Proceedings of the Institute of Acoustics

PROGRESS TOWARDS A UNIFIED MODEL FOR SPEECH PATTERN PROCESSING

Martin Russell (1) and Wendy Holmes (2)

(1) School of Electronic and Electrical Engineering, The University of Birmingham, Edgbaston, Birmingham B15 2TT
(2) Speech Research Unit, DERA Malvern, St Andrews Road, Malvern, Worcs WR14 3PS

1. INTRODUCTION

Although hidden Markov model (HMM) based systems have dominated speech recognition research for the past decade [1], there is a growing awareness of their limitations in the full context of acoustic speech pattern modelling. Of these limitations, the most frequently cited are *temporal independence* and *piecewise stationarity*. Consider an acoustic speech pattern y comprising a sequence of acoustic feature vectors $y=y_1,y_2,\ldots,y_T$. The temporal independence assumption states that there is no direct statistical dependency between each feature vector y_t and the other vectors in the sequence. Since the feature vectors are typically sampled every 10ms, and their function is to capture perceptually relevant information about the speech signal, the temporal independence assumption is clearly inappropriate. The piecewise stationarity assumption states that the underlying structure of the sequence y is piecewise constant as a function of time, with instantaneous transitions between the constant sections. Again this is quite inappropriate because speech is produced by a continuously moving physical system (the vocal tract) and the structure of speech patterns reflects this.

Segmental HMMs were introduced in an attempt to overcome these limitations by identifying the states of the underlying Markov process with sequences of feature vectors, rather than with individual feature vectors. By treating sequences of feature vectors as a homogeneous unit, it is possible to model speech pattern dynamics and to accommodate inter-vector correlation. A number of different types of segmental HMM have been studied, and an overview is presented in [2]. An example of a segmental HMM is the probabilistic trajectory model [3] in which the states of the underlying Markov process are associated with continuous paths, or trajectories, in the feature vector space. These trajectories model feature vector evolution throughout a particular segment of speech, and an observed sequence of feature vectors is then treated as a random function of that trajectory.

One shortcoming of this type of segmental HMM is that the trajectory is defined in the acoustic feature vector space Y, whose elements will typically represent spectra or cepstra. These are chosen because they provide a good representation of the salient 'instantaneous' properties of the speech signal. However, in general they provide a very poor representation of speech pattern dynamics, because although movement is seen as changes within a frequency band, its relationship with the motion of the major articulators is typically across frequency bands. It can also be argued that by representing the trajectory in the acoustic feature vector space one is simply providing an even more complex model of surface detail and still not addressing the mechanisms which give rise to that detail [5]. An obvious solution is to replace the spectrum-based representation with an articulatory, formant, or other type of

production-based representation. However the determination of this type of representation is a non-trivial pattern processing problem in its own right. Feature extraction errors at this level will be irrecoverable and their effects will be propagated into the speech recognition process. In other words, this approach is contrary to the principle of delayed decision making.

These arguments suggest an extension to the segmental HMM framework in which trajectories are defined in some intermediate space $\mathcal{G}$, which is suitable for modelling speech pattern dynamics and where proper constraints can be applied, and then mapped into the acoustic vector space Y. The resulting model could be referred to as a multiple-level segmental HMM. In addition, the use of such a model as a generator is an example of model-based speech synthesis and from this perspective it can legitimately be considered as a unified model for speech pattern processing. The desirability of such a unified model and its potential significance has already been noted [6], [5].

The notion of such a multiple-level, unifying model raises a number of fundamental issues. These include the choice of an appropriate intermediate representation, the characterisation of dynamics and statics within that representation, the mappings which determine the relationship between the intermediate representation and the state-level and surface-level representations, and the role of assumptions of randomness in these relationships. The purpose of this paper is to propose a mathematical framework for describing such a model, so that the above issues can be properly understood, and to consider how existing developments can be expressed in terms of this framework. Such a framework will also help to understand the relationship with conventional models and to identify particularly simple multiple-level models which might be most amenable to mathematical analysis.

2. CONVENTIONAL HMMS

Recall that an N-state HMM M is a statistical model defined by:

- An underlying N-state Markov model, specified by an NxN state transition probability matrix A and an initial state probability vector π, such that if σ_i denotes the ith state of M, then

  $a_{ij} = Prob(\sigma_j \text{ at time } t \mid \sigma_i \text{ at time } t\text{-}1)$, and $\pi_i = Prob(\sigma_i \text{ at time } t=1)$.

  The matrix A and vector π specify the sequential and durational structure of M.

- For each state σ_i a probability density function (pdf) b_i defined on the space of acoustic feature vectors Y, such that $b_i(o) = Prob(y_t = o \mid x_t = \sigma_i)$. Here, $y = y_1, \ldots, y_T$ is a sequence of acoustic feature vectors and $x = x_1, \ldots, x_T$ is a state sequence of length T. For simplicity, suppose that b_i is Gaussian with mean μ_i and covariance matrix v_i.

2.1 Relationship between the state space Σ and surface representation Y

From the above definition, M is a two-level model comprising a finite, sub-phonetic, state-level representation space $\Sigma = \{\sigma_1, \ldots, \sigma_N\}$, and a surface acoustic feature vector space Y. The pdfs b_i define

mappings in both directions between these two representations. Define $\varphi: \Sigma \rightarrow D(Y)$ by $\varphi(\sigma_i) = b_i$ (the notation $D(S)$ is used to denote the set of pdfs over an arbitrary set S). Equivalently, if the "synthesis" mapping $\varphi': \Sigma \rightarrow Y$ is defined by $\varphi'(\sigma_i) = \mu_i$, then one can think of φ as defined by $\varphi(\sigma_i) = \varphi'(\sigma_i) + w_i$, where w_i denotes zero mean Gaussian noise with covariance v_i. The "classification" mapping $\gamma: Y \rightarrow D(\Sigma)$ is defined by $\gamma(y) = (b_1(y), \ldots, b_N(y))$, the symbol $D(\Sigma)$ denoting the set of pdfs defined on Σ. These mappings are illustrated in figure 1(a).

3. MULTIPLE-LEVEL HMMS

From the perspective of speech pattern processing, the purpose of a multiple-level HMM is to accommodate intermediate representations of speech in between the state-level and surface-acoustic level representations. For simplicity this discussion will be restricted to models which include a single intermediate level ϑ between Σ and Y, as depicted in figure 1(b). From the discussion in section 1, the elements of ϑ should be thought of as vectors of articulatory, formant or other production-related parameters. This section is concerned with the definition of the relationships between these levels.

3.1 Relationship between the state space Σ and intermediate representation ϑ

In a hidden model, each state σ_i will correspond to a sub-set of ϑ. Hence the state space Σ is related to the intermediate level ϑ by associating each σ_i with a pdf b_i defined on ϑ, just as with a conventional HMM. For consistency with the previous section, the notation $\varphi: \Sigma \rightarrow D(\vartheta)$ and $\gamma: \vartheta \rightarrow D(\Sigma)$ will be used to denote the mappings between the state and intermediate levels in the in the multiple-level case. According to these definitions, the first level of a multiple-level HMM is just a conventional HMM.

3.2 Relationship between the intermediate representation ϑ and the surface representation Y

The relationship between ϑ and Y is more complex. If ϑ is finite, then in principle each of its elements can be associated with an element of Y or a pdf defined on Y. However, in general ϑ will not be finite and it will be necessary to define a mapping $\eta': \vartheta \rightarrow Y$ or $\eta: \vartheta \rightarrow D(Y)$. If ϑ is an articulatory or formant feature space, then η' corresponds to articulatory or formant synthesis respectively and is typically many-to-one. As usual, ambiguity and variability in the acoustic realisation of an element ι of ϑ can be accommodated by adding a random component, so that the probabilistic mapping becomes $\eta(\iota) \rightarrow \eta'(\iota) + w$, where w denotes zero mean Gaussian noise whose covariance ρ_ι is a function of ι. Equivalently, $\eta(\iota)$ is a Gaussian distribution over Y with mean $\eta'(\iota)$ and covariance matrix ρ_ι.

The 'inverse' mapping from Y to ϑ corresponds to the derivation of articulatory, formant or other production-based parameters from acoustic data and is typically complex and one-to-many [7], [8]. Indeed, if this were not the case then there would be no need for the type of framework which is currently being considered. If Y can be thought of as finite, for example as a result of vector quantisation, then the mapping can be defined explicitly by listing the element (or elements) of ϑ which correspond to each element of Y. This is the basis of the approach to formant analysis described in [7]. However, in general Y will not be finite and as in the case of a conventional HMM, it is necessary to think in terms of a more general mapping $\xi: Y \rightarrow D(\vartheta)$, where $D(\vartheta)$ denotes a set of pdfs defined on ϑ.

This is a generalisation, for example, of the derivation of multiple hypotheses in formant analysis [7]. The mapping ξ will be complex and non-linear. For example, such mappings have been defined using polynomial regression [9] and adaptive network techniques [8], [10]. Note that η and ξ may need to be differentiable for the parameter optimisation problem to be soluble by calculus based methods.

3.3 Probability calculation in a multiple-level HMM

For any acoustic feature vector $y \in Y$ and state $\sigma \in \Sigma$ the most basic HMM computations require the evaluation of the probability $P(y|\sigma)$. For any element $\iota \in \mathcal{I}$, the joint probability $P(y, \iota|\sigma)$ is given by $P(y, \iota|\sigma) = P(y|\iota, \sigma)P(\iota|\sigma)$. According to the assumptions of the previous paragraphs, $P(\iota|\sigma)$ and $P(y|\iota, \sigma)$ arise from Gaussian pdfs with means $\phi(\sigma)$ and $\eta(\iota)$ and covariance matrices γ and ρ_ι respectively, and hence can be computed. In reality the 'correct' value of ι is not known and so ι must be integrated out to obtain the required probability:

$$P(y|\sigma) = \int_{\iota} P(y, \iota|\sigma) = \int_{\iota} P(y|\iota)P(\iota|\sigma)$$

3.4 Summary

A simple framework for incorporating an intermediate representation into a conventional HMM has been proposed. The motivation is to provide a formalism where mechanisms which give rise to complex surface variability can be modelled. The model shares the limitations of a conventional HMM with respect to modelling speech dynamics and its main interest is as an initial testbed for development of the necessary mathematical tools. In order to model speech dynamics in the intermediate representation it is necessary to extend the notion of a multiple-level model to a *segmental* HMM framework.

4. MULTIPLE-LEVEL SEGMENTAL HMMS

The states of a segmental HMM ([2, 3]) are associated with sequences of acoustic feature vectors, rather than with individual feature vectors as in a conventional HMM. This is motivated by the belief that by modelling such sequences as units it will be possible to capture and exploit speech segment dynamics. Formally an N-state segmental HMM comprises an N-state Markov model, plus:

- For each state σ_i, there exists a probability density function (pdf) b_i defined on the space of sequences of acoustic feature vectors from Y, such that *if* $y = y_1, \ldots, y_T$ is such a sequence, then $b_i(y)$ is the probability of y given the state σ_i. The pdf b_i is a probabilistic segment model.

A number of alternative types of segment model have been proposed [2]. The current development focuses on the 'probabilistic trajectory' model [2] for two reasons, namely that the use of an underlying trajectory provides an explicit model of dynamics, and that the probabilistic trajectory formalism appears to be particularly suitable for a multiple-level framework.

A trajectory of length N in the space $\mathcal{I}$ is a function $\tau: \{1, 2, \ldots, N\} \rightarrow \mathcal{I}$, where $\{1, 2, \ldots, N\}$ denotes the ordered set of integers 1 to N. Intuitively τ is derived from a continuous mapping of the interval $[0, 1]$ into $\mathcal{I}$. Since speech segments exhibit temporal variability, any segment model must accommodate variations in trajectory length. In some approaches (e.g. [11]) this is achieved by 'time-warping' a segment model of fixed duration. However, a more natural approach is to include a statistical model of

variation in duration directly in the model. This approach is described in [3] and is assumed in the analysis which follows. However, for notational simplicity the issue of trajectory length is ignored in the discussions which follows. With this in mind, let $\Gamma(\mathcal{G})$ denote the set of trajectories of length N in $\mathcal{G}$.

Returning to the arguments from the introduction, intuitively, the goal of a multiple-level probabilistic-trajectory segmental HMM is to associate states of the underlying Markov model with subsequences of vectors in the acoustic feature vector space Y via 'continuous' trajectories in the intermediate representation $\mathcal{G}$. The motivation is that by introducing an intermediate level $\mathcal{G}$ which is particularly suitable for modelling speech dynamics, one can explicitly characterise some of the mechanisms which give rise to surface variability and lessen the reliance on the assumption that this variability is random.

4.1 Relationship between the state space Σ and the intermediate space $\mathcal{G}$

The first step is to define the relationship between the state space Σ and the intermediate representation $\mathcal{G}$. This is achieved by identifying each state $\sigma \in \Sigma$ with a pdf b_σ defined on the set $\Gamma(\mathcal{G})$ of trajectories in $\mathcal{G}$. Note that, since the set of trajectories of length 1 in $\mathcal{G}$ is identical with $\mathcal{G}$, this is consistent with 3.2. By analogy with 3.2, this mapping will be denoted by $\varphi: \Sigma \rightarrow D(\Gamma(\mathcal{G}))$ and $\gamma: \Gamma(\mathcal{G}) \rightarrow D(\Sigma)$ is defined by $\gamma(\tau) = (b_{\sigma 1}(\tau), \ldots, b_{\sigma N}(\tau))$. In the case where $\Gamma(\mathcal{G})$ is a parametric set of trajectories, the pdf b_σ may be defined on the parameter space. For example, in [4] a pdf over a set of linear trajectories is defined by assuming that the trajectory parameters, the slope and mid-point value, are normally distributed.

4.2 Relationship between the intermediate representation $\mathcal{G}$ and the surface representation Y

At the surface level, a segment-based framework is concerned with evaluating the probabilities of sequences of elements of Y rather than individual elements. Thus, if Y^N denotes the set of sequences of length N in Y, then the mappings of interest are between $\Gamma(\mathcal{G})$ and $D(Y^N)$.

There are at least two approaches to defining the mapping $\eta: \Gamma(\mathcal{G}) \rightarrow D(Y^N)$. One is to attempt to map elements of $\Gamma(\mathcal{G})$ directly into $D(Y^N)$, while another is to derive this mapping from simpler relationships between individual elements of $\mathcal{G}$ and individual elements of $D(Y)$. The first approach would appear to apply tighter constraints to sequences in Y at a cost of increased mathematical complexity. However, if the underlying mapping from $\mathcal{G}$ to Y is continuous, then continuity restrictions on the shape of the trajectory will translate into constraints in Y. In other words, if the mapping $\eta'': \mathcal{G} \rightarrow D(Y)$ is continuous, then the mapping $\eta': \Gamma(\mathcal{G}) \rightarrow D(Y^N)$ defined by $\eta'(\tau) = (\eta''(\tau(1)), \ldots \eta''(\tau(N)))$, will map trajectory continuity constraints into the surface representation. Hence this simpler approach is adopted here. As in section 3.2, η'' can be interpreted as articulatory or formant synthesis, depending on the nature of $\mathcal{G}$, and variability in the acoustic realisation of an element ι of $\mathcal{G}$ can be accommodated by adding a random component. Thus $\eta: \Gamma(\mathcal{G}) \rightarrow D(Y^N)$ is characterised by $\eta: \iota \rightarrow \eta'(\iota) + w_\iota$, where w_ι denotes zero-mean Gaussian noise with covariance dependent on ι.

The 'inverse' mapping from Y^N to $D(\Gamma(\mathcal{G}))$ corresponds to the derivation of a distribution of paths in articulatory or formant space from a sequence of acoustic feature vectors. Traditionally, the problem of deriving production-based parameters from acoustic data is alleviated by considering sequences of such data, because it is then possible to apply continuity constraints [7]. However, continuity

constraints are implicit in the current framework. An attractive approach to the definition of $\xi: Y^N \rightarrow D(\Gamma(\vartheta))$, is to base ξ on a continuous 'pointwise' mapping $\xi': Y \rightarrow D(\vartheta)$ and to apply continuity constraints as restrictions on the types of trajectory which are included in $\Gamma(\vartheta)$.

4.3 Probability calculation in a multiple-level probabilistic trajectory segmental HMM

For any sequence of acoustic feature vectors $y \in Y^N$ and state $\sigma \in \Sigma$ it is necessary to evaluate the probability $P(y|\sigma)$. For any trajectory $\tau \in \Gamma(\vartheta)$, the joint probability $P(y, \tau|\sigma)$ is given by $P(y, \tau|\sigma) = P(y|\tau, \sigma)P(\tau|\sigma)$. As previously, $P(y|\sigma)$ is then given by,

$$P(y|\sigma) = \int_{\tau} P(y|\tau, \sigma)P(\tau|\sigma)$$

The probability $P(y|\sigma)$ is obtained directly from the pdf $\varphi(\sigma)$. The calculation of the probability $P(y|\tau, \sigma)$ is derived from the pdf $\eta(\tau)$. However, this derivation requires further analysis in the case where the mapping $\eta: \Gamma(\vartheta) \rightarrow D(Y^N)$ is based on a 'pointwise' mapping $\eta'': \vartheta \rightarrow D(Y)$, so that the trajectory τ is mapped to the sequence $\eta'(\tau) = (\eta''(\tau(1)), \ldots \eta''(\tau(N)))$ in $D(Y^N)$. If it is assumed that the elements of y are independent, then,

$$P(y|\tau, \sigma) = \prod_{n=1}^{N} P(y_n|\eta''(\tau(n)))$$

Clearly, this is an extension of the approach to the definition of segment probabilities presented in [3].

5. RELATIONSHIP WITH OTHER MODELS OF SPEECH DYNAMICS

The multiple-level segmental HMM described above is an extension of the probabilistic trajectory segmental HMM described in [3], and is motivated by the need to model speech dynamics in a separate production-based representation space. This final section briefly considers its relationship with other models of speech dynamics.

The fixed-trajectory segment model is the simplest form of trajectory-based model. For a given segment duration, a state of the underlying Markov process is identified with a single trajectory in the space Y. Polynomial [12], exponential [13] and a 'noiseless' version of the dynamical system [14] fixed trajectory model have been studied. The fixed-trajectory model is mathematically simple and extensible to a multi-level framework, but has the disadvantage that variations in the acoustic realisation of a speech segment must be modelled as random variants of the given trajectory. Indeed, it was found that the performance of the dynamical system model was improved by adding a random element to the underlying dynamics, thus making the model more similar to a probabilistic-trajectory model [14].

A small number of alternative multiple-level models have been studied. The model proposed by Bakis [15] includes underlying trajectories which are interpreted in articulatory terms, with the notion of 'articulatory effort' influencing movement between target articulatory configurations for each phoneme. Intuitively, in this model the trajectory dynamics are a compromise between trajectory inertia and target attraction and determination of the dynamics is a problem in constraint satisfaction. More recently, this

type of model has been studied by Bridle and Richards [16]. Multiple-level approaches based on modelling articulation have also been described by Ramsay and Deng [17] and by Deng [18], however these studies deal with the application of a specific model to the problem of speech recognition rather than the development of a general framework. An alternative approach to multiple-level modelling which includes an explicit characterisation of speech dynamics is recognition-by-synthesis, and in particular recognition-by-formant-synthesis-by-rule [19].

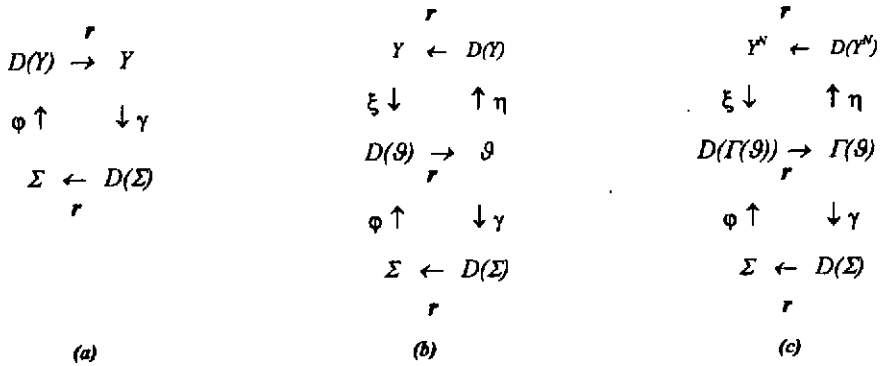

Figure 1: 'Mapping diagrams' for (a) a conventional HMM, (b) a three level HMM, and (c) a three level probabilistic-trajectory segmental HMM. The notation is as described in the text. The mapping 'r' denotes random generation of a sample from a given distribution.

6. CONCLUSIONS AND FUTURE WORK

The premise of this paper is that the capabilities of current segmental HMMs can be enhanced by modelling speech dynamics in an underlying production-related representation which lies between the symbolic state-level and surface acoustic representations. This paper has proposed a mathematical framework for describing such a multiple-level segmental HMM, so that its components can be identified and alternatives can be compared. The work is at an early stage, and significant problems remain. These include the demonstration that such models are useful from the perspectives of both speech science and computation, the development of a more complete mathematical theory, which must include algorithms for model parameter estimation and recognition, and proper testing.

Previous experience with probabilistic trajectory segmental HMMs has shown that even relatively small deviations from the conventional HMM formalism can lead to significant practical and theoretical issues which must be overcome in order to achieve improved speech recognition performance [18]. This suggests that a good understanding of the simpler models described in this paper, such as the multiple-level HMM, is needed before the more complex models can be successfully applied to the problem of automatic speech recognition.

7. REFERENCES

[1] STEVE YOUNG, "Large vocabulary continuous speech recognition: a review", Proceedings of the IEEE Automatic Speech Recognition Workshop, Snowbird, Utah, USA, pp 3-28 (1995)

[2] M OSTENDORF, V DIGALAKIS & O A KIMBALL, "From HMMS To Segment Models: A Unified View Of Stochastic Modelling For Speech Recognition", IEEE Transactions on Speech and Audio Processing, vol. 4, no. 5, pp 360-378.

[3] WENDY HOLMES & MARTIN RUSSELL, "Probabilistic Trajectory Segmental HMMs", to appear in Computer Speech and Language.

[4] MARTIN J RUSSELL AND WENDY J HOLMES, "Linear Trajectory Segmental HMMs", IEEE Signal Processing Letters, Vol 4, Number 3, March 1997

[5] M J RUSSELL, "Progress Towards Speech Models That Model Speech", Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara, pp 115-123, 1997.

[6] R K MOORE AND J S BRIDLE, "Speech Research at RSRE", Proc. Institute of Acoustics, Vol. 8: Part 7, pp. 480-483.

[7] J N HOLMES & W J HOLMES, "The use of Formants as acoustic features for automatic speech recognition"" Proc. Institute of Acoustics, vol. 18, part 9, pp 275-282, 1996.

[8] H RICHARDS, "The use of articulatory parameters for speech analysis", PhD Thesis, University of Swansea, 1997

[9] J W MÖLLER, B S ATAL & M R SCHROEDER, "Determination of articulatory parameters of the human vocal tract from acoustic measurements", J. Acoust. Soc. Am., 60, s77(A), 1976

[10] H RICHARDS, J S BRIDLE, M J HUNT & J S MASON, "Vocal tract shape trajectory estimation using MLP analysis-by-synthesis", Proc. IEEE ICASSP'97, pp 1287-1290, 1997.

[11] M OSTENDORF AND S ROUCOS, "A Stochastic Segment Model For Phoneme Based Continuous Speech Recognition", IEEE Trans. ASSP, Vol. ASSP-37, no. 12, pp 1857-1868, 1989.

[12] L DENG, "A generalised hidden Markov model with state-conditioned trend functions of time for the speech signal", Signal Processing, vol. 27, no. 1, pp 65-78, 1992.

[13] A WIEWIORKA AND D M BROOKES, "Exponential Interpolation Of States In A Hidden Markov Model", Proc. Institute of Acoustics, Vol. 18, Part 9, pp. 201-208, 1996.

[14] V DIGALAKIS, "Segment based stochastic models of spectral dynamics for continuous speech recognition", PhD Thesis, Boston University, 1992.

[15] R BAKIS, "Co-Articulation Modelling Within Continuous State Hmms", Proc. IEEE Workshop on Automatic Speech Recognition, Arden House, pp 20-21, 1991.

[16] J S BRIDLE & H RICHARDS, See website http://www.clsp.jhu.edu/ws98/projects/dynamic/presentations/hywel/presentation/index.html

[17] G RAMSAY & L DENG, "Maximum Likelihood Estimation for Articulatory Speech Recognition using a Stochastic Target Model", Proc. EUROSPEECH'95, Madrid, pp 1401-1404, 1995.

[18] L DENG, " A Dynamic Feature Based Approach to Speech Modelling and Recognition", Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara, 1997

[19] LAURIE MOYE, "Evaluation Of Speech Recognition By Synthesis", Proc. Institute of Acoustics, Vol 14: Part 6, pp 87-94 (1992).