Interspeech 2018
2-6 September 2018, Hyderabad

Implementing DIANA to model isolated auditory word recognition in English

Filip Nenadić[1], *Louis ten Bosch*[2], *Benjamin V. Tucker*[1]

[1]University of Alberta
[2]Radboud University Nijmegen
{nenadic, bvtucker}@ualberta.ca, l.tenbosch@let.ru.nl

Abstract

DIANA, an end-to-end computational model of spoken word recognition, was previously used to simulate auditory lexical decision experiments in Dutch. A single test conducted for North American English showed promising results as well. However, this simulation used a relatively small amount of data collected in the pilot phase of the Massive Auditory Lexical Decision (MALD) project. Additionally, already existing acoustic models were implemented. In this paper, we expand the analysis of MALD data by including a larger sample of both stimuli and participants. Acknowledging that most speech humans hear is conversational speech, we also test new acoustic models created using spontaneous speech corpora. Simulations successfully replicate expected trends in word competition and show plausible competitors as the signal unfolds, but acoustic model accuracy should be improved. Despite the number of responses per word being relatively small (never more than five), correlations between model estimates and participants' responses are moderate. Future directions in acoustic model training and simulating MALD data are discussed.

Index Terms: spoken word recognition, computational modelling, DIANA, auditory lexical decision, reaction times

1. Introduction

The last three decades of research on speech perception have been marked by the development of various models of spoken word recognition. Some good overviews of these models and the ways in which they can be compared or tested are given in [1, 2, 3, 4, 5]. In this paper we use DIANA [6], a recently developed computational model of spoken word recognition, to model responses collected as part of The Massive Auditory Lexical Decision (MALD) project [7]. DIANA resembles its predecessors in many regards: it is an activation and competition model, based mostly on bottom-up (phonetic) information as input (but potentially utilizing top-down information as well) for choosing the best candidate from a set of options stored in the lexicon. However, unlike most previous models such as TRACE [8, 9] or Shortlist [10, 11], DIANA does not require manually created lower-level abstract units (e.g., phonemes), instead using the acoustic signal itself as input. Additionally, even in comparison to models that use acoustic signal as input, such as Fine-Tracker [12, 13], DIANA also offers word/pseudoword decisions and estimates response latencies, comparable to those obtained in behavioral experiments.

DIANA has three components: the activation, the decision, and the execution components, as shown in Fig 1. The activation and decision components operate in parallel. The activation component analyzes the acoustic input by converting it into vectors of Mel-Frequency Cepstral Coefficients (MFCC). The acoustic characteristics of every phone (sub-lexical units used in our current setup) in the lexicon are represented by Gaussian

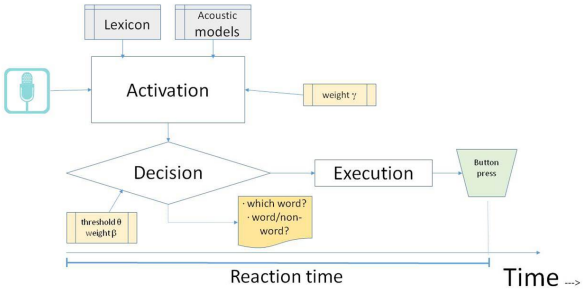

Figure 1: *The process of spoken word recognition as assumed by DIANA.*

mixture models specifying the distribution of MFCC vectors for three states in a hidden Markov model that each phone has. The matching is performed using a Bayesian framework, and calculated for every ten milliseconds of input. A controllable parameter determines the impact of bottom-up (acoustic) versus the top-down (prior probability i.e., frequency) information in selection of the winning candidate. The decision component selects the winner based on another controllable parameter determining the desired difference between the activation of the top candidate and the best runner-up. In the case that this difference is not attained at stimulus offset, yet another controllable parameter estimates the added time for the final winner decision, also potentially taking into account the number of remaining candidates. Finally, the execution component represents the time taken to actually perform the task of responding to a stimulus (i.e., pressing a button), and it is ordinarily fixed to e.g., 200 milliseconds.

Although it was used for modelling participant behavior in other tasks as well [14, 6], DIANA has predominantly been used to simulate auditory lexical decision responses in several studies, almost exclusively in Dutch. For example, DIANA was used to model responses to 613 disyllabic monomorphemic Dutch words made by 20 participants [15]. The error rates of the simulation were quite similar to those of actual participants, being 4% word and 7% non-word misclassification, whereas human subjects had 6% and 5% error rate respectively. Additionally, average correlation between model estimates and human subjects response latencies was $r = .47$, whereas the correlations in response latencies between subjects were ranging from $r = .1$ to $r = .3$, indicating that the model represents general (or average) tendencies of human subjects well.

Further implementations of DIANA tackled other issues, such as modelling the tendency of response latencies to a stimulus to correlate with response latencies to a number of previous stimuli [16]. These local speed effects (such as learning

10.21437/Interspeech.2018-2081

or fatigue [17]) affect response latency to a particular stimulus in addition to the underlying processing mechanism shared by all participants, long term effects (such as e.g., age or general cognitive abilities), and the deliberate participant strategy taken in the experiment. By applying a filter that takes into account local speed effects, the correlation between participants themselves and the correlation of DIANA to the average participant was shown to increase. Yet another procedure that took into account word frequency, required larger differences in activation between the winner and the runner-ups, and added extra choice time if there is a close competitor at the word offset, further increased the average correlation between DIANA's estimates and actual participant response latencies to $r = .76$ [18].

Even though DIANA is language-independent, it was implemented outside of Dutch only once, using pilot data from MALD [18]. This dataset included a total of 1200 word types with responses from 10 to 12 participants out of the 250 MALD pilot sessions. The participant sample was heterogeneous as it included native and non-native speakers of English. The results of the simulation still showed satisfactory performance of DIANA, with correlations with the average participant being $r = .45$.

1.1. The present study

The goal of the present study is to expand on the previous application of DIANA to English [18]. Our first aim is to develop acoustic models of Western Canadian English using spontaneous speech as a basis. We opted for spontaneous speech as our training material for three reasons. First, acoustic models are usually trained using careful speech corpora such as TIMIT [19], and we intend to compare models trained on these different types of corpora at a later time. Second, the speaker that recorded the MALD items speaks a Western Canadian variety of English, same as the speakers in the spontaneous speech corpora we used for training, and the participants in the behavioral experiment. Third, spontaneous or conversational speech is what listeners are most often exposed to [20], rather than careful speech – we deemed using spontaneous speech would better represent the kind of 'training' actual human listeners receive.

Our second aim was to test model performance when recognizing novel isolated word recordings and when simulating between-word competition as a function of time. Importantly, we also wanted to simulate the lexical decision task, observing both lexical decisions made and estimates of response latencies, and how they compare to participant performance in MALD.

2. Methods

2.1. Behavioral experiment

The auditory lexical decision experiment included 26,793 words and 9,592 pseudowords generated using Wuggy [21], which was adapted by the authors to create a phonetic database. We used the CMU Pronouncing Dictionary V0.6 [22] for pronunciation referencing, expanded for words which were missing entries. The stimuli were recorded by a single 28-year-old male speaker of Western Canadian English. The words were then split into 67 separate lists each containing 400 words, and paired with one of the 24 lists of 400 pseudowords each.

The participant pool included 231 monolingual native Canadian English listeners (180 female, 51 male, aged 17 to 29), forming a more homogeneous sample than in [18]. The participants were allowed to participate up to three times, never listening to the same list of words or pseudowords. Most participants only completed a single list, and the total number of MALD sessions was 284. Altogether, the database consisted of 227,179 participant responses.

A single session of the auditory lexical decision task contained 400 words and 400 pseudowords presented in random order. Each trial was initiated by a 500ms fixation point, followed by the sound stimulus presented over the headphones. Participants could respond during stimulus presentation and their response time was limited to 3000ms, after which the experiment would proceed to the next stimulus. Participants were instructed to use their dominant hand to respond to words, and their non-dominant hand to respond to pseudowords.

2.2. Model training

The training procedure was performed by automatic speech recognition training using Hidden Markov Model Toolkit (HTK; [23]). It was conducted in three steps, and at every step the training was performed in three iterations of re-estimation. We used two unpublished spontaneous speech corpora to create the initial models. The Western Canadian English spontaneous speech corpus (WCE) includes recordings of 11 subjects making telephone calls, and the Corpus of Spontaneous Multimodal Interactive Language (CoSMIL) contains recordings of conversations of 8 pairs of participants (16 participants total). The recordings were separated into sounds shorter than 10 seconds by their existing transcribed interval. Intervals longer than 10 seconds were manually split into two approximately equal intervals at the middle of a silent pause in speech. Intervals consisting of only silent pauses, laughter, breathing etc. were excluded. The number of separate sounds created in this manner was 20,086, with a total duration of just over nine hours. These sounds were downsampled to 16 kHz (see [23]). Despite procedures applied to avoid sound clipping, a small number of sounds (31) resulted in a warning, and these sounds were excluded. Training creates estimates for all sub-lexical units (in this case phones) as three-state hidden Markov models (HMMs), with their acoustic characteristics represented by Gaussian mixture models (GMMs). Since the training sounds included stretches of speech with two or more connected words, the models also need to correctly account for short pauses between them. The acoustic models generated on the spontaneous speech recordings were therefore further extended to include estimates for short pauses (forming the so-called 'sp models').

At this point every three-state HMM (i.e., phone) is represented by a single GMM. Increasing the number of GMMs per state has been shown to reliably reduce error rate when models are used for word recognition [24]. The second step in model training was then to increase the number of Gaussian mixtures per phoneme state to the usually recommended 32. This was done by doubling the number of states at each step — from 1 to 2, 4, 8, 16, and finally 32.

The final part of the training procedure was speaker adaptation, with actual recordings of the speaker the model will be used on implemented to realign the acoustic model estimates. We wanted to test how many recordings are minimally required to create adequate acoustic models, which would allow us to use more of the recordings in tests and simulations. We created separate models differing in the number of MALD word (not pseudoword) recordings used for adaptation. The model that was adapted on the smallest number of MALD words used only MALD list number 1 (400 words). The remaining 19 models each used one additional list with 400 words, with the model adapted on the largest number of MALD words adapted on lists

1 to 20 (8,000 words). Each list included just under 4 minutes of speech.

2.3. Simulations

The most common test for acoustic models is whether they can successfully recognize words stored in the lexicon in novel speech recordings (free word recognition). The procedure assigns an activation value to each word in the lexicon based on the probability of a certain word given the signal. In this case, our simulation used only bottom-up acoustic information. The considered lexicon comprised of all 26,793 MALD words. The models were tested on a total of 1,200 words from three MALD lists that were not used in the training phase (lists 65, 66, and 67). The best model (one adapted on 4,000 words, see Section 3) was used in all further simulations described below.

We observed the top three competitors besides the winner (by creating the so-called N-best lists, in this case it is a 4-best list), based on their activation strength. The number of competitors was selected arbitrarily and served to assert model plausibility beyond the winning word — misidentification might still include the target word as highly activated, and top competitors for correctly identified words should be similar to the winner. However, since we know speech unfolds over time, sounds were also split into 20ms frames, and an estimate was made upon addition of every new frame, with 10 top competitors being considered. Due to processing such gated sounds being demanding, we only considered a subset of the CMU lexicon, including all the words that have three phonemes or less and all the words that share the first three phonemes with the target word. We observed both changes in the estimated phone string and at top four competitors as the signal was unfolding.

We also performed a lexical decision simulation on the three lists and their corresponding pseudoword lists by comparing the best lexical activation and the activation obtained in a free sub-lexical unit (phone) loop. In the free phone loop, the language model (grammar) does not include words at all, only phones, and, optionally, probabilities from moving from one to the other. In this case, we created the so-called flat phone loop model (also called zerograms), as all possible transitions from one phoneme to the other were treated as equally possible. If a sound signal activates a free phone string not in the lexicon significantly better than it does for any lexical entry (the threshold is set by the researcher), the conclusion is made that the sound signal is comprised out of a string of segments that do not match any lexical entry well, making it a non-existing word i.e., a pseudoword. Conversely, when an actual word is given as input to the model, the difference between the free phone loop activation and the best lexical activation should ideally be zero – meaning that the segments activated by the free phone loop perfectly match the segments of a certain lexical entry. Such a result would be unwanted for a pseudoword as input, as it would indicate not only that the free phone loop activated wrong phones, but also that the activated phones matched a string existing in the lexicon. Therefore, one would expect a bimodal distribution of differences, in which the discrepancy between the free phone loop and the best lexical activation is small for one group of sounds (words), and large for the other (pseudowords), with minimum overlap between the two groups. When generating these estimates, since activation is cumulative, we divided the calculated difference in activation by the total number of phonemes in the signal word or pseudoword.

Finally, we estimated responses latencies and compared them to MALD participant responses. In this case, we used 17

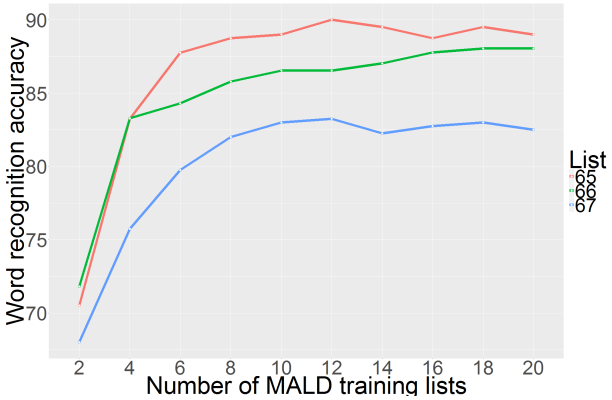

Figure 2: *Free word recognition accuracy of the 20 models adapted for speaker in the three MALD test lists.*

MALD lists (6,800 words, lists 51 to 67) and ran a gating simulation as described above, looking at top 20 competitors for any given signal. We estimated response latencies only for those words which appeared in the competitor list at word offset and which had less than 20 competitors remaining, to avoid ceiling effects when estimating the number of remaining competitors. Additionally, we included only the words for which there were at least three correct participant responses. The final number of retained words was 5,604.

The response latency estimation in DIANA is computed as the sum of activation, decision, and execution time. Since many words have minimal pairs only diverging at the end of the word and since the signal could change into a pseudoword at any point as it unfolds, both simulations and participant performance in auditory lexical decision experiments most often do not show a clear winner before word offset. Therefore, the model needs to account for the additional cognitive processing occurring after the acoustic information is no longer available. To do so, DIANA estimates the number of remaining plausible options (word and pseudoword alike) and, based on that number, the corresponding choice reaction time (using 'Hick's law'). In other words, response latency is depending on competitor activation, the number of plausible competitors at word offset, and a fixed increase in latency which accounts for execution time (200 ms). Estimates generated for words in this fashion were then compared to mean logged response times provided by 64 unique participants in the MALD experiment.

3. Results and discussion

Twenty models adapted on the different number of MALD words were compared (Figure 2) in free word recognition. The results show a large increase in accuracy up to 4,000 words (i.e., 10 MALD lists). After that, accuracy remains roughly the same and never reaches 90% for any of the lists. Therefore, we decided to use the models adapted on 10 lists for all future simulations.

Accuracy is not the same for the three test lists. The most probable explanation is that certain lists are more similar to each other when it comes to their acoustic content. For example, list 67 would then include a larger number of phones that are more often erroneously classified, due to them not occurring as often in training or adaptation material. Alternatively, certain lists may contain words with a larger number of close competitors, making the selection of the correct winner more difficult.

Table 1: *Activation of competitors at word offset for two example words.*

Target word	Competitor	Activation
BROWSE (correct)	BROWSE	-2,890.86
	BROWS	-2,890.86
	BROWNS	-2,938.98
	ROUNDS	-2,941.75
ASSURED (incorrect)	USHERED	-4,475.29
	ASSURED	-4,485.90
	ISSUED	-4,522.81
	PRESSURED	-4,549.67

We also noticed considerable volatility in free word recognition with even small changes in the adaptation material (see e.g., drops in accuracy for list 67 at training list 12 and 65 at training list 14) which we will not discuss in detail here, but which in addition to relatively low recognition accuracy warn that the current acoustic models need improvement.

However, even with these issues, Table 1 shows that competitors with high activation values tend to be similarly sounding words, regardless of whether the correct winner is selected or not. In the first case, the sound signal was the word *browse*, and the target word had the highest activation value at signal offset. However, word *brows*, which has identical pronunciation as described in CMU dictionary, has the same activation value. Close competitors included a word with an additional nasal (*browns*), and a word without the initial stop, but with an additional word-medial stop (*rounds*). In the second case, the correct word *assured* was not the winner, but it was a very close runner-up to a similar word *ushered*.

The simulations of competition as a function of time also provide the expected outcome — at first, there are no competitors that are distinguishable as potentially better than others, then a number of competitors rise and fall in activation, with very little difference between them, and, finally, a small subset of competitors starts separating from the group and rising in activation, with potentially one of them separating from this group as well, emerging as the clear front-runner.

Lexical decision task simulation was based on differences in activation in free word recognition and activation in the free phone loop. Words in comparison to pseudowords indeed tended to have a smaller difference in activations (Figure 3). However, the results of the simulation also revealed certain issues in the acoustic models. As can be seen in the figure, a number of pseudowords were matched with a lexical entry (having the difference of activations equal to zero), which should not happen unless the acoustic models make errors in the free phone loop i.e., in recognizing phones. Similarly, certain words have a difference in activations larger than 0, again indicating errors in the free phone loop.

The final step in the lexical decision simulation is to select the activation difference value that will serve as the threshold between the two choices participants can make. If a 'balanced response regime' is assumed, placing the threshold so as to have approximately the same amount of 'word' and 'pseudoword' responses, the error rates for the three lists are 20.87%, 20.01%, and 17.60%. MALD participants completing the same lists have an average error rate of 13.14%.

Finally, comparing the response latencies estimated by DIANA to those obtained by participants (averaged logged participant response latency) showed moderate correlation, similar to that observed in previous applications of the model ($r = .46$).

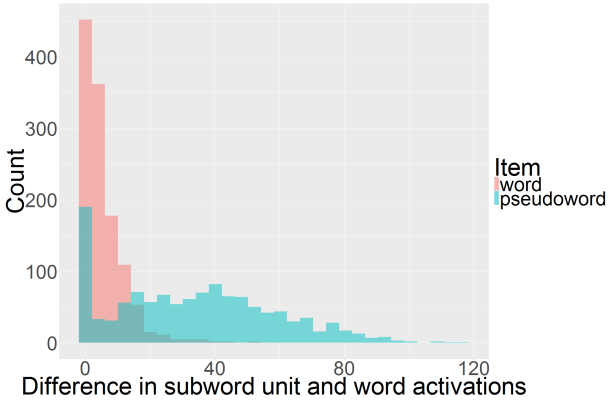

Figure 3: *The difference between free word and free phone loop activation values divided by the number of phones for words and pseudowords.*

Additionally, estimates also seem to describe the actual time required for the response well, as the mean response latency for our participants was 943 ms, while DIANA estimates average at 976 ms.

4. Conclusions

The acoustic models we created on the basis of spontaneous speech corpora seem to adequately capture the basic assumptions of spoken word recognition. In free word recognition, the models correctly recognized the target word from a lexicon of nearly 28,000 words in 85-90% of cases. The models also successfully simulate word competition at speech signal offset and as the speech signal unfolds. However, error rates are still relatively high, which is especially visible when simulating the lexical decision task i.e., the word/pseudoword response. Therefore, the models need improving before they can perform on par with existing acoustic models for North American English, either by including more recordings or by addition of careful speech corpora.

The reaction latencies estimated using DIANA remain satisfactory, even as they are correlated with averaged response times from only 3 to 5 participants. An additional source of variability that DIANA is unable to model yet are local speed effects (fatigue, distraction, attention fluctuation during the session, etc.) and many long-term and medium-term differences between participants (even whether a participant plays videogames or not). Our future attempts will also attempt to account for stimuli characteristics by including top-down weights i.e., word frequency effects, other lexical predictors, and considering morphological complexity of the target word.

5. Acknowledgements

This project funded by the Social Sciences and Humanities Research Council: Grant #435-2014-0678.

6. References

[1] A. Protopapas, "Connectionist modeling of speech perception." *Psychological Bulletin*, vol. 125, no. 4, p. 410, 1999.

[2] J. M. McQueen, "Eight questions about spoken-word recognition." Oxford University Press, USA, 2007, pp. 37–53.

[3] O. Scharenborg and L. Boves, "Computational modelling of spoken-word recognition processes: Design choices and evaluation," *Pragmatics & Cognition*, vol. 18, no. 1, pp. 136–164, 2010.

[4] A. Weber and O. Scharenborg, "Models of spoken-word recognition," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 3, no. 3, pp. 387–401, 2012.

[5] J. S. Magnuson, D. Mirman, and H. D. Harris, "Computational models of spoken word recognition." Cambridge University Press Cambridge, UK, 2012, pp. 76–103.

[6] L. ten Bosch, M. Ernestus, and L. Boves, "DIANA: An end-to-end computational model of human speech processing," in preparation.

[7] B. V. Tucker, D. Brenner, D. K. Danielson, M. C. Kelley, F. Nenadić, and M. Sims, "Massive auditory lexical decision: Toward reliable, generalizable speech research," *Behavior Research Methods*, in print.

[8] J. L. McClelland and J. L. Elman, "The trace model of speech perception," *Cognitive Psychology*, vol. 18, no. 1, pp. 1–86, 1986.

[9] T. J. Strauss, H. D. Harris, and J. S. Magnuson, "jtrace: A reimplementation and extension of the trace model of speech perception and spoken word recognition," *Behavior Research Methods*, vol. 39, no. 1, pp. 19–30, 2007.

[10] D. Norris, "Shortlist: A connectionist model of continuous speech recognition," *Cognition*, vol. 52, no. 3, pp. 189–234, 1994.

[11] D. Norris and J. M. McQueen, "Shortlist B: A bayesian model of continuous speech recognition," *Psychological Review*, vol. 115, no. 2, pp. 357–395, 2008.

[12] O. Scharenborg, "Modelling fine-phonetic detail in a computational model of word recognition," in *the 9th Annual Conference of the International Speech Communication Association*. ISCA Archive, 2008, pp. 1473–1476.

[13] ——, "Using durational cues in a computational model of spoken-word recognition," in *10th Annual Conference of the International Speech Communication Association [Interspeech 2009]*. ISCA Archive, 2009, pp. 1675–1678.

[14] L. ten Bosch, G. Giezenaar, L. Boves, and M. Ernestus, "Modeling language-learners' errors in understanding casual speech," *Errors by humans and machines in multimedia, multimodal, multilingual data processing*, pp. 7–121, 2016.

[15] L. Ten Bosch, L. Boves, and M. Ernestus, "Towards an end-to-end computational model of speech comprehension: Simulating a lexical decision task," in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 2822–2826.

[16] L. ten Bosch, M. Ernestus, and L. Boves, "Comparing reaction time sequences from human participants and computational models," in *Interspeech 2014: 15th Annual Conference of the International Speech Communication Association*, 2014, pp. 462–466.

[17] M. Ernestus and R. Baayen, "The comprehension of acoustically reduced morphologically complex words: The roles of deletion, duration, and frequency of occurrence," in *Proceedings of the 16th International Congress of Phonetic Sciences*, 2007, pp. 773–776.

[18] L. Ten Bosch, L. Boves, and M. Ernestus, "DIANA, an end-to-end computational model of human word comprehension," in *18th International Congress of Phonetic Sciences (ICPhS 2015)*. University of Glasgow, 2015.

[19] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report*, vol. 93, 1993.

[20] N. Warner, "Reduction," in *The Blackwell Companion to Phonology: General issues and segmental phonology*, M. van Oostendorp, C. J. Ewen, E. Hume, and K. Rice, Eds. John Wiley & Sons, 2011, pp. 1866–1891.

[21] E. Keuleers and M. Brysbaert, "Wuggy: A multilingual pseudoword generator," *Behavior Research Methods*, vol. 42, no. 3, pp. 627–633, 2010.

[22] R. Weide, "The carnegie mellon pronouncing dictionary [cmudict. 0.6]," 2005. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[23] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The htk book (version 3.4)," *Cambridge University Engineering Department*, 2006. [Online]. Available: http://htk.eng.cam.ac.uk/

[24] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Cambridge, United Kingdom: Cavendish Laboratory, Tech. Rep., 2006.