Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 16 (1992), 27-34.

ON THE RELATION BETWEEN VOWEL DURATION IN HMM-MODELS AND THE TRANSITION PROBABILITIES IN PHONE-LIKE UNITS

Louis F.M. ten Bosch

Abstract

The relation will be discussed between the duration probability density function (pdf) of a segment, as observed in training speech material for an HMM recognition system on the one hand, and the transition and emission probabilities in a specific phone-like unit (PLU) on the other. First, a theoretical relation between the duration pdf of Dutch vowels and the HMM *transition* probabilities of the corresponding speech segments is reconsidered. Next, the theoretical basis of this relation is extended with respect to the incorporation of the HMM *emission* probabilities.

1. Introduction

In speech recognition applications based on hidden Markov modelling (HMM), usually a large number of HMM-parameters are to be trained in order to model the speech segment characteristics in a speech database. The training of the HMM-model, i.e. the iterative adjustment of the transition and emission probabilities in all the phone-like units by e.g. the Baum-Welch algorithm, is time-consuming. For example, the connected-speech recognition system REXY described by Van Alphen (1992), contains only 39 phone-like units and already more than 30,000 unknown parameters. One overall update of all its parameters takes several hours of CPU-time on a μ-VAX.

This paper concerns the question, how phonetic knowledge can be implemented into the statistically-based HMM speech recognition systems. Such an implementation may be useful in order to substantially reduce the required training. After a successful training procedure, an HMM-based system contains 'statistical knowledge' about the characteristics of speech segments. This implicit 'statistical' knowledge, which is encoded in the transition and emission probabilities, should at least be in line with the explicit 'phonetic knowledge' about the segments. In this paper, we will focus on the relation between aspects of phonetic knowledge on the one hand, and explicit statements about the HMM parameters on the other, in the case of Dutch vowels. Such phonetic knowledge is available in e.g. the description of many phonetic details about the influence of speaking rate on the realization of speech segments (c.f. Strange, 1989; for Dutch vowels: Van Son & Pols, 1990).

Suppose two speech databases contain identical texts carefully spoken by the same speaker under identical circumstances, the only difference being, for example, the

speaking rate. Two independent HMM training sessions on these databases will result in two different HMM parameter sets. However, the phonetic similarity between both databases will be reflected in some kind of similarity between both HMM parameter sets. This similarity provides a possibility of an intelligent training procedure, such as a 'warm' start, if an HMM is to be trained on a speech database spoken at fast rate while the HMM parameters are already known for the normal speech rate. The study of the similarity between both HMM parameter sets gives insight in the potential duration modelling of HMM. If such a similarity between the parameter sets cannot be found in this case (in which only one one-dimensional parameter is varied), one might doubt whether an intelligent incorporation of phonetic knowledge in general into an HMM recognition system is ever possible.

The differences between corresponding HMM parameters in the two parameter sets reflect the statistical differences between the realizations of the speech segments. The transition and emission probabilities are related (in a complex way) to the statistical properties of the segment duration as well as of the spectral details. However, from a theoretical point of view, the modelling of duration by an HMM can be understood more easily than the spectral modelling. One reason is that the classical theory of Markov chains can be used to explicitly relate the modelled segment duration and the HMM *transition* parameters. Secondly, in the literature, several attempts have been described to model duration in HMM by using additional techniques (e.g. clustering: Picone, 1989; Lerner & Mazor, 1990, and references therein).

In ten Bosch (1991), an explicit relation was formulated between the transition probabilities within one phone-like unit on the one hand, and the duration probability density function (duration pdf) of the corresponding segment on the other hand. This analysis was hypothetical in the sense that the *emission* probabilities in the HMM have not been taken into account. However, these emission probabilities play an important role in the actual modelling of duration, as is clear from the structure of the Baum-Welch-algorithm (c.f. Lee, 1989). In this paper, we will consider how the original theoretical set-up in ten Bosch (1991) can be extended with respect to these emission probabilities. This extended theory relates the HMM parameters of one phone-like unit (i.e. its transition and emission probabilities) to one explicit phonetic parameter (i.e. the modelled segment duration).

2. Shortcomings of the previous theoretical model

The theoretical model in ten Bosch (1991) specifies a relation between the duration pdf of a speech segment, and the transition probabilities in the HMM phone-like unit (PLU). This relation reads as follows. Suppose P(N) denotes the probability of a segment having a duration between Nd and (N+1)d (d denoting the 'bin width' in a duration histogram; this parameter typically has a value of about 10 ms). The corresponding *generating function*, defined by $F(X) = \sum P(N) X^N$ can be interpreted as a transfer function of the PLU by means of its Padé expansion: $F(X) = A(X)/B(X)$. In this formula, X denotes a formal variable. It was shown in ten Bosch (1991) that a *topological* structure of the corresponding PLU can be found on the basis of the *algebraic* structure of the rational function A(X)/B(X). Here, 'topological structure' must be understood in the wide sense, i.e. not only with respect to the 'qualitative' topological connections between hidden states, but also with respect to 'quantitative' details such as the transition probabilities. The overall tranfer function F(X) of the PLU is composed of (many) transfer functions of the simple linear form $a_{ij}.X$, a_{ij} denoting the transition probability from state S_i to state S_j. The actual calculation is rather technical but conceptually straightforward; for details the reader is referred to ten Bosch (1991). Essential here is the observation that the duration pdf specifies a transfer

function that fully depends on the transition probabilities a_{ij}.

As a consequence, these transition probabilities can be calculated on the basis of the duration pdf of a segment. In the present discussion, we will use the 7-state, 12-transition PLU-model as used by Lee (1989), Van Alphen (1992), and Wang *et al.* (1992) (figure 1). This PLU is transition-assigned. It has been shown in ten Bosch (1991) that this particular PLU-model is capable of modelling a large class of duration pdf's that can be described by the existence of a *linear* relation between P(N), P(N+1), P(N+2) and P(N+3) for all N > 3. More precisely, all pdf's P(N) with the property that $P(N+3) = a_1P(N+2) + a_2P(N+1) + a_3P(N)$ for all N > 3 such that $X^3 = a_1X^2 + a_2X + a_3$ has three real roots between -1 and 1, can exactly be modelled by the particular 7-state, 12-transition PLU. This seems to be a rigid constraint, but it appears that most of the vowel duration pdf's observed in the REXY training set can adequately be modelled by this 7-state, 12-transition PLU.

By using the theoretical relation between the duration pdf and the PLU-topology, the 'optimal' PLU of 13 Dutch monophthongs (i.e. the transition probabilities) have been calculated from the observed duration pdf's. This was performed by minimizing the cross entropy between the observed duration pdf and the pdf modelled by the PLU. The results are presented in table I. It must be remarked that the emission probabilities were tied: only three different emission pdf's (corresponding to 'begin', 'center' and 'end' of the speech segment) were used. These pdf's are also tied to the first, second and third selfloop, respectively.

From table I, it can be observed that there exist substantial differences between the 'predicted' model data and the actual observed data. The column 'theoretical a_{ii}' presents the values of the *theoretical* selfloop probabilities of a_{22}, a_{33} and a_{44}. The column 'actual a_{ii}' shows the values as actually found after the Baum-Welch optimization. For most of the vowels (all but schwa, denoted by '@'), the theoretical selfloop probabilities are equal (within one PLU) up to a deviation less than 1 percent, ranging from 0.81 for a long vowel /ø/ to 0.53 for a short /I/. In case of the schwa, the selfloop probabilities read 0.19, 0.32 and 0.36, which means that the schwa has a 'deviant type' pdf and a relatively short mean duration. The actual data show a tendency $a_{22} > \max(a_{33}, a_{44})$, the ordering between a_{33} and a_{44} being less consistent. Also, the phonetic duration of the vowels can be traced back in the theoretical data (a larger selfloop probability yields in general a longer duration).

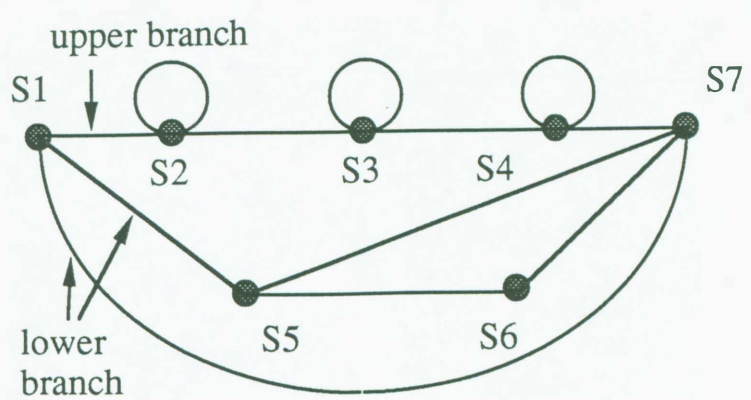

Figure 1. The 7-state, 12-transition phoneme-like unit (PLU) as used by Lee (1989), Van Alphen (1992) and Wang *et al.* (1992).

However, it is clear that the model data, which are not based on any prior knowledge of the emission probabilities, do not reflect specific details present in the actual data. This within-PLU discrepancy between model data and actual data can be understood by considering an extended theoretical model, as will be shown in the next section. The extended analysis reveals a relation between the emission pdf's, the set of feasible sequences of acoustic observations of the speech segment, and the transition probabilities in the PLU.

3. Extension of the theoretical model

In order to extend the theoretical model, we will study the influence of the emission probabilities on the PLU by having a look at the so-called forward algorithm. The forward algorithm evaluates the probability $P(O \mid \lambda)$, O and λ denoting the sequence of observations $\{O_1, ..., O_N\}$, and the particular PLU-model, respectively. Here, and throughout the following sections, N will denote the *length* (number of observations) of the observation sequence O.

The PLU model has an upper branch, consisting of a series of three selfloops, as well as a lower branch, consisting of parallel paths. The upper branch is capable of modelling durations *exceeding* 3d, whereas the lower branch models durations d, 2d and 3d. The upper branch consists of a serial combination of 5 states S_1, S_2, S_3, S_4, S_7 and is topologically specified by the four 'simple' transitions S_1S_2, S_2S_3, S_3S_4, S_4S_7 and three selfloops S_2S_2, S_3S_3 and S_4S_4. As the bulk of the pdf-data has maxima between 8d and 13d, we will consider the contribution of the upper branch to $P(O \mid \lambda)$.

Tabel I. The 'predicted' and actual selfloop probabilities of S_2S_2, S_3S_3 and S_4S_4. In the theoretical case, these values differ less than 1%, except for the schwa ('@'). All figures are multiplied by 1000.

phoneme symbol	as in Dutch	selfloop probabilities a_{22}, a_{33} and a_{44} theoretical a_{ii}			actual a_{ii}		
a	'paal'	738			793	878	759
ɑ	'pal'	568			751	728	619
e	'beet'	750			854	843	756
ɛ	'bed'	560			760	689	628
i	'piet'	710			806	717	742
I	'pit'	528			750	642	515
u	'boek'	606			739	533	642
o	'poot'	718			862	792	779
ɔ	'pot'	584			697	737	580
y	'muur'	627			762	738	716
ø	'peut'	810			817	848	804
œ	'put'	553			592	481	579
@	'de'	193	321	358	656	565	451

Let $\alpha(t, i)$ denote the probability of the joint event of observing O_t during time step t, having already observed the observations O_1, ..., O_{t-1}, and arriving in hidden state S_i. Furthermore, b_{tij} denotes the emission probability of observation O_t during transtition S_iS_j, and a_{ij} denotes the transition probability of transtition S_iS_j. Then, in the present case, by the forward algorithm:

$$\begin{aligned} \alpha(0, i) &= 1 && (i = 1) \\ &\quad \text{or } 0 && \text{(for all other i)} \\ \alpha(t, i) &= \textstyle\sum_j \alpha(t\text{-}1, j) \,.\, a_{ji} \,.\, b_{tji} && \text{(for all i,} \\ & && 1 \le t \le N) \end{aligned} \qquad (1)$$

such that $P(O \mid \lambda)$ equals $\alpha(N, 7)$. All the observations O_t are supposed to be member of a (finite) codebook.

In this formulation of the forward algorithm, it is not straightforward to derive an explicit formula for the duration pdf of a modelled segment. If $O = \{O_1, ..., O_N\}$ runs through all possible sequences of N observations, $P(N) = \sum_O P(O \mid \lambda)$ denotes the probability of observing a sequence of length N. If all the pdf's b_{tji} were independent of t (so $b_{tji} \equiv b_{ji}$), then $\sum_O P(O \mid \lambda)$ could easily be evaluated by using the previous theoretical model using the transfer function F(X) of the upper branch. The transfer function of a transition S_iS_j is then given by the linear polynomial $a_{ij}.b_{ij}.X$. However, if the emission pdf's depend on t (which is a much more natural case), the coefficients of the transfer function have to be modified in a more complex way. We will present a solution for the resulting modified transfer function which allows an interpretation of our previous observations in a natural way. This solution is given by the following algorithm:

1) Restrict for each transition S_iS_j a set $\mathbf{O}_{ij}$ of 'possible' observations O_k, e.g. by selecting those O_k from the codebook for which b_{kij} exceeds a certain threshold. This is to have a localization and limitation of the feasible observation space on the basis of the emission pdf's. (It makes no sense to deal with all possible sequences without any spatial correlation between subsequent observations.) For example, if the emission pdf's of a PLU are tied to three different ones, each corresponding to a selfloop, we can construct three (different) sets $\mathbf{O}_{22}$, $\mathbf{O}_{33}$, and $\mathbf{O}_{44}$, corresponding to the three different emission pdf's. In general, the sets $\mathbf{O}_{ij}$ are just special subsets from the codebook used. By this construction, the three emission pdf's define a certain 'metric' on the set of observations;

2) Set $\gamma_{ij} = \sum_{Ok \in Oij} b_{kij}$, i.e. the sum of the emission probabilities of the observations O_k in the set $\mathbf{O}_{ij}$ corresponding to transition S_iS_j;

3) Allot to each transition S_iS_j a weighting w_{ij} equal to the product of the transition probability a_{ij} and the corresponding γ_{ij} found in step 2: $w_{ij} = a_{ij}\,\gamma_{ij}$;

4) Evaluate the transfer function of the upper branch F(X) by using the newly defined transfer function $w_{ij}.X$ of each transition S_iS_j;

5) Normalize F(X) to $F_{norm}(X)$ according to the additional constraint $F_{norm}(1) = 1$ by putting $F_{norm}(X) := F(X)/F(1)$.

We observe that this algorithm relates its output F(X) (or $F_{norm}(X)$) to three different aspects of the HMM: (a) the sets $\mathbf{O}_{ij}$, i.e. the specification of all possible sequences O, (b) the transition probabilities, and (c) the emission probabilities. The framework of the algorithm in step 4 is essentially identical to the framework used in ten Bosch (1991). The definition of the newly defined transfer function can be considered as an extension of the definition of the transfer function in the previous case,

since the old transfer function can be obtained from the extended function by simply setting $\gamma_{ij} = 1$ for all transitions S_iS_j in the extended model, in other words: by not clamping the preferred location of the observations O_t.

Secondly, we make a remark about the interpretability of the transfer functions corresponding to a transition S_iS_j. These functions are all of the form $F(X) = w_{ij}X$, where w_{ij} denotes the normalized weighting of the corresponding transition. In the same way as we have seen in the old model can the transfer function (and accordingly the weightings w_{ij}) be evaluated on the basis of the known actual duration pdf of the segment being modelled by the PLU. However, in the extended version the optimized parameters w_{ij} of the transfer function $F(X)$ do not denote the transition probabilities a_{ij}, but they denote a product $a_{ij}\ \gamma_{ij}$ instead. In other words, given the duration pdf of a segment, there is

1) a trade-off between the actual transition probability a_{ij} and the value of γ_{ij} (their product w_{ij} being determined by the duration pdf), and

2) a degree of freedom with respect to the actual values of b_{kij} within one set O_{ij} (their sum γ_{ij} indirectly being determined by the relation $w_{ij} = a_{ij}\ \gamma_{ij}$).

These two points make clear how the information on spectral modelling and duration modelling can interact within the PLU. The durational information of a segment clamps the set of feasible PLU's (i.e. the set of all possible PLU's with identical topology that are potentially capable of modelling the segment) to a specific subset. By these points we are able to understand the discrepancies found between model data and observed data in the previous section. On the one hand, we observed that, in general, the selfloop probability a_{22} exceeds a_{33} as well as a_{44}. On the other hand, the weightings w_{ii} were found to be equal (within each PLU) for almost all vowels, which yields the following equalities: $w_{22} = w_{33} = w_{44}$, so $a_{22}.\gamma_{22} = a_{33}.\gamma_{33} = a_{44}.\gamma_{44}$ for almost all vowels. Since a_{22} tends to be larger than a_{33} or a_{44}, γ_{22} tends in general to be smaller than γ_{33} or γ_{44}, which means that the probability sum of the observations in the first set O_{22} is allowed to be smaller than are the probability sums of the observations in O_{33} and O_{44}. In other words, either (a) the set O_{22} is small and only a few observations that occur in the initial part of the observation sequence O are allowed to be member of O_{22}, or (b) the set O_{22} contains many elements with low probability (e.g. O_{22} is spatially extended and has many elements 'remote' from its center if a Gaussian emission pdf is assumed), or (c) a combination of these effects occurs. From this point of view, the modelling of the initial vowel segments is less adequate than is the modelling of the center part or the final part.

With respect to point 2, nothing can be specified without additional, detailed information on the emission pdf's used.

The derivation of the algorithm is straightforward. For simplicity, we consider the case of three tied emission pdf's: $\gamma_{12} = \gamma_{22}$; $\gamma_{23} = \gamma_{33} = \gamma_{34}$; $\gamma_{44} = \gamma_{47}$. Let O_{22}, O_{33} and O_{44} denote the three appropriate subsets chosen from the discrete codebook. (This construction can easily be modified in the case of a continuous-density HMM.) If $O = \{O_1, O_2, ..., O_N\}$ is a specific observation sequence of length N, where the first $k_1 + 1$ observations are to be (arbitrarily) chosen from O_{22}, the subsequent $k_2 + 2$ observations from O_{33}, and the last $k_3 + 1$ observations from O_{44}, then $P(O \mid \lambda)$ is given by the product

$$P(O \mid \lambda) = (a_{12}\ a_{22}^{k1}\ a_{23}\ a_{33}^{k2}\ a_{34}\ a_{44}^{k3}\ a_{47}).\ (\gamma_{22}\ \gamma_{22}^{k1}\ \gamma_{33}\ \gamma_{33}^{k2}\ \gamma_{33}\ \gamma_{3}^{k3}\ \gamma_{44}) \qquad (2)$$

where γ_{ii} reads $\gamma_{ii} = \sum_{Ok \in Oii} b_{kii}$ $(i = 2, 3, 4)$.

The probability of observing an arbitrary observation sequence O of length N, with its initial part (of length ≥ 1) in the set $\mathbf{O}_{22}$, the center part (of length ≥ 2) in $\mathbf{O}_{33}$, and its final part (of length ≥ 1) in $\mathbf{O}_{44}$, reads

$$P(\{\text{length}(O) = N\} \,\&\, \{O \in \mathbf{O}_{22}\,\mathbf{O}_{33}\,\mathbf{O}_{44}\} \mid \lambda) = \sum P(O \mid \lambda) \qquad (3)$$

where O runs though all sequences of length N, i.e. the summation in the right-hand side is over all cases of formula (2) where $k_1 + 1 + k_2 + 2 + k_3 + 1 = N$ with $k_i \geq 0$. However, this sum can be interpreted in a different way, viz. as the coefficient of X^N in the Taylor expansion around $X = 0$ of the following rational function F(X):

$$F(X) = a_{12}\gamma_{12}X\frac{1}{1-a_{22}\gamma_{22}X}a_{23}\gamma_{23}X\frac{1}{1-a_{33}\gamma_{33}X}a_{34}\gamma_{34}X\frac{1}{1-a_{44}\gamma_{44}X}a_{47}\gamma_{47}X \qquad (4)$$

where, according to the tied emissions, $\gamma_{12} = \gamma_{22}$; $\gamma_{23} = \gamma_{33} = \gamma_{34}$; $\gamma_{44} = \gamma_{47}$. Expression (4) itself can in turn be interpreted as the most general transfer function of the upper branch of the 7-state, 12-transition PLU without tied emission pdf's, in which the old transfer functions corresponding to individual transitions $a_{ij}\,X$ have been replaced by the extended versions $a_{ij}\gamma_{ij}\,X$.

This derivation explains steps 1, 2, 3 and 4 of the scheme. Step 5, the normalization, is necessary in order to evaluate the probability of an arbitrary observation sequence with, in the tied case, its initial part in the set $\mathbf{O}_{22}$, the center part in $\mathbf{O}_{33}$, and its final part in $\mathbf{O}_{44}$, of having an overall length equal to N, more specifically, the probability

$$P(\{\text{length}(O) = N\} \mid \{O \in \mathbf{O}_{22}\,\mathbf{O}_{33}\,\mathbf{O}_{44}\} \,\&\, \lambda) \equiv$$

$$\{\text{coefficient of } X^N \text{ in } P_{norm}(X)\}$$

Step 5 is based on Bayes, as well as the observation that the number F(1) denotes the probability of an observation sequence of *arbitrary* length to have its initial part in the set $\mathbf{O}_{22}$, the center part in $\mathbf{O}_{33}$, and its final part in $\mathbf{O}_{44}$:

$$\{\text{coefficient of } X^N \text{ in } P_{norm}(X)\} =$$

$$P(\{\text{length}(O) = N\} \mid \{O \in \mathbf{O}_{22}\,\mathbf{O}_{33}\,\mathbf{O}_{44}\} \,\&\, \lambda) =$$

$$P(\{\text{length}(O) = N\} \,\&\, \{O \in \mathbf{O}_{22}\,\mathbf{O}_{33}\,\mathbf{O}_{44}\} \mid \lambda) \,/\, P(\{O \in \mathbf{O}_{22}\,\mathbf{O}_{33}\,\mathbf{O}_{44}\} \mid \lambda) =$$

$$P(\{\text{length}(O) = N\} \,\&\, \{O \in \mathbf{O}_{22}\,\mathbf{O}_{33}\,\mathbf{O}_{44}\} \mid \lambda) \,/\, F(1) =$$

$$\{\text{coefficient of } X^N \text{ in } P(X)\} \,/\, F(1)$$

where we used equation (3) and the generating-function interpretation in the last step. As a consequence, $F_{norm}(X) = F(X)/F(1)$, which is step 5 of the algorithm.

3. Conclusion

In this paper, we studied a possible extension of a previous algorithm described in ten Bosch (1991). That previous algorithm provides an optimal PLU (topology including transition probabilities) for modelling a given duration probability density function. The emission probabilities have not been taken into account.

The previous theory does not cope with details found in observed data. First we show that substantial differences may exist between the actual transition probabilities and the predicted ones. This was shown in the case of 13 Dutch vowels, by comparing the REXY data and the predicted values.

Next, a theoretical extension has been proposed in order to interpret these differences by considering the emission pdf's. It appears that this extension is genuine in the sense that it is compatible in all cases in which the old approach is applicable.

From the structure of the forward algorithm and the proposed theoretical extension, it follows that knowledge of the duration pdf is not sufficient to explicitly evaluate the transition probabilities as well as the emission probabilities within one PLU. In the equal-pdf case, the transition probabilities can be evaluated. In the general case, the duration pdf only specifies a subset of feasible PLU's. By using the extended theory, the precise structure of this PLU-subset is known.

In the future, this study will be extended to more segments, viz. all 39 segments used in the REXY-system. Moreover, a study of systematic deviations from the case where a_{22}, a_{33} and a_{44} play a more symmetric role is in preparation. In particular, we will study the results of an HMM training session of a speech database in *reversed* time.

References

Bosch, L.F.M. ten (1991). "On relations between phone models, segment duration, and the Padé-expansion". *Proceedings of the Institute of Phonetic Sciences*, Univ. Amsterdam, the Netherlands, **15**: 61-77.

Wang, X., ten Bosch, L.F.M., and Pols, L.C.W. (1992). "Dimensionality and correlation of observation vectors in HMM-based speech recognition". This volume.

Lee, K.F. (1989). *Automatic speech recognition: the development of the SPHINX system*, Kluwer Academic Publ., Boston.

Lerner, S., and Mazor, B. (1990). "Issues related to the estimation of time variant hidden Markov Models". *Proceedings ICASSP* 1990, 553- 557.

Picone, J. (1989). "On modeling duration in context in speech recognition". *Proceedings ICASSP* 1989, 421-424.

Strange, W. (1989). "Dynamic specification of coarticulated vowels spoken in sentence context". *J. Acoust. Soc. Am.* 85(5), 2135-2153.

Van Alphen, P. (1992). *HMM-based continuous speech recognition: Systematic evaluation of various system components*, PhD.-thesis, Univ. Amsterdam, the Netherlands.

Van Son, R., and Pols, L.C.W. (1990). "Formant frequencies of Dutch vowels in a text, read at normal and fast rate". *J. Acoust. Soc. Am.* **88(4)**: 1683-1693.