PRESERVING FINE PHONETIC DETAIL USING EPISODIC MEMORY: AUTOMATIC SPEECH RECOGNITION WITH MINERVA2

Roger K. Moore and Viktoria Maier

Dept. Computer Science, University of Sheffield, UK
r.k.moore@dcs.shef.ac.uk, v.maier@dcs.shef.ac.uk

ABSTRACT

Previous research has demonstrated competitive recognition results using a simulation of episodic memory - 'MINERVA2' - on the Peterson & Barney corpus of vowel formant data. This paper presents a modified implementation designed to work on real speech data, and results are reported on isolated-word recognition experiments conducted using the TI-ALPHA corpus. It is shown that access to fine phonetic detail is critical for achieving high recognition accuracy, whether it is provided by the episodic model or by hidden Markov models incorporating large numbers of Gaussian mixture components. However it is confirmed that, although MINERVA2 offers a powerful means for generalizing by accessing the fine detail retained in *all* the training data, it is severely hampered by its inability to model temporal sequence. It is concluded that a new episodic model is needed that is based on the principles of MINERVA2 but which overcomes such limitations.

Keywords: episodic memory, exemplar-based ASR, MINERVA2.

1. INTRODUCTION

Automatic speech recognition (ASR) is a field of research that has matured over a period of more than half of a century. In that time many different approaches have been investigated and contemporary large-vocabulary continuous-speech recognition (LVCSR) systems represent a considerable improvement over the first isolated-word recognisers (IWR). It can be argued that it has been the introduction of hidden Markov models (HMMs) in the 1980s that has been the main catalyst for these improvements. However, HMMs are not without their shortcomings; many assumptions are made about the nature and structure of speech signals, and a number of these are patently false. Nevertheless, the advantages of using probabilities and statistics to model our lack of knowledge [12] about the detailed structure and dependencies in speech, currently far outweigh the disadvantages arising from poor approximations to reality.

However, it has become apparent that the performance of current state-of-the-art ASR systems is in danger of asymptoting to a level of recognition accuracy that falls significantly short of that which is required to support many advanced applications [14] let alone being comparable with the capabilities of a human listener [8]. As a consequence, a number of researchers are exploring the field of human speech recognition (HSR) in order to better understand the nature of speech, and to investigate the possibility that a simulation of the human speech recognition system might lead to more competitive and robust ASR [13].

In particular, there is growing interest in the possible implications of 'episodic' memory for perceptual tasks, and a number of HSR researchers are investigating an 'exemplar based' approach [1][2][3][4][6][9][16]. The main reason for this rise in interest is that the flexibility of HSR is not able to be modelled adequately with an architecture relying on pre-abstracted representations. An exemplar-based approach offers a mechanism for retaining and accessing the 'fine phonetic detail' [4] that would be discarded in purely abstract representations (such as HMMs).

In a previous paper [10] the authors presented a vowel recognition system based on Hintzman's computational multiple-trace (episodic) memory model known as 'MINERVA2' [5]. The episodic system performed very well in comparison to conventional pattern classifiers such as a support-vector-machine (SVM), a Gaussian mixture model (GMM) and a k-nearest-neighbour classifier. This study was conducted on the Peterson & Barney vowel formant data [15], but it was a relatively simple speech-related task since it involved *no*

temporal information. This paper presents new results based on a comparison between a modified version of MINERVA2 and a standard HMM classifier using the TI-ALPHA isolated-word database [7].

2. MINERVA2

MINERVA2 simulates episodic memory by first storing 'traces' (records of individual memory experiences or episodes). Inputs to the system - 'probes' - are compared to *all* of the traces in memory, and the retrieved 'echo' (essentially a weighted composite of the stored traces) returns a vector containing additional knowledge that is unspecified in the input, e.g. its class. The weights are determined by the similarity between the input and each stored trace. Hintzman [5] showed that such a model is able to create abstract representations of stored data, and that by probing repetitively with the abstracted representations (a process referred to as 'echoes of echoes'), it is possible to refine the response and exploit the implicit relationships between individual stored traces.

The main parameters of the model are (i) the feature representations, (ii) the similarity function, (iii) the weighting function (also called the activation function) and (iv) the echo retrieval function.

2.1. Model parameters

2.1.1. Feature representations

In the implementation discussed here, the feature vector consisted of the standard representation used in ASR tasks – mel frequency cepstral coefficients (MFCCs) and their derivatives. The class labels (i.e. the identities of the lexical items) are stored as blocks of features in the same way as outlined in [10].

2.1.2. Similarity measure

As in our previous work, the similarity between the input and stored traces has to be computed using an intermediate step that is different to Hintzman's original binary approach. In our implementation, the distance measure used is the Euclidean Distance (*ED*):

$$ED_{I,t} = \sqrt{\sum_{i=1}^{n} \left(| I_i - t_i | \right)^2} \qquad (2.1)$$

… where I_i is the i^{th} feature of the input vector and t_i is the i^{th} feature of the trace t.

The similarity between the input I and the trace t is then computed by:

$$sim_{I,t} = 1 - \left(ED_{I,t} / \max\left(ED_{I,t} \right) \right) \qquad (2.2)$$

… where $ED_{I,t}$ is the vector of length n, with n equal to the number of features, and $max(ED_I)$ is the maximum value in the vector. It is necessary to normalize *ED* in order to ensure that the range of $sim_{I,t}$ is between 0 and 1.

Due to the non-binary nature of the features in this adapted version of MINERVA2, the similarity function is in fact calculated across the standard feature set only, and does not include the class label features. This is necessary because, although the Euclidean Distance (ED) across the class label features is a static value containing no useful information for the comparison of a new input with the traces in memory, it does change the derived similarity values. In particular, because of the normalisation of the similarity values, the bigger the constant value of the class label features that goes into the ED, the smaller the range that the derived values will occupy in the similarity range {0 - 1}. In practice this means that this static value would counter the effect of the power factor (defined in section 2.1.3 below).

2.1.3. Activation function

To gain the final weighting w of the traces with respect to input I, the similarity measure is raised to the power of p. As Hintzman noted [5], this in effect gives more weight to the most similar traces and less to those traces that are not similar.

$$w_{I,t} = sim_{I,t}{}^{p} \qquad (2.3)$$

Hintzman sets the value of the power factor p to three. He states, however, that other values are permissible as long as the sign of $sim_{I,t}$ is retained. Thus in Hintzman's MINERVA2, p is restricted to odd values. However, this restriction is only necessary if negative similarity measures are possible. Since the similarity measure in the adapted version is always positive, p is permitted to have any value, odd or even.

2.1.4. Echo Intensity

Echo intensity is a measure of how much activation has been triggered. The more traces that match the input, and the more similar they are to the input, the greater the value of I. Echo intensity

can be used to judge frequency and familiarity; it is defined as follows:

$$\text{int}_I = \sum_{t=1}^{T} w_{I,t} \quad (2.4)$$

…where I is the input, T is the total number of traces stored.

2.1.5. Echo retrieval

The echo is the derived abstraction of the stored traces as a response to the input. This is accomplished by computing a weighted sum of all traces in memory. The echo then becomes:

$$echo_I = \left(\sum_{t=1}^{T} w_{I,t} \cdot trace_t \right) / \text{int}_I \quad (2.5)$$

… where $w_{I,t}$ is the weight on trace t for input I, and T corresponds to the number of stored traces. Note that in our adapted approach, a normalisation of this value is necessary for numeric reasons.

2.2. Handling multiple frames

MINERVA2 is essentially a single-frame classifier; hence moving from the Peterson & Barney vowel data to an isolated word corpus requires the addition of a mechanism for handling variable-length tokens. However, such a step constitutes a fundamental change in the underlying methodology. Handling temporal sequence appropriately requires the derivation of an entirely new temporal episodic memory model, and this is the subject of ongoing research [11].

Prior to the development of a fully functional temporal episodic model, several intermediate solutions present themselves. In this study, a 'bag-of-frames' (BoF) approach was adopted as the configuration that involves the least number of assumptions about the temporal evolution of speech patterns. BoF simply means that a word is classified according to the accumulated response of all of its constituent frames *regardless of the order in which they occurred*. In the long term, this is unlikely to represent a realistic configuration; however in the short term, it allows whole-word recognition experiments to be conducted using the adapted version of MINERVA2.

$$bagsClass = \arg\max_{W \in C} \left(\sum_{n=1}^{N} echoClassesVals_n \right) \quad (2.6)$$

… where *bagsClass* is the class that is attributed to the whole 'bag of frames' constituting an utterance, W is a class from the set of all classes C, n is the index of which frame of the utterance, and *echoClassesVals* are the values that the echo returns for all possible classes.

3. EXPERIMENTS AND RESULTS

The database chosen for this investigation was the TI-ALPHA isolated word corpus. The data consists of 16 speakers (eight male and eight female) uttering the 26 letters of the US English orthographic alphabet ("A", "B", "C", etc.). The complete test set consists of 6628 utterances, and the complete training set consists of 4142 utterances. This standard ASR database was chosen because of (i) the high confusability of the vocabulary, hence its high sensitivity to alternative recognition approaches, and (ii) its relatively small size, hence allowing manageable recognition experiments. Although primarily designed to be used for multi-speaker (MS) experiments, it was also possible to partition the data for speaker-independent (SI) tests.

All experiments were conducted using standard MFCC features and their first and second derivatives, giving rise to a total of 39 features per frame. A 25ms frame was taken every 10ms. The classes corresponded to whole-word labels.

Results were also obtained using a standard whole-word HMM baseline that employed left-to-right HMMs with three emitting states per model. A further HMM model was trained with only one emitting state in order to emulate the same 'temporally-invariant' model as in the BoF scheme. All HMM models were trained by incremental mixture splitting. The number of components per mixture was optimized for best performance. All references to the number of states in an HMM refer to emitting states only.

3.1. MINERVA2 vs. single-state HMMs

In statistical pattern recognition, the process of generalization is achieved by combining information during training. For example, in state-of-the-art classifiers such as HMMs or GMMs, training data is used to find the mean and variance of a single- or multi-component Gaussian mixture distribution of the data. In direct comparison, MINERVA2 does something very similar – it also computes the mean of similarity-weighted data; however, there is no overall mean as the similarity weighting attempts to substitute a general distribution for one that best fits the current input. Hence, MINERVA2 models the various classes to be

expressed using only one value per feature. The consequence is that the use of such similarity-weighted training data allows the constructed models to take into account the fine-phonetic similarity found within a frame.

Therefore, the first hypothesis to be tested is as follows: does the use of similarity-weighted training data enhance the model's recognition performance using the minimum number of model parameters? If so, then one would expect that MINERVA2 would outperform a one-state single-Gaussian HMM, if it makes sense to take the similarity of such fine details into account. As can be seen from the results shown in Table 1, MINERVA2 clearly outperforms the single-state HMM. (Note that error rate differences of more than 0.5% are statistically significant.)

Table 1: Comparison between a single-Gaussian and MINERVA2 model. Multi- Speaker (MS) and Speaker-Independent (SI) recognition results..

Classifier	Error Rate
MS:HMM S1 (single-Gaussian)	35.4 %
MS: Episodic Model (p=29)	11.3 %
SI:HMM S1 (single-Gaussian)	40.0 %
SI: Episodic Model (*p=29*)	27.5 %

3.2. MINERVA2 vs. multiple-state HMMs

HMMs typically use GMMs (rather than single Gaussians) in order to allow data belonging to one class to be modelled using different distributions. In effect, the training data is split up and clustered during training to a previously defined number of Gaussian distributions. This means that in the subsequent testing stage, partially clustered training data is compared to the unknown input. However, in direct contrast, MINERVA2 is based on the assumption that an online comparison of the input data to *all* of the training data leads to a more appropriate weighting of the information, and this may offer an advantage in recognition accuracy.

However, the standard MINERVA2 architecture has an inbuilt disadvantage with respect to multiple-state models such as HMMs in that it can not flexibly model temporal information. Nevertheless, it is interesting to find out just how well/badly MINERVA2 would perform in comparison to HMMs using GMMs and/or multiple states.

The first experiments were run on the complete test- and training data in multi-speaker mode, and the results are presented in Table 2. As expected, the best recognition performance was obtained using the three-state-HMM with 120 Gaussians per state.

Table 2: Multi-speaker recognition results. S1 (S3): HMM with one (three) emitting states.

Classifier	Error Rate
HMM S3 (120 GMM)	3.9 %
HMM S3 (60 GMM)	4.1 %
HMM S3 (1 GMM)	29.0 %
HMM S1 (300 GMM)	4.2 %
HMM S1 (3 GMM)	49.4 %
Episodic Model (p=29)	11.3 %
Episodic Model (p=61)	10.6 %

A speaker-independent training set was established for each test utterance by removing the respective test speaker's utterances from the training data. This meant that there was about 6% less training data for the speaker-independent tests in comparison with the multi-speaker condition. Table 3 presents the speaker-independent results.

Table 3: Speaker-independent recognition results. S1 (S3): HMM with one (three) emitting states.

Classifier	Error Rate
HMM S3 (30/60 GMM)	11.7 %
HMM S3 (1 GMM)	33.4 %
HMM S1 (60 GMM)	11.9 %
HMM S1 (3 GMM)	52.6 %
Episodic Model (p=29)	27.5 %

As can be seen from the results listed in both Tables 2 and 3, the addition of GMMs to both one-state and three-state HMMs gives rise to better performance than obtained using MINERVA2. Even the one-state HMM/GMM (which does not have the advantage of being able to model temporal information) was able to achieve better recognition results than the episodic model. However, it is noticeable that the number of Gaussians needed per state in the HMMs for optimal performance is rather high, (given that there are only about 160 utterances per class for training in the MS condition and about 150 utterances in the SI condition). This suggests that the individual Gaussians in the mixture are less generalized than the echo response of MINERVA2, and hence the decision is based on even less information than the echo acquired by MINERVA2.

Another interesting observation is that the single-Gaussian three-state-HMM does not outperform MINERVA2, despite the HMM's advantage in having some capacity to model temporal information. In fact, in the multi-speaker

condition (Table 2), the recognition performance obtained using three-state HMMs is only slightly better than that obtained using one-state-HMMs (3.9% vs. 4.2%).

In order to assess whether this indicates limited use of temporal information, it is necessary to compare single-Gaussian three-state HMMs with one-state HMMs which have an equal total number of Gaussians (i.e. three-state HMMs with a one-component GMM vs. one-state HMMs with three-component GMMs). The results in Tables 2 and 3 indicate that the one-state HMMs lead to significantly worse recognition results, hence confirming (as one would expect) that temporal information is indeed a key feature for accurate speech recognition.

On the other hand, a further comparison of the slight differences in recognition performance between the one- and three-state HMMs with a high number of Gaussians per state (i.e. S1-300GMM vs. S3-120GMM in Table 2) suggests that the modelling of fine spectral detail in the speech signal carries more significance for good classification than some limited ability to model the temporal evolution of the patterns.

3.3. MINERVA2's sensitivity to training data

Additional experiments were conducted to investigate the dependency of MINERVA2 on the type and amount of training material. The test and training data used for these experiments were relatively small subsets of the complete test and training set, however they were selected to cover the complete set of classes (i.e. the whole alphabet). The following experiments were conducted:

- **Experiment A (multiple-speaker)**
  Train: 208 utterances, two speakers (M1, F8)
  Test: 194 utterances, all 16 speakers
- **Experiment B (multiple-speaker)**
  Train: 208 utterances, *four* speakers (M7 - F2)
  Test: 194 utterances, all 16 speakers
- **Experiment C (multiple-speaker)**
  Train: 416 utterances, four speakers (M7 - F2)
  Test: 194 utterances, all 16 speakers
- **Experiment D (speaker-independent)**
  Train: 416 utterances, four speakers (M7 - F2)
  Test: ~194 utterances, all 16 speakers

Table 4 illustrates the results for these different experimental configurations (p=11):

Table 4: Results of experiments A-D

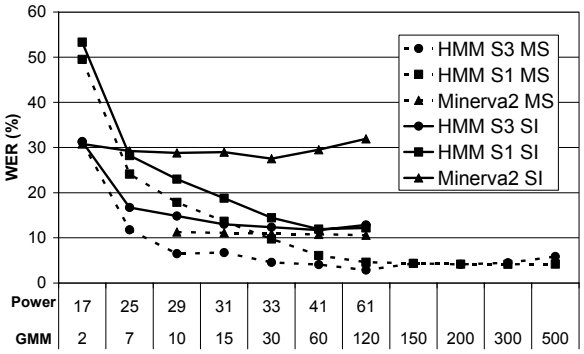
Classifier	Error Rate
A	42.3 %
B	48.0 %
C	29.9 %
D	45.8 %

These results seem to indicate that the number of training utterances may be more important than the number of training speakers that are in the system; the performance improved considerably by adding more training tokens, but it stayed approximately the same when tokens were replaced using examples from another speaker.

3.4. Parameter dependency

A further observation was that the HMM and MINERVA2 models behaved with different sensitivity to changes in the values of the free parameters (i.e. the number of Gaussians per state for the HMMs, and the value of *p* for MINERVA2). The results of experiments varying these parameters are shown in Figure 1.

Figure 1: Word error rates (WER %) for multiple-speaker (MS) and speaker-independent (SI) HMM and MINERVA2 conditions as a function of power value/GMM components.



The results presented in Figure 1 suggest that MINERVA2 is *less* dependent on its parameter settings than the HMMs. Further, although not yet tested, there is reason to believe that it should be possible to automate the setting of the free variables in MINERVA2 by relating them to the size of the frames per class and number of total classes. In contrast, even after extensive research into HMM-based ASR, finding the correct topology for the models is still a matter of hand tuning on an evaluation set.

4. DISCUSSION

The results presented in this paper provide support for two main conclusions. First, the comparison between MINERVA2 and single-Gaussian classifiers indicates that access to fine-phonetic information can indeed improve the performance of automatic speech recognition. This is supported by the fact that a high number of Gaussian mixture components have the best recognition results. Second, the results have shown that the MINERVA2 model is not able to perform as well as state-of-the-art HMM classifiers on the TI-ALPHA data, primarily due to its lack of temporal structure.

Therefore, the overall conclusion would seem to be that HMMs are not only able to retain fine-phonetic information, but they seem to be able to use it better than the current MINERVA2 model. However, MINERVA2 still holds a big advantage over the HMM models – it retains *all* the information present in the training data. An HMM, even when using high numbers of Gaussians per mixture, will always have to sacrifice some of the fine detail in frequency (by using fewer Gaussians than examples) and in time (by using fewer states than frames) in order to be able to generalize. Once lost, this information is lost forever in an HMM.

There are two very obvious shortcomings in MINERVA2 which probably cause the current implementation to be inferior to HMMs. The first is the combination of *unweighted* Euclidean Distance with MFCC features in the modified MINERVA2 model. In a GMM, each feature in a feature vector is modelled independently in a separate dimension, and hence, if the variances of some of the features are very different to others, there is no direct impact on the other dimensions; differences are automatically scaled by the variances. However, the unweighted Euclidean Distance in the modified version of MINERVA2 does not compensate if some of the features have a larger range of values than others. For example, if the range of values for one MFCC feature is ten times larger than another, then it would carry more importance for the computation of the similarity between trace and probe. Thus, either different (or normalised) features have to be used, or the model has to be able to account for such differences to ensure that all features are given equal weight. Some initial experiments with normalized features seem to confirm this weakness.

Of course the main disadvantage of MINERVA2 compared with HMMs is that while the HMM has the *ability* to model temporal sequence, MINERVA2 cannot. This arises from the fact that the system echo is not attributable to any particular trace, but is rather a collected response from *all* traces in memory independent of the sequence that they occurred in. Two solutions present themselves:

1. The sequence of frames should be encoded into the traces by using a multi-frame context window.
2. The sequence should be superimposed on the front-end features in a manner that is similar (but more explicit) to the derivatives used with MFCC features.

The first option was tried on the same data as in experiment B. By increasing the context from one to 20 frames, a steady gain in performance was achieved, and for a 20-frame context the error rate decreased by 5.7% to 42.3%.

However, neither of these solutions treats sequence as an elementary property of the recognition process. Hardwiring context information into the features means that the temporal dimension of the model is fixed and the model would only compare corresponding features with each other. For example, this would mean that speaking rate or other aspects of temporal dynamics would be fixed in the trace and could rule out traces that are very similar but have different timing structures. Instead, an adequate model for speech recognition should be able to use the information of sequence directly, and thus allow for dynamics in all dimensions. Disregarding sequence means that key information available in each episode is not exploited by the MINERVA2 model as it stands.

5. CONCLUSION

This paper reports new results for an adapted computational multiple-trace memory model known as MINERVA2 on an isolated word recognition task. Unlike the previously reported results (which were on parameterised vowel data), the experiments reported here used real speech data encoded using standard MFCC features (and their time derivatives). In order to introduce the least number of assumptions about the temporal evolution of speech patterns, a 'bag-of-frames' approach was introduced to handle multiple-frame

utterances. The recognition results have been compared with standard HMM classifiers.

Unlike the previously reported results on the Peterson and Barney database, MINERVA2 shows a clear lack of performance on isolated words compared with that obtained using HMMs, and there are reasons to believe that this could result from the combination of MFCC features with the Euclidean Distance used in MINERVA2. However, the main shortcoming of the model is its inability to incorporate sequence information. This weakness can be overcome to some extent if sequence is encoded as part of the features supplied to the system. Unfortunately this approach means that there is no flexibility for comparing temporal information between probe and trace by the system, which is a definite weakness of MINERVA2. Further development of the MINERVA2 model is needed in order to overcome such constraints while at the same time maintaining its positive features and preserving its appealing simplicity.

Nevertheless, the value of using fine-phonetic information seems to have been confirmed by the experiment reported here. HMMs are obliged to average over such information in order to generalize, and the fine detail is inadvertently lost in the process. On the other hand, an episodic model such as MINERVA2 is not only able to generalize by using all the training data (in contrast to HMMs which typically use the data from one class at a time), but it does so in a way that exploits the detail by retaining the original data.

In conclusion, despite the relative simplicity of the experiments reported here, the results confirm that the retention of fine phonetic detail is important to accurate speech recognition, and that episodic modelling offers a compelling alternative to HMMs as a means for exploiting such detail. However, it is also clear that MINERVA2 is severely limited in its ability to exploit temporal sequence information. Hence a new episodic model is required that is based on the principles of MINERVA2 but which overcomes such limitations. Such a model is currently the subject of ongoing research [11] and early results suggest that this new 'temporal episodic memory model' (TEMM) is capable of outperforming HMMs on a speech recognition task.

6. REFERENCES

[1] Bybee, J. 2001. *Phonology and Language Use*. Cambridge University Press.

[2] Goldinger, S.D. 1996. Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition* 22: 1166-1183.

[3] Goldinger, S.D. 1998. Echoes of echoes: an episodic theory of lexical access. *Psychol. Rev.* 105, 251-79.

[4] Hawkins, S., Smith, R. 2001. Polysp: a polysystemic, phonetically-rich approach to speech understanding. *Italian Journal of Linguistics - Rivista di Linguistica*, 13, 99-188.

[5] Hintzman, D.L. 1986. Schema-abstraction in a multiple-trace memory model. *Psychological Review* 93, 411-427.

[6] Kirchner, R. 2004. Exemplar-based phonology: it's about time. *Proc. 23rd West Coast Conference on Formal Linguistics*.

[7] Liberman et al. 1993. TI46-Word, *LDC Catalog No.* LDC 93S9.

[8] Lippmann, R. 1997. Speech recognition by machines and humans. *Speech Communication* 22, 1-15.

[9] Luce, P.A., Lyons, E.A. 1998. Specificity of memory representations for spoken words. *Memory and Cognition* 26, 708-715.

[10] Maier, V., Moore, R.K. 2005. An investigation into a simulation of episodic memory for automatic speech recognition. *Proc. InterSpeech* Lisbon, 1245-1248. (http://www.dcs.shef.ac.uk/~roger//publications/Maier Moore Interspeech 2005 revised.pdf)

[11] Maier, V., Moore, R.K. 2007. Temporal episodic memory model: an evolution of MINERVA2. Submitted to: *InterSpeech*, Antwerp.

[12] Makhoul, J., Schwartz, R. 1986. Ignorance modeling. In: *Invariance and Variability in Speech Processes*. J. Perkell, D.H. Klatt (Eds.), Erlbaum.

[13] Moore, R.K., Cutler, A. 2001. Constraints on theories of human vs. machine recognition of speech. *Proc. SPRAAC Workshop on Human Speech Recognition as Pattern Classification*, Max-Planck Institute for Psycholinguistics, Nijmegen, 145-150.

[14] Moore R.K. 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. *Proc. Eurospeech* Geneva. 2582-2584.

[15] Peterson, G.E., Barney, H.L. 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175-184.

[16] Tulving, E. 2002. Episodic memory: from mind to brain. *Annu. Rev. Psychol*. 53, 1–25.