Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 15 (1991), 61-77.

ON RELATIONS BETWEEN PHONE MODELS, SEGMENT DURATION, AND THE PADÉ-EXPANSION

L.F.M. ten Bosch

Abstract

In this paper, we will consider a relation between three different numerical aspects that play a role in speech research models, viz. (1) the phone model as applied in HMM speech recognition algorithms, (2) the statistical distribution of the duration of the speech segments to be modelled, and (3) a mathematical tool, known as the Padé-approximation. To each phone model a transfer function will be assigned. By means of this assignment, a correspondence between the *topological* structure of phone models on the one hand and the *algebraic* properties of the corresponding transfer function on the other hand can be established. By this correspondence, an admissable class of network topologies of phoneme-like units can be calculated effectively on the basis of segment duration data.

1. Introduction

As is well known, so-called Hidden Markov Models (HMM) are frequently used in Automatic Speech Recognition. In the HMM-approach, the speech signal is considered to be the acoustic outcome of a (finite, first order, stationary) Markov process (Holmes, 1988; Lee, 1989) A concatenation of acoustic events (for example in terms of spectral characteristics of frame-like speech segments) is modelled by a sequence of non-observable ('hidden') states in a Markov-chain. Between these hidden states, transition probabilities are given by means of a *transition matrix A*. An *emission matrix B* specifies the probability density of acoustic vectors for each transition between hidden states. These two matrices *A* and *B* completely specify the Markov model[1].
The transitions that are in principle possible, i.e. those which may have a probability larger than zero, are *admissable*. The set of admissable transitions determines the *topology* of the Markov chain (= Markov 'network').
A *speech segment* can be modelled by many different Markov chains, sometimes called phone-like units (PLU's). In figure 1, two PLU's with different topology are presented. Only the admissable transitions are depicted. On the left-hand side, a simple model is shown in which a segment is represented by three hidden states. On its left-most and right-most side, the *initial* and *final* state are represented by the symbols I and F, respectively. Transitions from and towards these default states are non-emitting, i.e. they do not correspond to an acoustic output.

[1]The recognition rates in case of connected word recognition also depend on the *grammar*. In the strict sense, this grammar does not belong to the underlying Markov model.

On the right-hand side of figure 1, a more elaborate network consisting of seven states is shown. As this network contains twelve transitions, this latter network is capable of representing a much more difficult sequence of acoustic events.

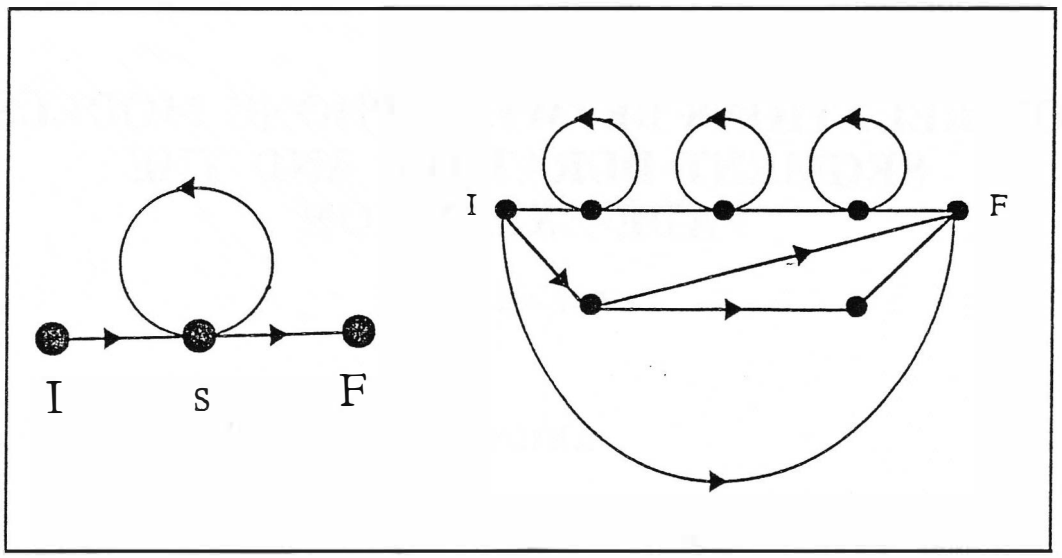

Fig. 1. (Left) A three-state network with three transitions (of which one represents a self-loop). (Right) A seven-state network with twelve transitions.

A *path* through the network represents the evaluation over time of a (through vector quantizing discretized) speech parameter vector, each transition corresponding to one fixed time step. In current HMM-algorithms, networks are 'feed-forward', which means that time-reversal is not allowed[2]. The duration of the modelled speech segment is defined by the number of transitions (i.e. time steps) required to reach the final state F from the initial state I, which is, in turn, determined by the matrix A of state transition probabilities. In other words, A fully determines the duration distribution[3] of a segment.

In this paper, we will go into detail on the relation between the matrix A and the duration distribution and, more specifically, the inverse problem of how to derive A (together with the network topology) from the observed distribution of the segment duration. It will be shown that every duration distribution determines a class of networks that all are equivalent with respect to their power to model the statistical behaviour over time of that particular segment. These networks, however, may differ with respect to their topology and the transition probabilities, and may consequently behave differently with respect to the capability of *spectral* modelling. With respect particularly to modelling duration, our method is capable to point to a specific exemplar out of this class which is the 'most simple PLU' in some sense.
In the next section, we informally introduce the preliminaries required for the present problem, and we consider many examples how the network topology determines the

[2]In the present exposition, hoewever, there is no *theoretical* objection to time-reversal HMM-networks. Time-reversality will not be studied.
In practical HMM systems, *place*-reversality is not allowed, which means that the time evaluation along the states a-b-a is impossible within one phone model. In this paper, we generally meet this restriction; in a few cases we will study specific examples of place-reversality.

[3]That is, the probability of the time interval having a specific duration. We use the term 'distribution' rather than the term 'probability density function' (pdf) in order to avoid confusion with respect to pdf's defined on codebooks.

duration probability function. In section 3, the formal relations between network topology, duration distribution and the Padé-approximation will be dealt with.

2. Preliminaries

A central notion will be the *generating function*. Let f(k) denote an arbitrary (real) function defined on the set of integers N = {0, 1, 2, 3, ...}. Then the function Gf(X) defined by

$$Gf(X) = \sum_{k \geq 0} f(k)\, X^k$$

is called the *generating function* of the function f. In this formula, X denotes a formal variable. If f(k) denotes a probability density function on N, then it follows that Gf(X) is non-negative and monotonically increasing on the interval [0,1], and Gf(1) = 1. The expectation of a stochast with f(k) as a pdf is formally given by (Gf)'(1).
Generating functions are handy tools when dealing with the durational behaviour of networks.

Example. The self-loop (figure 1) contains three states I, s and F. There are three admissable transitions (I, s), (s, F), and the self-loop (s, s). We consider a process, starting at t=0 in the initial state I, which terminates as soon as F is reached. With every transition, a probability and a print action are defined, according to the following table:

transition	probability	action
(I, s)	1	print ''
(s, s)	a	print 'T'
(s, F)	1-a	print ''

in which '' denotes the empty string. A full specification of all transition probabilities between the states (I, s, F) is specified by the matrix A:

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & a & 0 \\ 0 & 1\text{-}a & 0 \end{pmatrix}$$

In this matrix, the column and row indicate the source and target state, respectively[4]. After the process has terminated in the final state F, the resulting length L of the output string of T's equals the number of times the system passed the self-loop (s, s). Let P(L=k) denote the probability of this length L equalling exactly k. We have

$$P(L=k) = a^k (1 - a)$$

By substitution, its generating function GP(X) reads

[4]In the literature, the transposed notation is sometimes applied. The advantage of the present notation is that the ordinary matrix products retains an interpretation in terms of probabilities.

$$GP(X) = \sum_{k\geq 0} P(L=k)\, X^k = \sum_{k\geq 0} a^k\,(1-a)\,X^k = (1-a)\sum_{k\geq 0}(aX)^k = \frac{1-a}{1-aX}$$

In the last step, we essentially applied the power series expansion of the function 1/(1-X).
This function GP(X) is uniquely determined by the network and by the probability density function P(L=k). GP(X) is an example of a so-called *rational function*, i.e. a quotient of two polynomials. We will encounter rational functions throughout the present theoretical exposition.
Any network can be assigned a generating function, by first evaluating P(L=k) and next GP(X). However, in general this is not the most handy way to evaluate GP(X). A more elegant method is to directly evaluate GP(X) from the topology of the network, without explicit reference to the probability density function P(L=k). The by-pass is depicted in figure 2.
The next example shows the direct formal evaluation of GP(X) in case of the network N_2 (figure 3).

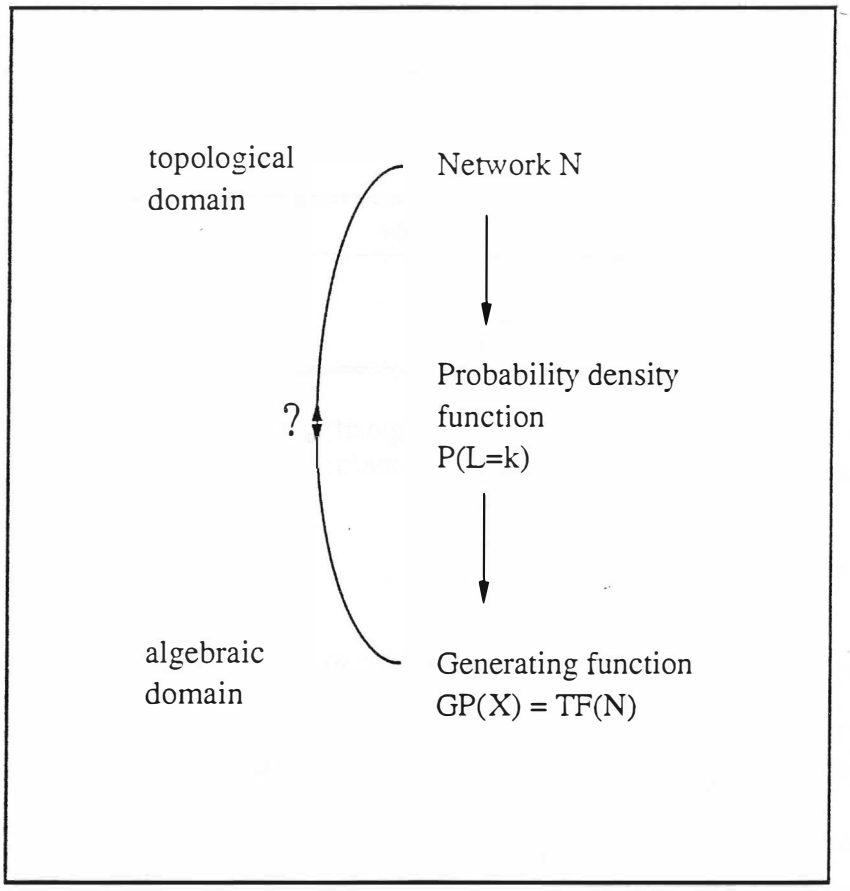

Fig. 2. A conceptual overview of the model. The by-pass is represented by the question mark.

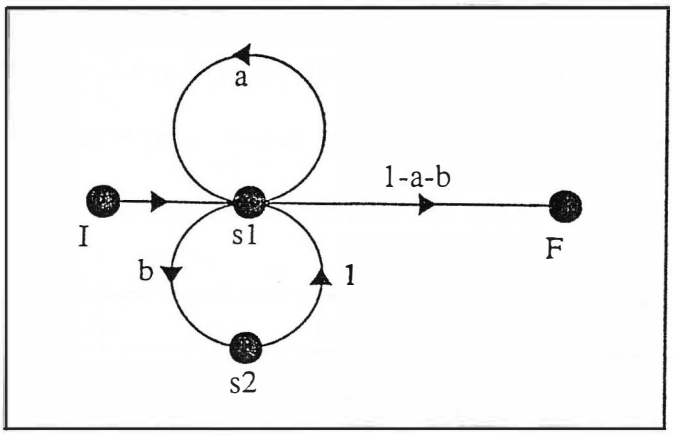

Fig. 3. The network N_2. For an explanation see the text.

Example. In order to formally assign a generating function GP(X) to the network N_2, a rational function $R(s_0, X)$ will be assigned to each state s_0 of the network N_2 by solving a specific system of linear equations. The algebraic structure of the system of linear equations is based on the topological structure of the network. Each of the linear equations relates $R(s_0, X)$ to a weighted 'incoming' sum of R(s, X) over all states s that are adjacent to s_0:

$$R(s_0, X) = \sum_{\{\text{all s adj. to } s_0\}} (\text{weighting polynomial})\ R(s, X)$$

After having solved the set of equations, GP(X) is defined as the quotient

$$\frac{R(F, X)}{R(I, X)}$$

The weighting polynomials determine whether or not the corresponding state transition is taken into account in the duration[5]. Their construction is clarified by the following table:

edge	probability	variable	weighting polynomial
(I, s_1)	1	$X^0 = 1$	1
(s_1, s_1)	a	$X^1 = X$	aX
(s_1, s_2)	b	$X^1 = X$	bX
(s_2, s_1)	1	$X^1 = X$	X
(s_1, F)	1-a-b	$X^0 = 1$	1-a-b

The first column contains all admissable state transitions. In the second column, the actual state transition probability is given. (These numbers are shown in figure 4). The third column contains new information, viz. the contribution of each transition for the

[5] By using variables $X_1, ..., X_k$, it is possible to assign a k-ary polynomial to an arbitrary network. We will not go into details.

overall duration. In this column, the exponent of the variable X equals the contribution of the corresponding state transition to the number of time steps required to pass from I to F. As the transitions (s_1, s_1), (s_1, s_2), and (s_2, s_1) have X^1, these transitions take one time step.
The fourth column contains the product of the entries in the second and third column.

According to the topology of the network, the following set of four linear equations is constructed:

State s	Weighted 'incoming' sum	Result R(s, X)
I	1	= R(I, X)
s_1	1.R(I, X) + aX.R(s_1, X)+ X.R(s_2, X)	= R(s_1, X)
s_2	bX.R(s_1, X)	= R(s_2, X)
F	(1-a-b).R(s_1, X)	= R(F, X)

After direct substitution and some combination, it can easily be shown that N_2 has the following generating function:

$$GP(X) = \frac{R(F, X)}{R(I, X)} = \frac{R(F, X)}{1} = \frac{1\text{-}a\text{-}b}{1\text{-}aX\text{-}bX^2}$$

The same result can be obtained by directly considering the matrix equation Mx + e = x, or, equivalently, (M-I)x = -e, in which

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & aX & X & 0 \\ 0 & bX & 0 & 0 \\ 0 & 1\text{-}a\text{-}b & 0 & 0 \end{pmatrix} \text{ and } I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

with

$$e = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ and } x = \begin{pmatrix} R(I, X) \\ R(s_1, X) \\ R(s_2, X) \\ R(F, X) \end{pmatrix}$$

Straightforward calculation shows that det(M-I) = $1\text{-}aX\text{-}bX^2$, which equals the denominator in GP(X) shown above. Observe that this determinant never vanishes (i.e. becomes zero), and accordingly, that the solution x is always unique.

In the next section, we will see that the generating function GP(X) might be allotted the character of a 'transfer function' related to the network. To emphasize this relation to the underlying network N, GP(X) will also be denoted TF(N), the 'transfer function' of N.

The assignment of the rational function TF(N) = GP(X) to the network N is unique. It fulfils three elegant properties which relate the *topological* properties of the network to the *algebraic* properties of the generating function. We will see, however, that the correspondence is not one-to-one, but it might be helpful to formulate a statement about one domain on the basis of arguments from the other domain.
The three properties will be referred to as 'correspondence relations'.

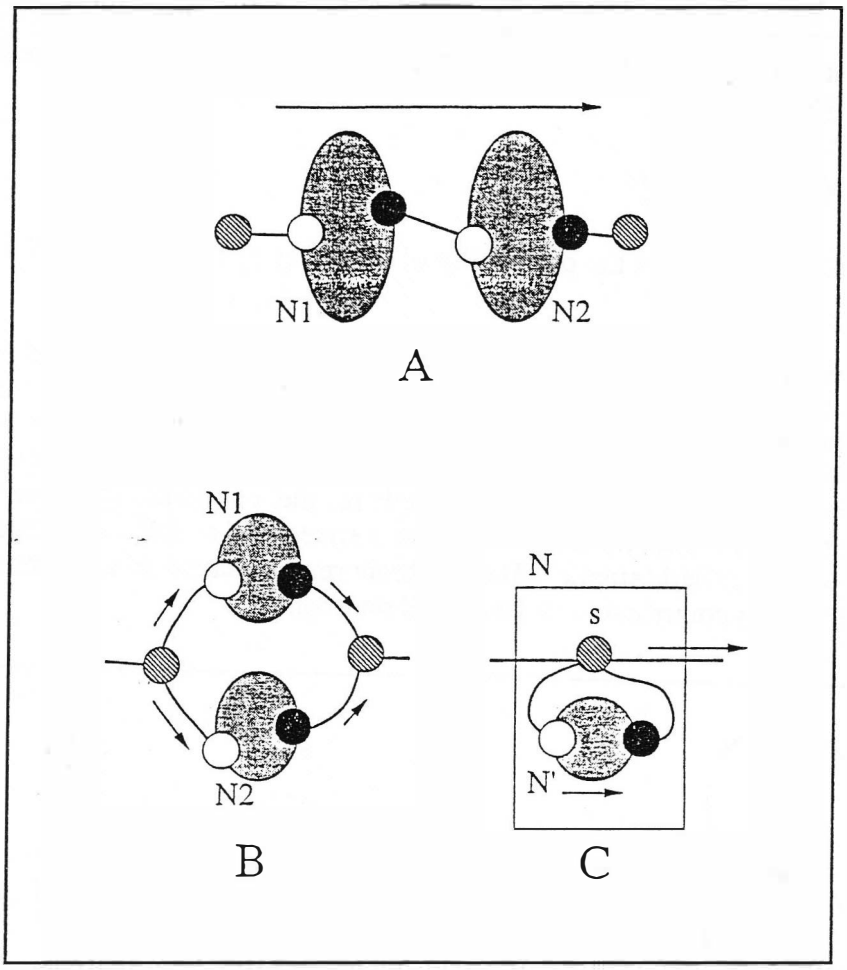

Fig. 4. Three different topological compositions of networks. The grey areas represent networks that are 'building blocks' to be integrated. Their initial and final states are represented by white and black circles, respectively. States of the integrated network are grey.
(A) Graphical representation of two networks in series.
(B) Graphical representation of a weighted parallel connection.
(C) Graphical representation of network nesting (N' into N).
For a more detailed explanation see the text.

Correspondence relation 1 (networks in series):
Let two networks N_1 and N_2 be given in a 'series' as shown in figure 5(A). A new network N_2N_1 is defined by identifying the final state of network N_1 with the initial state of network N_2. Then

$$TF(N_1N_2) = TF(N_1)\ TF(N_2)$$

Correspondence relation 2 (networks in parallel):
Let two networks N_1 and N_2 be given 'in parallel' as shown in figure 5(B). The networks now have one common initial state and one common final state. From the initial state, the transition towards network N_1 has probability a_1; the transition towards network N_2 has probability a_2, where $a_1 + a_2 = 1$. If we denote the resulting network formally by $a_1 N_1 + a_2 N_2$, then

$$TF(a_1 N_1 + a_2 N_2) = a_1\ TF(N_1) + a_2\ TF(N_2)$$

Correspondence relation 3 (nesting):
Let TF(N) denote the transfer function of a network N as illustrated in figure 5(C). If TF(N') denotes the transfer function of the nested subnetwork N', the overall transfer function TF(N) reads

$$TF(N) = \frac{1-a}{1-aTF(N')}$$

where a ($0 \leq a < 1$) denotes the probability of 'entering' the subnetwork N' from state s.

These relations are not difficult to prove, and the verification will be omitted here. They allow us to calculate the generating function of an arbitrary network by cutting and chopping the network into smaller pieces that yield more elementary generating functions. Finally, we arrive at the simplest networks of which the generating function can be easily obtained: the self-loops, and simple parallel networks. The ultimate simple networks consist of exactly one transition between two different states. These 'networks' will be called 'atomic'. The algebraic representation of an atomic network equals X^n if n is its contribution to the overall duration.

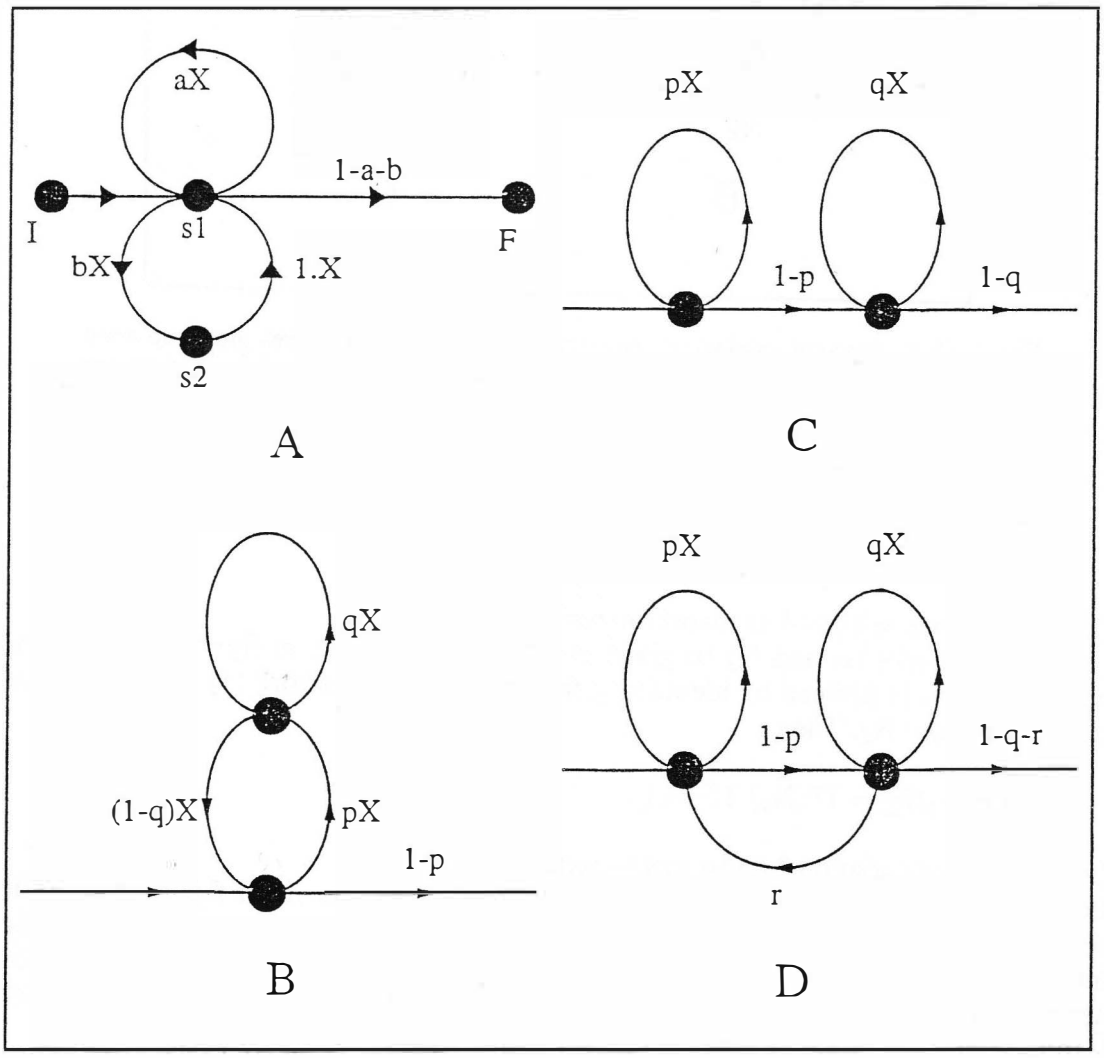

Fig. 5. Four different networks. For a discussion see the text.

has a discriminant $D = (p+q)^2 - 4pq(1-r+pr) = (p-q)^2 + 4pqr(1-p) \geq 0$. Moreover, since $(p+q)/(2pq) > 0$ and $(1-(1-p)r) > 0$, H(X) will generally have positive zeroes z_1 and z_2. Since $r < 1-q$, $H(1) > 0$, and both these zeroes will exceed 1. Consequently, their inverses w_1 and w_2 fulfil $0 < w_1, w_2 < 1$, so $H(X) = (1-w_1X)(1-w_2X)$ (up to a constant) which yields the identity

$$TF(N) = \frac{(1-p)(1-q-r)}{(1-pX)(1-qX) - (1-p)r} = \frac{1-w_1}{1-w_1X}\frac{1-w_2}{1-w_2X}$$

which corresponds to a series of two self-loops. We conclude that there exist networks with a slightly different place behaviour (i.e. they allow slightly different state sequences) with the same duration pdf. In the place-reversal network, the HMM may swap between two different spectral states, in the other network, HMM first settles down in the first state and *next* in the other. We observe that spectral considerations are here of primary importance.
We finally have a closer look at the elaborate PLU, which was already presented in figure 1. This network has twelve transitions of which eight are independent. (Independency is lost in every state -- the outgoing transitions are subject to one normalization constraint). It can be verified that the corresponding rational function is of the form:

$$a_1X + a_2X(b_1X+b_2X^2) + a_3X^4 \frac{1-c_1}{1-c_1X}\frac{1-c_2}{1-c_2X}\frac{1-c_3}{1-c_3X} = R(X, a_1, ..., c_3)$$

where a_1, a_2, a_3, b_1, b_2, c_1, c_2, and c_3 denote the independent transition probabilities occurring in the network. The problem of finding these transition probabilities can be reformulated into the problem of finding these coefficients by appropriate minimization of the difference between $R(X, a_1, ..., c_3)$ and a given generative function GP(X). For example, the a_i and b_i can be found by low order approximations of GP(X).

In this section, we used the concept of generating function in order to introduce the 'transfer function' of a network. The transfer function can be evaluated in a direct way by first calculating the corresponding distribution P(L=k). However, we have indicated a more elegant method to obtain the transfer function directly from the topology of the underlying network. It was shown that this assignment can be formalized. This assignment fulfils three properties that relate the network topology on the one hand with the algebraic properties of the generating functions on the other hand.
The evaluation of TF(N) from N was one-way. In the next section, we will introduce a second tool, known as Padé-approximation, in order to find the network topology from its transfer function.

3. Network topology and the Padé-expansion

In the previous section, we have considered a method to derive a transfer function from the topology of a network. In this section, we will present a method to deal with the inverse problem, and we will encounter many examples of the many-to-one character of the assignment.
One of the mathematical tool involved is known as the Padé-expansion. In the following section, this expansion will be introduced informally. A more detailed analysis can be found in Lindemann (1981). The Padé-expansion can loosely be interpreted as the inverse operation of the power series expansion.

Examples. In figure 5, four different networks are presented. The default initial and final states are shown only in the first network.
The first network (indicated by A; equal to the network N_2 we studied before) represents a nesting of a parallel combination of two subnetworks, viz. an atomic network with probability a, and a series combination of two atomic networks with probability b and 1, respectively. If these atomic networks each contribute once for the duration, the nested subnetwork has as its transfer function (correspondence relation 1 and 2):

$$TF(N') = aX + (bX).X = aX + bX^2$$

Consequently (relation 3):

$$TF(N) = \frac{1-a-b}{1-aX-bX^2}$$

which is in accordance with our previous results for N_2.
The second network B with a nested loop has as its transfer function

$$TF(N) = \frac{(1-p)(1-qX)}{1-qX-p(1-q)X^2}$$

The third network C consists of a sequence of 2 self-loops connected serially. The self-loops have probability p and q, respectively. From correspondence relation 1, it follows that

$$TF(N) = \frac{1-p}{1-pX}\,\frac{1-q}{1-qX}$$

In the next section we will see that, although the first and third networks A and C seem very different from a topological point of view, they are very close in an 'algebraic' sense.
If one specific place-reversal step is introduced in the third network C, we obtain the situation shown in the fourth network. The probability of the backward transition from the second state towards the first state equals r. If we assume that the place-reversal does not take time, it can be shown by a straightforward calculation that the resulting transfer function reads:

$$TF(N) = \frac{(1-p)(1-q-r)}{(1-pX)(1-qX) - (1-p)r}$$

which evidently reduces to the transfer function of the second network in the case of r=0. Surprisingly, this generating function has an interpretation as generating function of some network *without* a place-reversality! (This is an example of the many-to-one assignment of generating functions.) The mathematical argument is that the denominator

$$(1-pX)(1-qX) - (1-p)r = pqX^2 - (p+q)X + (1-(1-p)r) = H(X)$$

Let F(X) denote some real function, defined on an interval [a, b], and assume that all its derivatives exist. Let c denote some intermediate point such that $a \leq c \leq b$. Then F(X) can be (d_1, d_2)-Padé-approximated by a quotient

$$\frac{P_1(X-c)}{P_2(X-c)}$$

P_1 and P_2 denoting polynomials of degree *at most* d_1 and d_2, respectively. The approximation is found by adequate minimization of the expression

$$\max |F(X)\, P_2(X) - P_1(X)|$$

by varying the coefficients, and considering the maximum over some adequate interval containing c.

Example. The (0, 1)-Padé-approximation around 0 of the power series

$$1 + aX + (aX)^2 + (aX)^3 + (aX)^4 + \ldots$$

is given by

$$\frac{1}{1-aX}$$

Any smooth function with continuous derivatives has a (d_1, d_2)-Padé-approximation for any combination (d_1, d_2).

A second tool that we may need in order to relate the transfer function TF(N) with the underlying network N is known from elementary algebra. Any rational function

$$\frac{P_1(X)}{P_2(X)}$$

with $\deg(P_1) \leq \deg(P_2)$ can be decomposed into a sum of partial fractions, i.e. a sum of 'simpler' rational functions

$$\sum_i \sum_n \frac{a_{in}}{L_i^n(X)} + \sum_j \sum_n \frac{L'_{jn}(X)}{Q_j^n(X)}$$

in which a, L(X), and Q(X) denote constants, linear, and quadratic polynomials in X, respectively. Here the quadratic polynomials $Q_j(X)$ are irreducible over R, which means that they have no real zeroes. The real zeroes of $P_2(X)$ equal the zeroes of the linear expressions $L_i(X)$; each pair of complex conjugate roots of $P_2(X)$ corresponds to one of the quadratic denominators $Q_j(X)$. The inner summation with respect to n runs up to the order of the respective zero of $P_2(X)$.
An example will yield more insight. The statement asserts that any quotient with arbitrary $P_1(X)$ with $\deg(P_1) \leq 7$ of the following form:

$$\frac{P_1(X)}{(1-X)^2(1+X^2)^3}$$

can be written as

$$\frac{a_1}{(1-X)^1} + \frac{a_2}{(1-X)^2} + \frac{a_3 + a_4X}{1+X^2} + \frac{a_5 + a_6X}{(1+X^2)^2} + \frac{a_7 + a_8X}{(1+X^2)^3}$$

for some specific $a_1, \ldots a_8$.

This decomposition of the rational function

$$\frac{P_1(X)}{P_2(X)}$$

into partial fractions corresponds to the decomposition of the underlying overall network into a set of simple subnetworks.

Examples.
The observed duration distribution for some speech segment may be found to approximate a specific function. Such a proximity can effectively be tested as soon as an appropriate 'distance' on duration distributions is defined. In this example we consider the case of an exponential decay in k, i.e.

$$P(L=k) = c\, a^k$$

with c denoting some normalization constant such that $\Sigma\, P(L=k) = 1$. Then

$$GP(X) = \sum_k c\, a^k X^k = \frac{c}{1-aX}$$

It directly follows that an appropriate network might consist of one selfloop[6].
A more complicated example is given when P(L=k) is such that *positive* constants a, b and c can be found such that

$$P(L=k) = c\frac{a^{k+1} - b^{k+1}}{a - b}$$

It can readily be shown that the corresponding GP(X) reads

$$\frac{1-a}{1-aX}\,\frac{1-b}{1-bX}$$

which can be modelled by a series of two self-loops with probability a and b.

The following examples deal with the case that duration distributions are member of a specific class, viz. $P(L=k) = k^n c^k$, where[7] n is an integer and c a real number between 0 and 1. Observe that this class contains the exponentially decaying distributions (take n=0). For $n > 0$, P is unimodal with a maximum at (-n/log(c)), and P tends to zero if k

---

[6] Or, more prudently, that the corresponding network should at least contain one self-loop with probability a.

[7] For simplicity, we here omit the normalization. Consequently, the results hold up to one specific constant. This scaling is irrelevant for all essential steps in the theory, such as the Padé-expansion.

tends to infinity. Unimodal distributions might be approximated by a distribution out of this class.

n = 0: We now have the exponentially decaying distribution $P(L=k) = c^k$, corresponding to one self-loop.

n = 1: Here $P(L=k) = kc^k$. Padé-expansion yields

$$X\frac{(1-c)^2}{(1-cX)^2}$$

corresponding to a series of one atomic network and two self-loops.
n = 2: Here $P(L=k) = k^2c^k$. Padé-expansion yields

$$\frac{cX + c^2X^2}{c + c^2}\ \frac{(1 - c)^3}{(1 - cX)^3}$$

corresponding to a series of a parallel network and three self-loops.

n = 3: Here $P(L=k) = k^3c^k$. Padé-expansion now yields

$$\frac{cX + 4c^2X^2 + c^3X^3}{c + 4c^2 + c^3}\ \frac{(1 - c)^4}{(1 - cX)^4}$$

corresponding to a series of two subnetworks: one consisting of three parallel transitions, and a second consisting of four self-loops in series.

In general, the distribution $P(L=k) = k^n c^k$ corresponds to a network which consists of a series of two subnetworks N_a and N_b. The network N_a consists of at most (n+1) parallel transitions with durations 0, 1, ..., n and specific probabilities. This network is related to the *nominator* polynomial in the Padé-expansion of GP(X).
The network N_b consists of a series of (n+1) self-loops, all with the same probability c. The resulting network is the 'product network' N_aN_b of the subnetworks N_a and N_b. This network corresponds to the *denominator* polynomial in the Padé-expansion of GP(X).

For an *arbitrary* smooth distribution P, it is always possible to find pairs (n_i, c_i) such that P is arbitrarily well approximated by a sum of distributions of the form described above. This is assured by the fact that P(L=k) can always arbitrarily well be approximated by a particular solution of a difference equation of arbitrary (but finite) order --which is the similar reason behind the existence of LPC-coding of a speech signal. As a conclusion, an arbitrary distribution can be obtained by a network consisting of arbitrarily many (but a finite number of) product networks N_aN_b. In practice, however, one should be careful in interpreting this result: nothing is specified about the 'recognition behaviour' of the networks. It might be the case that a network that is suboptimal with respect to duration modelling performs optimally with respect to overall recognition rates. Unfortunately, these rates are, up to now, hardly susceptible for a theoretical analysis, contrary to duration modelling.

In all these examples, the crucial step is the interpretation of the Padé-expansion as transfer function of a network[8]. Several cases of ambiguity can be observed. Firstly, there exists a principal ambiguity in the network interpretation of the algebraic terms a^nX^n where $a > 0$ and $n > 1$: the corresponding network may equally be interpreted as a transition of duration n with probability a^n, or a series of more than one transition, their total duration equalling n and the product of their probabilities eqalling a^n.
Another type of ambiguity is related to the summation as specified in the correspondence relation 1. The simple algebraic equation $X = 0.5\,X + 0.5\,X$ already yields a topological ambiguity between the atomic network with duration 1 and a parallel network of atomic networks with duration 1.
Precisely these ambiguities yield a 'class' of networks, all corresponding to the same transfer function, but with a different topological structure. Consequently, these networks may differ with respect to spectral modelling. On the basis of the duration, however, the search for an appropriate network topology may be inspired by a detailed study of the duration distribution. The following examples show that the interpretation deserves some care.

Example. Suppose we have the following transfer function

$$\frac{1-a}{1-aX}\frac{1-b}{1-bX}$$

As we have seen, if $0 \leq a, b < 1$ then this transfer function corresponds to the distribution

$$P(L=k) = c\frac{a^{k+1} - b^{k+1}}{a - b}$$

corresponding to a series of two self-loops (see the third network C in figure 5). However, not necessarily need a and b to be *positive* in order to render the resulting rational function interpretable as a transfer function! For example, if $a < 0$ and $(-a) < b$,

$$\frac{1-a}{1-aX}\frac{1-b}{1-bX} = \frac{1-(b+a)+ab}{1-(b+a)X+abX^2} = \frac{1-(b-(-a))-(-a)b}{1-(b-(-a))X-(-a)bX^2} = \frac{1-c_1-c_2}{1-c_1X-c_2X^2}$$

with $c_1 = b-(-a) > 0$ and $c_2 = (-a)b > 0$. And this rational function corresponds to the first network A presented in figure 5 with $c_1 = a$ and $c_2 = b$. We observe that these seemingly very different networks are algebraically rather close: it depends on the *sign* of just one parameter (in this case: a) which networks fulfils best.

Another very fine example of topological ambiguity is provided in figure 6. It is based on the algebraic identity

$$\frac{1}{2}X + \frac{1}{2}\frac{1 - 1/3}{1 - 1/3\,X} = \frac{1}{3} + \frac{2}{3}\left[X\left\{\frac{11}{12} + \frac{1}{12}\left\{X\,\frac{1 - 1/3}{1 - 1/3\,X}\right\}\right\}\right]$$

Networks may show *fractal* properties. There exist arbitrarily complicated ones; one simple example is shown in figure 7. It is based on the identities

---

[8]By the author, an algorithm is being developed to facilitate this interpretation. It parses an input function according to a grammar defined by the three correspondence relations.

$$\frac{1-a}{1-aX} = (1-a)\,\frac{1-aX+aX}{1-aX} = (1-a) + aX\left[\frac{1-a}{1-aX}\right]$$

We have considered a method to construct a network on the basis of an arbitrarily specified duration distribution. We conclude this section by presenting another universal network that is able of modelling any arbitrary duration distribution (figure 8). The verification that this network is indeed 'universal' is left to the reader.

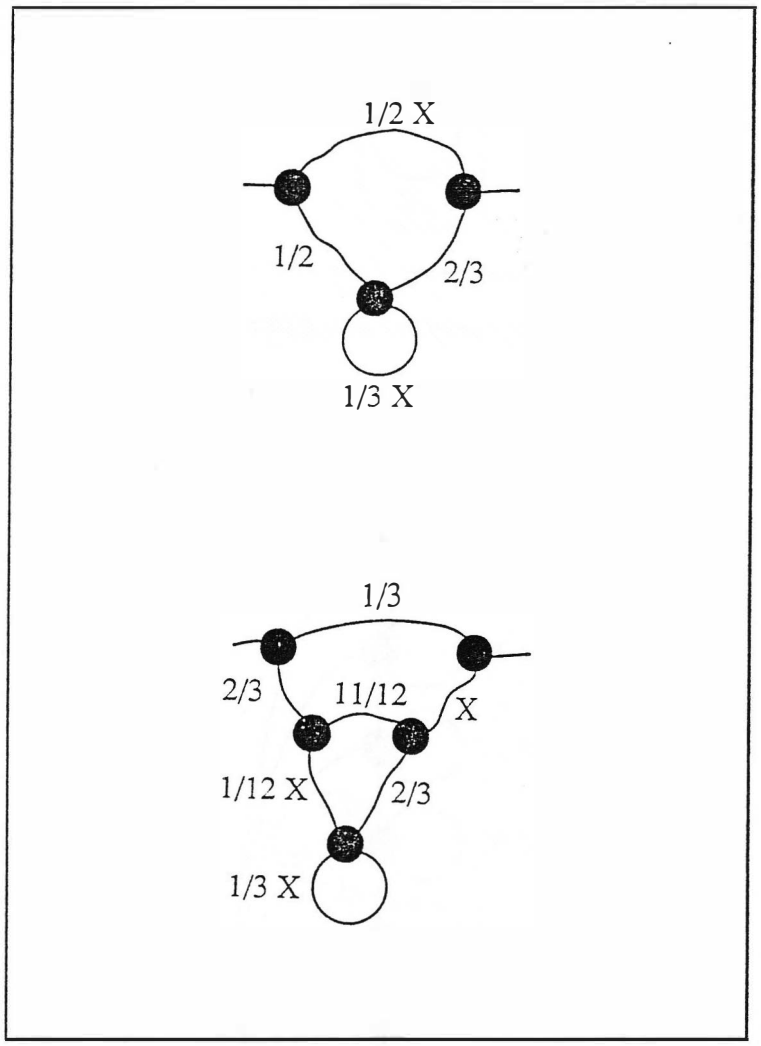

Fig. 6. Another example of topological ambiguity.

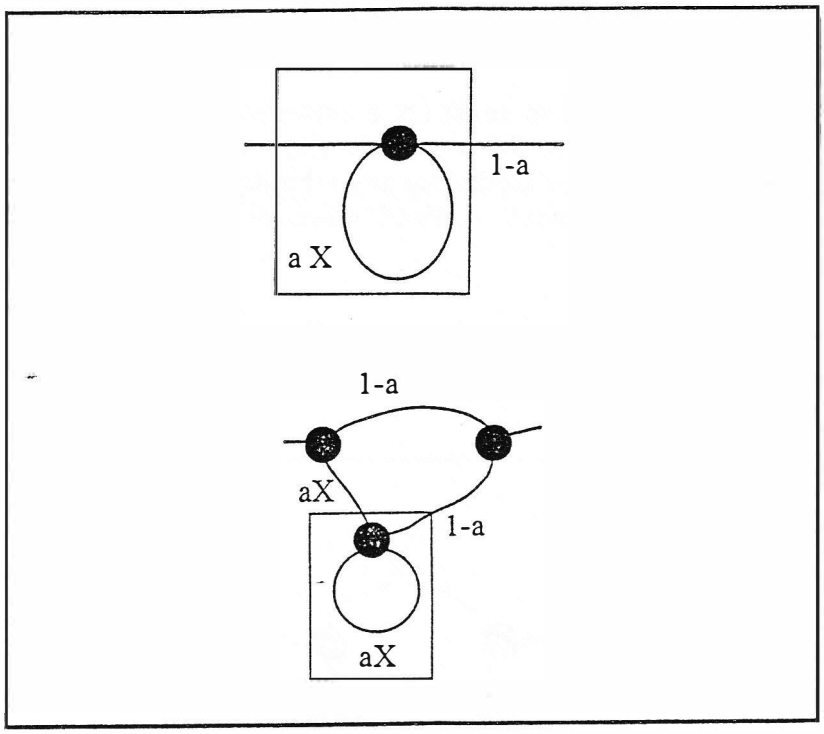

Fig. 7. An example of fractallike networks.

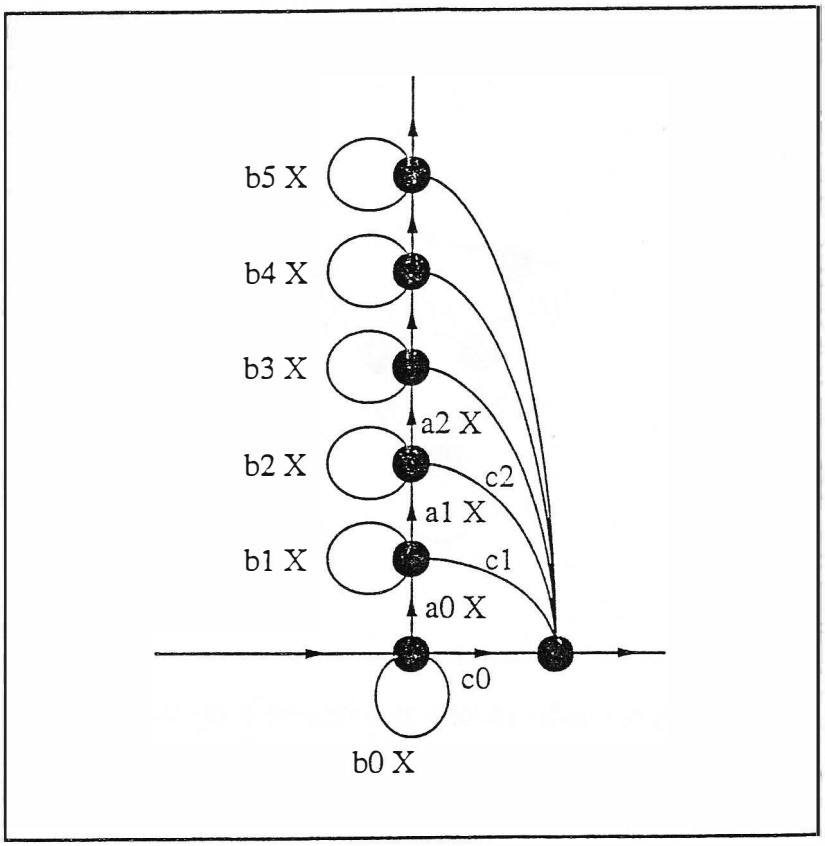

Fig. 8. An example of a universal network, capable of modelling any arbitrary duration distribution.

4. Conclusion

In this paper, a method was presented for assigning a 'transfer function' to an arbitrary network. The transfer function is always a so-called rational function, i.e. a quotient of two polynomials. By using concepts such as the duration probability density function, the generating function and the Padé-expansion, three 'correspondence relations' have been formulated. These relations define a relation between the topological structure of phone-like units on the one hand, and algebraic properties of the transfer function on the other. On the basis of arguments in one domain, results can be derived in the other domain.

The assignment of the transfer function to a network is not one-to-one. A transfer function does not define one specific network, but rather a 'class' of networks that are all equivalent with respect to their ability of durational modelling. The spectral modelling of these networks may differ; this aspect may be of importance in further optimization of network topologies. The topological correspondence between networks within one class has been studied in the paper.

By the author, an algorithm is being developed in order to facilitate the interpretation of given rational functions as a transfer function of some underlying network. Research in the near future will be focussing on the possibility of additional restrictions on the class of networks on the basis of their potential of *spectral* modelling.

The results may be of interest for a recently started AIO-project on 'Duration modelling in the HMM-approach' at our Institute by X. Wang. This project follows the AIO project on more general HMM recognition and implementation, also at the Institute, by P. van Alphen.

References

For a detailed introduction to HMM, the reader is referred to:

Holmes, J.N. (1988) *Speech Synthesis and Recognition*. Van Nostrand Reinholt (UK), Wokingham, England.

Lee, K.F. (1989). *Automatic Speech Recognition. The development of the Sphinx system*. Kluwer Academic Publishers. Boston, Dordrecht, London.

For an account of the Padé-approximation and partial fractions, introductions are provided by:

Ledermann, W. (1981). *Handbook of applicable mathematics. Volume I: Algebra.* (R.F. Churchhouse, ed.) John Wiley and Sons.

Ledermann, W. (1981). *Handbook of applicable mathematics. Volume III: Numerical methods.* (R.F. Churchhouse, ed.) John Wiley and Sons.

Errata

Page 65, line 2 from below. Read 'figure 3' instead of 'figure 4'.

Page 69, line 6 from below. Read 'third' instead of 'second'.

Page 70, line 3 from above. One additional argument is lacking, viz. that the average of both zeroes $(p+q)/(2pq)$ equals or exceeds unity if $pq > 0$: As $0 < p, q < 1$, $0 < pq \leq p$ and $0 < pq \leq q$. By addition, $0 < (2pq) \leq (p+q)$. Accordingly, $(p+q)/(2pq) \geq 1$.