent confusions have a good chance to be corrected during a following language level processing, since they may be phonetically predicted.

4.3 Phoneme level classification using expert segmentation

This evaluation is based on the smoothed output activations of the phone network. An example of this is presented in Fig. 24. After observing all test material using this representation, the following significant conclusions could be drawn:

- The peak of the phone with the maximum activation has an average value of 0.47 inside the phoneme borders. The second and third largest peeks have a value of 0.27 and 0.20, respectively. Thus, the first candidate regularly has a considerably higher value than the others.
- The activation levels of adjacent phonemes cross each other close to the phoneme borders. This fact forms the basis for our automatic segmentation network.
- The phone candidates proposed by the net and the manual labels are the same in the most cases. The recognition rate values are summarized in Table VIII, in which each epoch includes one presentation of each pattern in the full training set.

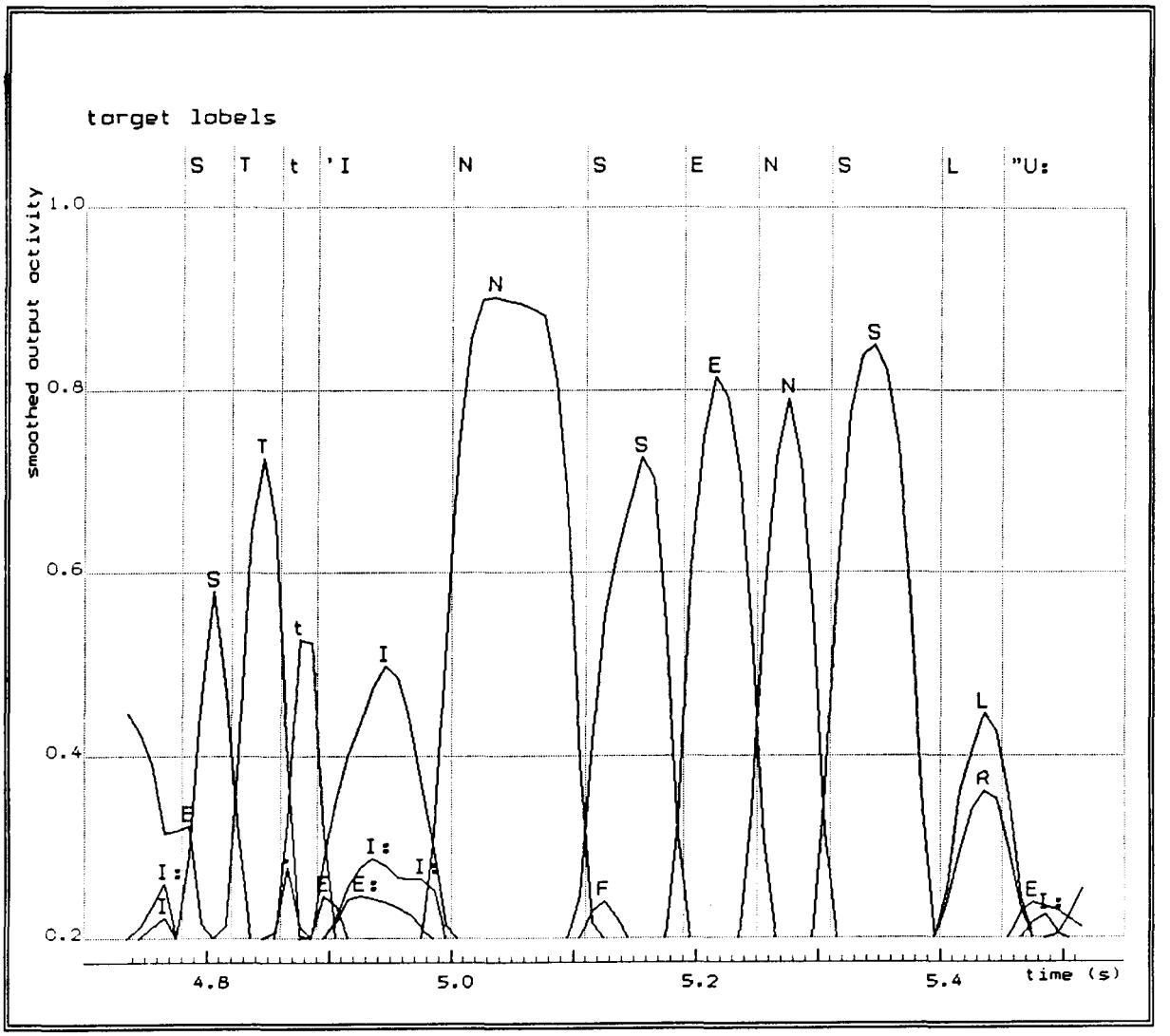

Fig. 24. An example of the smoothed phone net output activations.

The results for Hungarian show the evolution of the performance as a function of the phone net training. The performance shows some signs of levelling off at the end. The better scores for Swedish can probably be attributed to its smaller phoneme set.

Table VIII. Phoneme recognition in percent for the Swedish and Hungarian phone network when using the segmentation of the human expert. Results for correct phoneme within the top one, two or three best phone candidates proposed by the phone net. Performance increase during training shown for the MAMO-network.

Material	*INTRED*	*M*	*A*	*M*	*O*		
No. of epoches	2500	50	500	1000	1500	2000	2500
1st corr.	64.4	23.2	42.2	46.7	46.5	49.6	50.4
1st or 2nd corr.	78.1	32.9	58.4	65.3	67.3	69.7	69.9
1st, 2nd or 3rd corr.	82.2	41.1	68.1	71.5	74.3	75.6	76.3

The peak value of the smoothed phoneme activations is a good measure of how reliable a network classification is. The relation between the recognition rate and the peak of the smoothed activation is shown in Fig. 25. It shows that the probability of making a correct classification is fairly well approximated by a linear function of the smoothed activation peak. However, sometimes wrong candidates were proposed even though the neural network had a high activation value (0.7-0.8). This must, of course, be taken into consideration although one could argue that similar phenomena occur also in human communication.

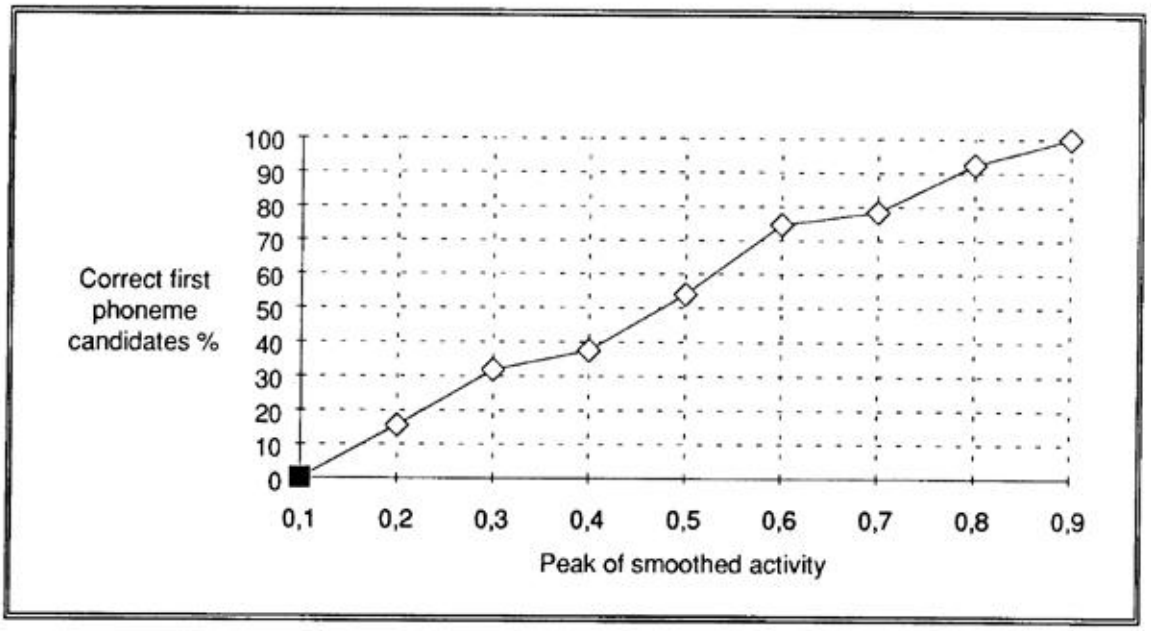

Fig. 25. The probability of correct phoneme recognition as a function of the smoothed activation peak values in the MAMO-material.

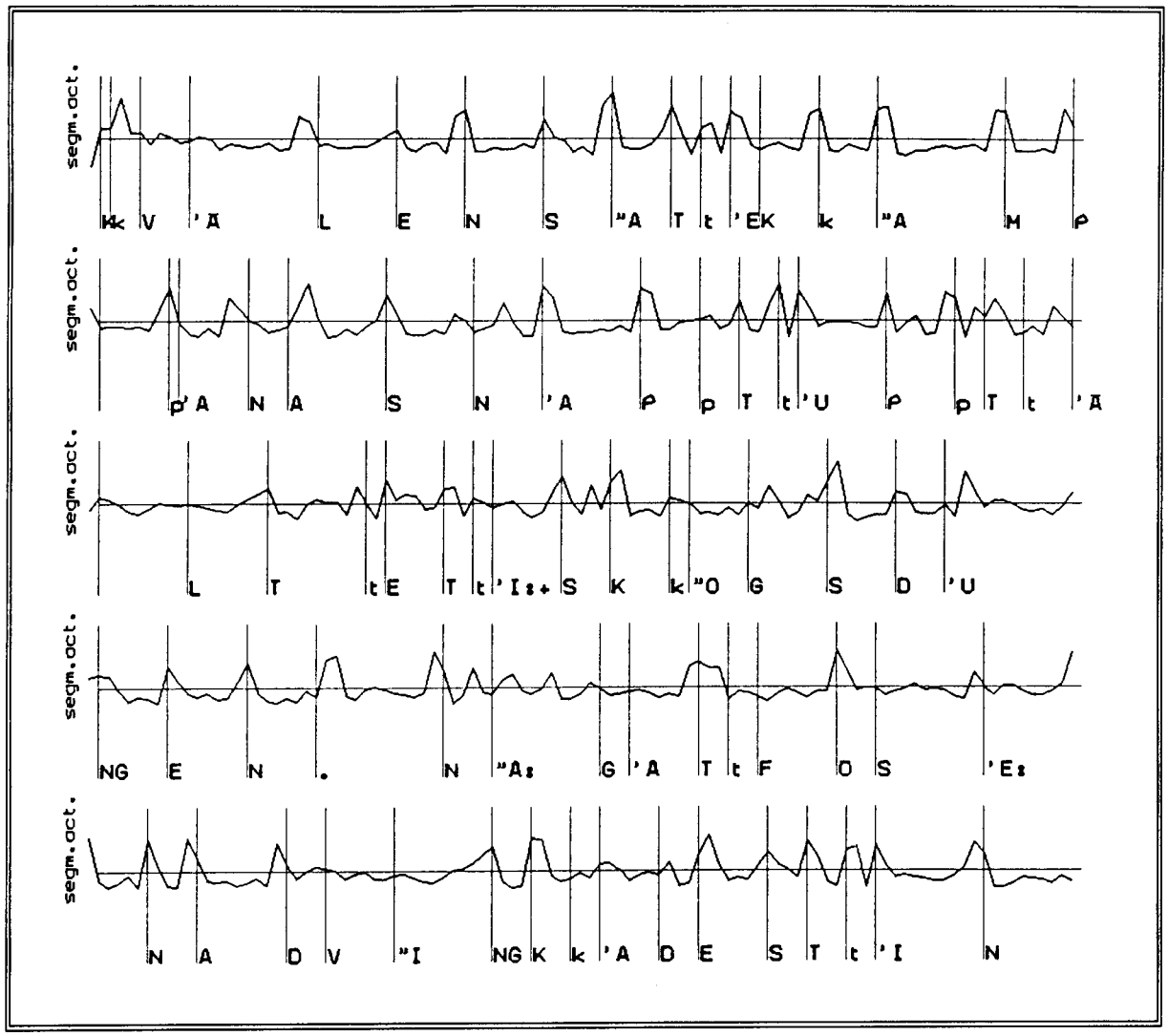

Fig. 26. *The output activation of the segmentation network as a function of time. The vertical lines mark the expert segmentation. The straight horizontal lines indicate the threshold level.*

4.4 Recognition of phoneme borders

The single output of the segmentation network is trained to have a high value in the first frame of each phoneme, see Fig. 26. The system detects a phoneme segment when the activation has a peak above the decision threshold. Increasing this threshold decreases the number of detected segments and vice versa. The optimum value can only be determined in cooperation with a language level processing. In our evaluation, a medium level threshold was chosen, where the number of errors did not vary very much when perturbing the threshold value. For the INTRED-material a threshold of 0.2 was used and the value for the MAMO-material was 0.16. The results are summarized in Table IX.

Table IX. Performance of the segmentation network.

	MAMO %	*INTRED %*
Segmentation in the frame marked by human	39.3	35.9
Segmentation within +/- 1 frame	82.0	81.6
Lost segments	18.0	18.4
Extra segments	57.4	54.3

4.5 Phoneme recognition using automatic segmentation

The phoneme recognition process is based on the peaks of the smoothed phone net output activations within the segments detected by the segmentation net. The first three candidates were compared to the target phoneme in the evaluation. To reduce some of the problems due to lost and extra phoneme borders, phonemes in automatic segments placed plus or minus one (expert) segment from the correct target phoneme segment were allowed when comparing to the target phoneme. It is conjectured that these errors should not degrade an upper level recognition performance too much. Moreover target segments were compared to phoneme candidates in multiple segments that overlapped them in time. Frequently these inserted segments were correctly labelled, which would reduce the degradation in recognition performance caused by them. Having these evaluation criteria in mind, we note that we get about the same level of performance as when using manual expert segmentation, see Table X below.

Table X. Results for phoneme recognition when using automatic segmentation.

	MAMO %	*INTRED %*
Correct first candidate	50.3	64.4
Correct first or second candidate	64.5	74.8
Correct first. second or third candidate	72.0	80.8

4.6 Experiments with the base-line reference phone network

The base-line phone network was trained by the same training data and the same number of epoches as the standard phone net for the INTRED-material. The number of connections is considerably less compared to the standard phone network. Since we have less free parameters, this should mean that the training level is higher compared to the standard network. The risk for over-training is small considering the large variation in the training data and we did

not observe any effect like that. In this test, only the spectral data of each speech frame was used as input to the phone net. The results are summarized in Table XI. The results on the frame level are included for comparison. They are based on a frame-by-frame comparison of the phone net outputs to the target labels and include no segmentation at all. The table shows that the recognition scores are considerably lower for the base-line net when compared to the standard net, indicating that the coarse feature part of the dual window plays an important role in the recognition process.

Table XI. Phoneme recognition on frame and segment level using the base-line phone network and the original phone network using the INTRED-material.

		System performance using base-line network (%)	*System performance using original phone network (%)*
Frame level recognition	correct first candidate	40.8	54.2
Expert segment borders	correct first candidate	41.3	64.4
	correct first or second candidate	54.9	78.1
	correct first. second or third candidate	60.9	82.2
Automatic segment borders	correct first candidate	39.8	62.4
	correct first or second candidate	52.4	74.6
	correct first. second or third candidate	59.4	80.8

4.7 Speaker and language dependency of the system

Speaker and language independence of the coarse feature set was a main objective when planning the system structure. We only have had the possibility to test these characteristics for a limited number of speech materials. The results are summarized in Fig. 27. For the INTRED - JONSSON cross-speaker test the language was the same but both the speaker and the text were changed. In the INTRED - MAMO and MAMO - INTRED cross-language tests both the language and the speaker were changed. According to the figure, the recognition rate of the coarse features has changed only a little compared to the standard test results, which shows that these features indeed are quite robust.

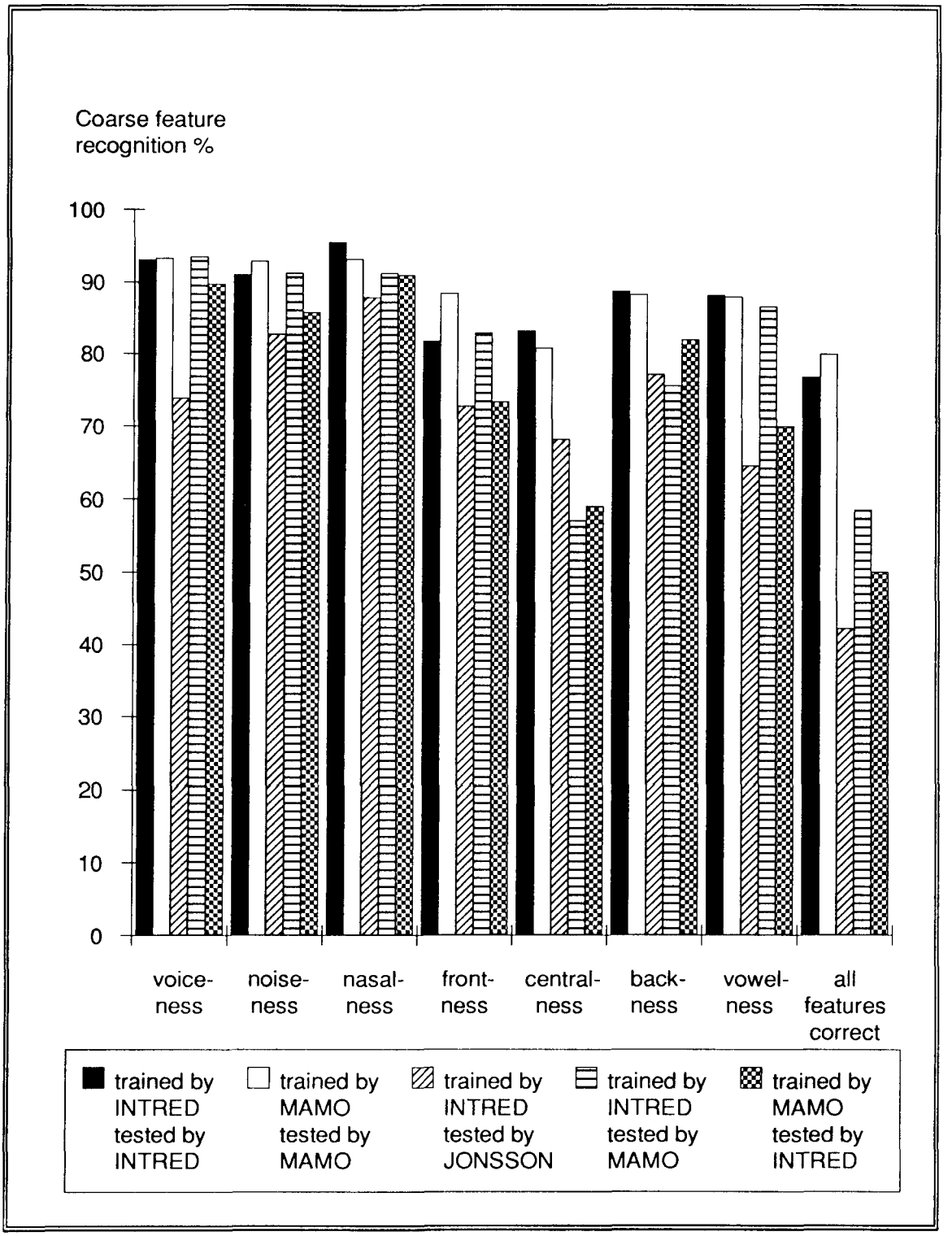

Fig. 27. Cross-speaker and cross-language recognition results of the coarse features.

The phoneme level results of the cross-tests are shown in Table XII. It is evident that the complete system is highly adapted to the training speaker, which substantially reduces the performance for other speakers. This is not very surprising considering how much the acoustic speech signal varies between speakers.

The Swedish and the Hungarian phoneme sets are considerably different why cross-language tests at the phoneme level would have no meaning.

Table XII. Recognition of phonemes using another speaker on frame and segment level. The networks were trained by the INTRED-material and tested by the JONSSON-material.

Frame level recognition	correct first candidate	17.2 %
Expert	correct first candidate	18.3 %
segment	correct first or second candidate	25.1%
borders	correct first. second or third candidate	31.5%
Automatic	correct first candidate	21.0 %
segment	correct first or second candidate	29.8 %
borders	correct first. second or third candidate	35.6 %

5. SOME CHARACTERISTIC PHONEME RECOGNITION ERRORS

An analysis of the phoneme recognition errors has shown that in most cases there are several reasons for the confusions. The classification only allows phoneme candidates to be either correct or false, but a more realistic modelling would introduce probabilities associated with the labels. The recognition errors have been classified into some different categories below. The categories are not independent of each other.

5.1 Recognition system specific errors

Low-frequency phonemes are represented by fewer samples in the training material and this will in turn lead to a less effective training of their representations in the phone network. Substitutions and deletions caused by this effect are referred to as training-specific errors. Updating the weights after each pattern was supposed to counter-effect this somewhat, but the most frequent phonemes still have a large influence on the weights. The sensitivity of the network to the less frequent phonemes does not increase until the more frequent ones are recognized with small output errors. One third of the phoneme recognition errors in the MAMO-material is in this class of training-specific errors. In evaluations based on the human segmentation, these errors appear as substitutions, but when using automatic segmentation, they are counted as deleted phonemes. The sensitivity of the phone network will be low and the segmentation network will not be able to find the segment borders for low frequency phonemes.

Another typical system-specific error category is the deletion of the burst phases in stops. The coarse feature network generally has strong peaks in these cases (compare the noiseness feature in Fig. 13). However, at the output of the phone network, the peaks are reduced, since the bursts are short (usually only one single frame), and the smoothing filter will push them below the activity of the surrounding phonemes.

5.2 Errors related to the speech material

Reduced articulation is an integral part of human speech. The last part of a word is often not clearly articulated. Frequently phonemes are not completely deleted but there are still some indications of their identity by various acoustic phenomena. The acoustic manifestation occurs only as modulations of neighbouring sounds. This is an apparent problem also when doing manual labelling of speech.

Another interesting phenomenon is the overlapping articulation of two sounds, where the phonetic elements are hard to separate. The smoothed output activations of the phone network are, however, capable of displaying these conflicting evidence, as may be seen in Figs. 28 and 29. In the MAMO-material (with text "..hússal rakta.."), an overlapping articulation of the L- and R-phonemes occurred. The phonetician did not insert any separate label for the L-sound, but the phone network gave a very specific response: the L- and R-outputs have simultaneous but weak peaks. The stronger activations of their neighbours mask the simultaneous L-R peeks, but it is intriguing to see this embryo of recognition.

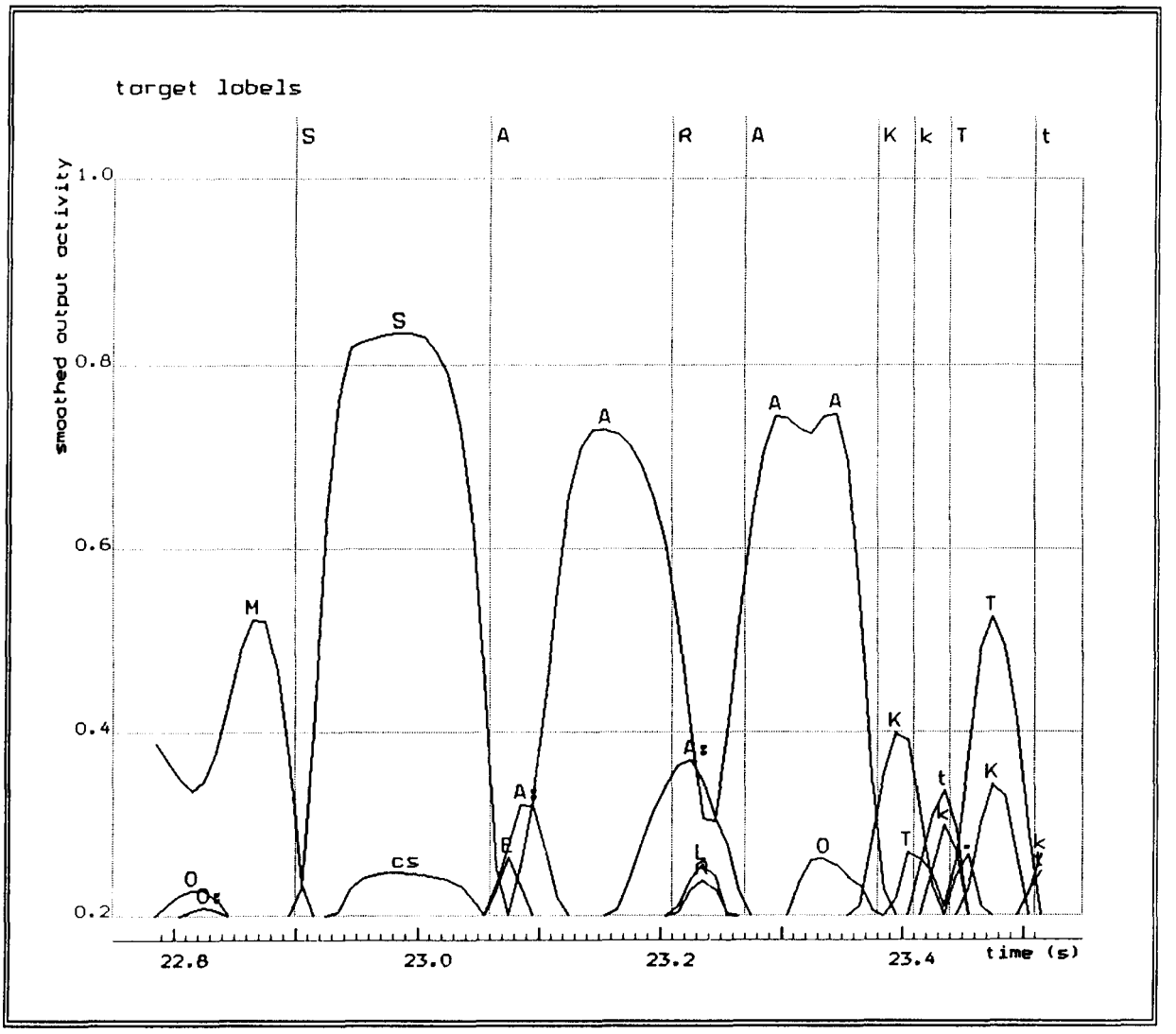

Fig. 28. Example of smoothed activations in the case of reduced and overlapping articulation, MAMO-material.

In the MAMO-material, 70% of the affricates are classified as stop-fricative-combinations having the same place of articulation as the target affricate. In the evaluation they were

counted as substitution errors. The question whether to perceive them as one compound sound or a combination of two sounds, is still discussed among Hungarian phoneticians.

The results discussed in the previous section and the error examples in this section show that the whole system has a good phone labelling capability. It has also demonstrated a capability to represent ambiguous and overlapping events by parallel output activations.

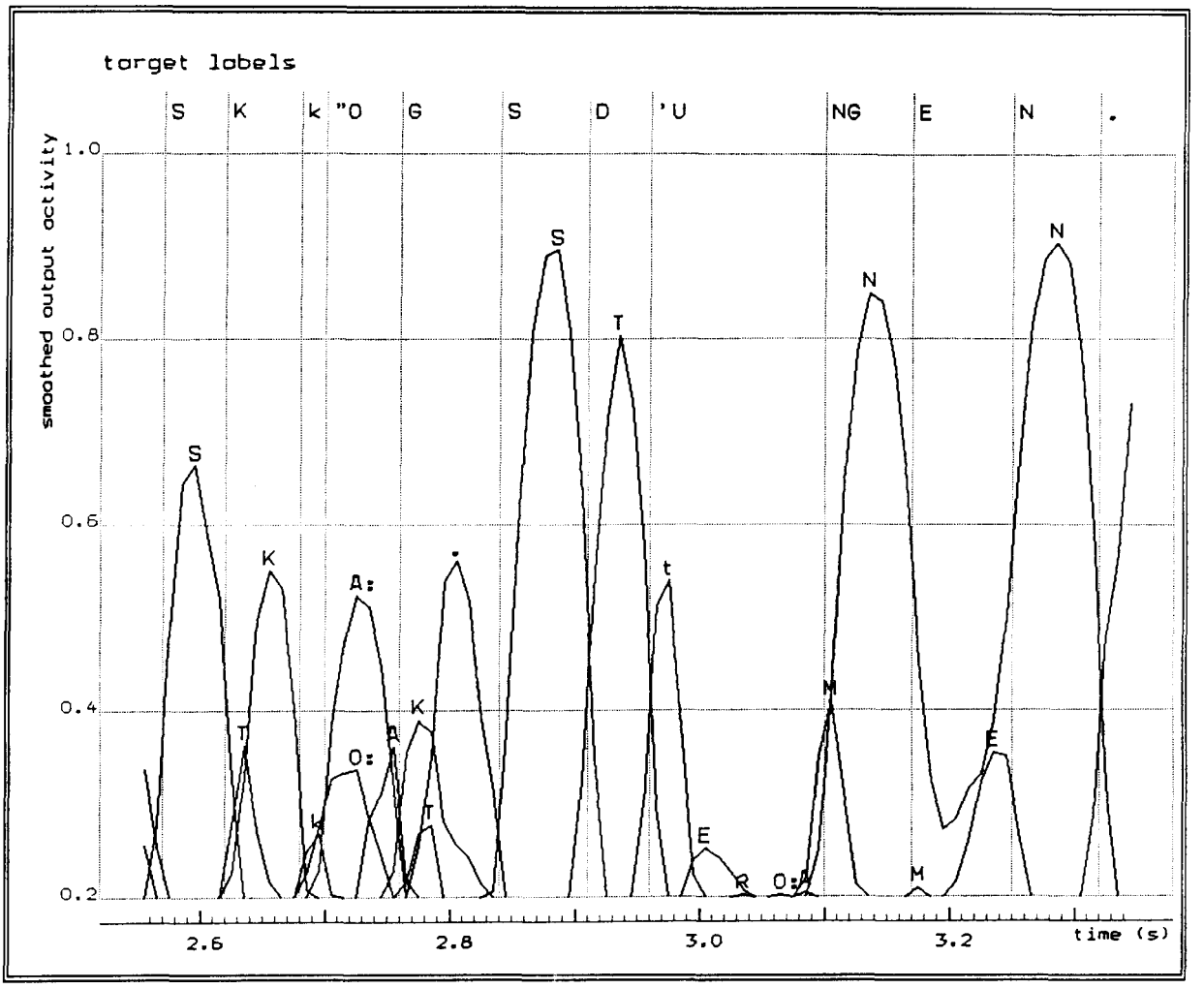

Fig. 29. Example of smoothed phone activations in the cases of unvoiced "G" and "D" and in the case of a reduced "E" (Swedish).

A short speech sample from the INTRED-material in Fig. 29 shows several errors related to the speech material. In the text "skogsdungen", the G-K and D-T substitutions are regarded as errors by the evaluation program. A traditional spectrogram representation of the speech signal shows that these voiced stops are realized by unvoiced segments, which is natural considering coarticulation effects. However, the phonemic labelling scheme used for this material is the real origin of these errors and the network interpretation is quite natural in this case. These problems may be resolved by introducing higher level linguistic components or having enough training samples for a context sensitive net to learn that voiced stops become voiceless in certain surroundings. Returning to reduced articulations, the "E"-phoneme in "skogsdungen" is a typical example of this. Its smoothed activity has a peak value of 0.36 only, but still the phoneme is clearly there. However, it is only the second candidate since the activations of strong surrounding nasals are masking it (the E-vowel is naturally heavily nasalized).

6. CONCLUSIONS

A phone-labelling system for continuous speech has been constructed and evaluated. Besides using an earlier established Swedish database a Hungarian speech data-base has been assembled for training and testing the system.

The coarse features have been shown to be quite robust, when changing speaker and language. It would also be interesting to test a feature set where the place of articulation features were replaced by some features related more to the spectral characteristics of sounds (i.e., compact, diffuse and flat, compare Fant, 1973). The results in Fig. 16 where some Swedish front vowels were regarded as central by the net is an indication of this.

The output activations of the network have an inherent ability to indicate parallel and overlapping speech events. They also have been shown to develop continuous values in spite of binary targets to discriminate between phones with similar target features, and in at least one case these output values have a meaningful phonetic interpretation, as can be seen in the vowel plots of Figs. 18 and 19.

The lexical search in the language level processing should be based on the most probable phoneme candidates. Since the activation values of the current system are closely related to the probability of correct labelling (compare Fig. 25), our system is able to provide this additional information. The further lexical processing resembles steps undertaken when trying to solve crossword puzzles. In different studies it has been shown that even rather a crude phonetic description of a word will drastically reduce the number of possible word candidates also for vocabularies having thousands of words (Carlson, Elenius, Granström and Hunnicutt, 1986; Huttenlocker & Zue, 1983).

Many ideas regarding how to increase the system performance appeared during the evaluation of the system. The system is quite complex and has thousands of free parameters. Due to limited time only a few parameter variations were tested. Below, some ideas are listed concerning parameters that either could improve the recognition rate or speed up the training process:

- the number of nodes in the hidden layers of the networks (coarse, phone, segmentation)
- the training parameters of networks (momentum term, learning rate)
- the selection of the training data (perhaps by representing rare phonemes more frequently than in natural speech and by omitting ambiguous elements)
- optimizing the smoothing filter parameters
- optimizing the feature window length
- optimizing the segmentation window length
- managing stops as single events in the phone classification.

The size of the speech material used is a limitation, that we have some difficulties in estimating the effect of. Compared to the practically unlimited variations possible in a language, the representation of speech by 50 sentences, or 2800 phonemes, or 21000 frames, is very fragmentary, but compared to similar systems reported in other papers the size of the speech material is similar to many of them.

There have been some speculations about the necessary acoustic-phonetic recognition accuracy in continuous speech recognition. It is a well known fact in telecommunication (Flatcher, 1953) that an 80% logatom intelligibility in a transmission system means practically error-free communication (99 % sentence intelligibility). According to Klatt (in Lea, 1980) some researchers argue that 60-70% accuracy is what we could expect from machines, while Klatt rather thinks that 90% is the needed performance target.

It is difficult to compare this system to other recently published systems. There are only a few results reported for complete phoneme sets, and the working principles, the speech mate-

rials used, and the evaluation methods are different. The speaker independent recognizer of AT&T has been reported to have a 52% phoneme recognition rate (Levinson, Liberman, Ljolje & Miller, 1989). Systems based on the Kohonen feature map report a 75-90 % recognition rate depending on the speaker (Kohonen, Torkkola, Shozakai, Kangas & Väntää, 1987). Many systems reporting recognition rates above 90 % process a subset of phonemes only or use presegmented phoneme samples. Considering that the above systems probably have been more elaborately tuned, we consider our efforts quite promising.

Future work includes evaluation of different input parameters, varying the sizes of input windows, speeding up the training and introduction of a lexical component. It would also be interesting to test recurrent nodes and to do comparisons with self-organizing nets like those proposed by Kohonen, using the same speech material.

Acknowledgements

This project was supported by the Swedish Institute and by the Swedish Board for Technical Development.

References

Blomberg, M. (1989): "Synthetic phoneme prototypes and source adaptation in a speech recognition system," *STL-QPSR* No. 1, pp.131-135.

Chomsky, N. & Halle, M. (1968): *The Sound Pattern of English*, Harper & Row, Publ., New York.

Elenius, K. & Blomberg, M. (1982): "Effects of emphasizing transitional or stationary parts of speech signal in a discrete utterance recognition system," pp. 535-538 in *Proc. ICASSP-Paris.*

Elert, C-C. (1989): *Allmän och svensk fonetik*, Norstedts Förlag AB, Stockholm.

Fant, G. (1973): *Speech Sounds and Features*, The MIT Press, Cambridge, MA.

Fletcher, H. (1953): *Speech and Hearing in Communication*, D. Van Nostrand Company, Princeton, NJ.

Hunnicutt, S. (1987): "Acoustic correlates of redundancy and intelligibility," *STL-QPSR* No. 2-3, pp. 7-14.

Huttenlocker, D. & Zue, V.W. (1983): "Phonotactical and lexical constraints in speech recognition," *MIT Speech Communication Group, Working Papers, Vol. III.*

Jacobson, R., Fant, G. & Halle, M. (1963): *Preliminaries to Speech Analysis. The Distinctive Features and Their Correlates*, The MIT Press, Cambridge, MA.

Kohonen, T. (1984): *Self-Organization and Associative Memory*, Springer-Verlag, Berlin.

Kohonen, T. (1988): "The 'NEURAL' phonetic typewriter," *IEEE Computer* **3**, pp. 11-22.

Kohonen, T., Torkkola, K., Shozakai, M., Kangas, J., & Väntää, O. (1987): "Microprocessor implementations of a large vocabulary speech recognizer and phonetic typewriter for Finnish and Japanese," pp. 377-380 in *European Conf. on Speech Technology, Edinburgh.*

Komori, Y., Hatazaki, K., Tanaka, T., Kawabata, T. & Shikano, K. (1989): "Phoneme recognition expert system using spectrogram reading knowledge and neural networks," pp. 549-552 in (J. Tubach & J.J. Mariani, eds.) *Proc. Eurospeech 89, Paris, Vol. II*, CPC Consultants Ltd., Edinburgh.

Krause, A. & Hachbarth, H. (1989): "Scaly artificial neural nets for speaker-independent recognition of isolated words," pp. 21-24 in *Proc. ICASSP-Glasgow.*

Lea, W.E. (ed.) (1980): *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, p. 266.

Levinson, S.E., Liberman, M.Y., Ljolje, A., & Miller, L.G. (1989): "Speaker independent phonetic transcription of fluent speech for large vocabulary," pp. 21-24 in *Proc. ICASSP-Glasgow.*

Lippmann, R.P. (1987): "An introduction to computing with neural nets," *IEEE ASSP Magazine* **4**:2, pp.4-22.

Lippmann, R.P. (1988): "Neural nets for computing," pp. 1-6 in *Proc. ICASSP-New York.*

Mariani, J. (1989): "Recent advances in speech processing," pp. 429-440 in *Proc. ICASSP-Glasgow.*

McClelland, J.L. & Rumelhart, D.E. (1988): *Explorations in Parallel Distributed Processing* The MIT Press, Cambridge, MA.

McDermott, E. & Katagiri, S. (1989): "Shift-invariant multi-category phoneme recognition using Kohonen's LVQ2," pp. 81-84 in *Proc. ICASSP-Glasgow.*

Niles, L., Silverman, H., Tajchman, G., & Bush, M. (1989): "How limited training data can allow a neural network to outperform an 'optimal' statistical classifier," pp. 17-20 in *Proc. ICASSP-Glasgow.*

Nord, L. (1988): "Acoustic-phonetic studies in a Swedish speech data bank," pp. 1147-1152 in *Proc. SPEECH'88, Book 3* (7th FASE Symposium), Institute of Acoustics, Edinburgh.

Rumelhart, D.E. & McClelland, J.E. (1986): *Parallel Distributed Processing, Vol. 1-2*, The MIT Press Cambridge, MA.

Singh, S. (1976): *Distinctive Features Theory and Validation*, University Park Press, Baltimore.

Treleaven, P (1989): "Neuroncomputers," *Int.J. Neurocomput.* **1**:1, pp. 4-31.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. (1989): Phoneme recognition using time-delay neural networks," *IEEE ASSP* **37**:3, pp. 626-631.