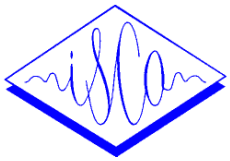

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

PRONUNCIATION MODELING FOR SPONTANEOUS SPEECH BY MAXIMIZING WORD CORRECT RATE IN A PRODUCTION-RECOGNITION MODEL

Ming-yi Tsai, Lin-shan Lee

Graduate Institute of Communication Engineering
National Taiwan University, Taipei, Taiwan, Republic of China
Email: pancho@speech.ee.ntu.edu.tw, lsl@iis.sinica.edu.tw

ABSTRACT

In this paper, we develop a new method for compiling a pronunciation dictionary to model pronunciation variation in spontaneous speech recognition. The pronunciation dictionary is assembled by iteratively selecting pronunciations from a data-driven word confusion table, based on directly maximizing the word correct rate simulated by a production-recognition model such that the optimal performance of recognition can be achieved. In other words, the compiled pronunciation dictionary can not only accommodate as many as necessary pronunciations but also avoid possible introduced confusion during recognition. The simulation of word correct rate is performed with a novel human-machine communication model, consisting of a human speech production module and a machine speech recognition module. Our experimental results on LDC Mandarin Call Home and Call Friend corpora showed that significant improvement is achieved with this new approach. Furthermore, the framework and theory presented here are applicable to other languages.

1. INTRODUCTION

It is well known that the pronunciation variation in spontaneous speech could seriously deteriorate the performance of ASR systems, if not dealt carefully. Usually, pronunciation variation is handled by enumerating probable pronunciations for each word in a pronunciation dictionary with a prior probability for each pronunciation. However, simply adding several pronunciation variants to the dictionary would increase the homophone rate and hence may not be helpful for the overall recognition performance. It might also fail in selecting an adequate pronunciation set to avoid the introduced confusion that merely emphasizing on generating credible pronunciation variants through time-alignment[1], phone duration[2], acoustic likelihood or confidence measure [3-5], etc.. To optimize the performance, one needs to add the variants which maintain a reasonable balance between solving original recognition errors and introducing new confusion during recognition. Although, some researchers have taken this issue into account in modeling pronunciation variation, few studies have been systematically carried out concerning the explicit relationship between pronunciation confusion and recognition performance. For example in [6], penalties were assigned to certain pronunciation variants trying to prevent the potential confusion. However, how to systematically implement the penalty function is still an open question, and the introduced penalty has been suggested as a possible cause to the worse performance. On the other hand, some studies on measuring pronunciation confusion show that the overall recognition performance can be improved by avoiding possible confusion [7-11], but a direct correlation between these approaches and the overall recognition performance is not observed. For this reason, we attempt to develop a pronunciation variation modeling approach with direct correlation to the recognition performance in a systematical way. We model the pronunciation variation using a pronunciation dictionary which is assembled by iteratively selecting pronunciations from a data-driven word confusion table with a criterion of explicitly maximizing the word correct rate simulated by a production-recognition model such that a reasonable balance between solving original recognition errors and introducing new confusion during recognition can be achieved. As a result, the assembled pronunciation dictionary can not only accommodate as most as necessary pronunciations but also avoid possible confusability.

In order to simulate the word correct rate of recognition, a human-machine communication model is derived as a hybrid of theory in psycholinguistics (human speech production module) and techniques in automatic speech recognition (machine speech recognition module). Accordingly, the word correct rate during the human-machine communication process can be simulated in either a context-independent or a context-dependent way with this model.

In our opinion, the modeling framework proposed in this study provides a useful reference for researchers attempting to optimize a component in ASR system explicitly correlating with the recognition performance, and for psycholinguists interested in the computational aspects of psycholinguistic modeling.

This paper is organized as follows. In section 2, we describe the procedure to acquire the data-driven word-level confusion table. Section 3 presents the human-machine communication model and the formulation of the simulated word correct rate in both context-independent and context-dependent manners. In section 4, the algorithm for selecting pronunciations is outlined based on a criterion of maximizing the simulated word correct rate. Section 5 describes the corpus used for experiments and discusses the experimental results of our new method. Finally, the section 6 offers brief concluding remarks.

2. PRONUNCIATION GENERATION

Figure.1, illustrates the construction process for a data-driven word-level confusion table [7]. The surface forms of training data are obtained by unconstrained phone recognition. After aligning them with the canonic forms by dynamic programming, a phone-level confusion table is produced and converted to an augmented phone-unit dictionary with prior probabilities of pronunciations for each phone. It is then used for phone-level forced alignment of the same training data against its orthographic phone transcriptions to obtain the more reliable surface forms. Afterwards, a word-level confusion table consisting of the probable pronunciations and prior probabilities is obtained.

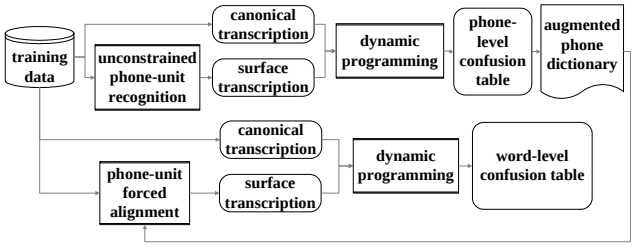

Figure.1 *The word-level confusion table construction process*

3. THE HUMAN-MANCHINE COMMUNICATION MODEL

Figure.2 illustrated our proposed human-machine communication model consisting of a human speech production module and a machine speech recognition module.

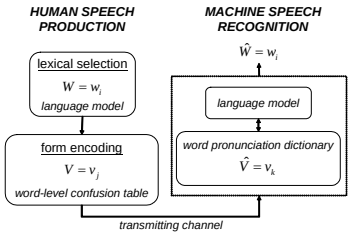

Figure.2 *The human-machine communication model*

3.1. The Human Speech Production Module

Although the process of how human retrieve and use words in communication is extremely complicated, a staged and feed-forward description is generally accepted by psycholinguists [12]. Models of lexical access have been always conceived as process models of normal speech production. For example, a serial two-system architecture of lexical access was presented [13], in which a speaker produces a word by first selecting an appropriate item from his mental lexicon (the lexical selection system), next the selected item's articulatory shape is prepared (the form encoding system). Here, we postulate our human speech production module with the two same stages, lexical selection and form encoding, as shown in the left of ***Figure.2***. But, unlike those in psycholinguistic studies, this module is not intended to imitate the true human behavior, but to systematically simulate the statistics of words and pronunciations produced by human. Therefore, we use a language model, trained on the training corpus, to model the lexical selection process and a word-level confusion table, consisting of possible pronunciations and prior probability for each word, to model the form encoding process.

3.2. The Machine Speech Recognition Module

As shown in the right of ***Figure.2***, the automatic speech recognition process is modeled by a machine speech recognition module. In order to simulate the recognition results of an automatic speech recognizer, it is composed of a language model, trained on the training corpus, and a pronunciation dictionary, composed of a certain set of pronunciations for each word.

3.3. Simulation of Word Correct Rate in the Human-machine Communication Process

In the human-machine communication process as indicated in ***Figure.2***, a speaker retrieves a word $\boldsymbol{W}$ in the lexical selection stage and the corresponding pronunciation $\boldsymbol{V}$ for this word is then assembled in the following form encoding stage. Next, the transmitting channel conveys the pronunciation $\boldsymbol{V}$ from the speaker to the machine speech recognition module, where the incoming pronunciation $\boldsymbol{V}$ is recognized as the pronunciation $\hat{V}$ and the corresponding word $\hat{W}$. Accordingly, our goal is to select an optimal pronunciation set from a word confusion table into the word pronunciation dictionary in the Machine Speech Recognition Module, such that the simulated correct rate of the recognized word $\hat{W}$ is maximized.

Based on the statistics from the underlying components including language model, word-level confusion table, and word pronunciation dictionary, the word correct rate of recognition can be simulated systematically. Let $\boldsymbol{\Gamma}$ be a finite set of words, and is referred to as the *vocabulary*, and $|\boldsymbol{\Gamma}|$ be the number of distinct words in the vocabulary. Furthermore, let $\boldsymbol{\Omega_i}$ be the set consisting of all pronunciations (or phoneme sequences) realized for the word $\boldsymbol{w_i \in \Gamma}$ according to the data-driven word-level confusion table, and $|\boldsymbol{\Omega_i}|$ be the number of distinct pronunciations realized for $\boldsymbol{w_i}$. Accordingly, the simulated word correct rate of the human-machine communication process is defined as

$$Cor = \sum_{w_i \in \Gamma} P\left(W = w_i, \hat{W} = w_i\right). \tag{1}$$

Furthermore, it can be interpreted as

$$\begin{aligned} Cor &= \sum_{w_i \in \Gamma,\ v_j, v_k \in \Omega_i} P\left(W = w_i, V = v_j, \hat{W} = w_i, \hat{V} = v_k\right) \\ &= \sum_{w_i \in \Gamma,\ v_j, v_k \in \Omega_i} P(W = w_i) \cdot P\left(V = v_j \mid W = w_i\right) \cdot P\left(\hat{V} = v_k \mid V = v_j, W = w_i\right) \cdot \\ &\qquad P\left(\hat{W} = w_i \mid \hat{V} = v_k, V = v_j, W = w_i\right). \end{aligned} \tag{2}$$

Since only the pronunciation $\boldsymbol{V}$ is transmitted to the machine, it is assumed that the probability of recognizing the pronunciation $\hat{V}$ and its corresponding word $\hat{W}$ is independent of the produced word $\boldsymbol{W}$. Consequently,

$P\left(\hat{W} = w_i \mid \hat{V} = v_k, V = v_j, W = w_i\right) = P\left(\hat{W} = w_i \mid \hat{V} = v_k, V = v_j\right)$

and $P\left(\hat{V} = v_k \mid V = v_j, W = w_i\right) = P\left(\hat{V} = v_k \mid V = v_j\right)$.

In addition, we assume that the produced pronunciation (or phoneme sequence) $\boldsymbol{V}$ can be always correctly recognized such that $P\left(\hat{V} = v_k \mid V = v_j\right) = 1$ *iff* $j = k$.

Thus, the simulated word correct rate becomes

$$Cor = \sum_{w_i \in \Gamma,\ v_j \in \Omega_i} P(W = w_i) P(V = v_j \mid W = w_i) P(\hat{W} = w_i \mid \hat{V} = v_j) \cdot \quad (3)$$

3.3.1. Simulating the Word Correct Rate in a Context-independent manner

Here, we attempt to iteratively simulate the word correct rate of recognition with the available contextual-independent statistics. From ***eq.* (3)**, the simulated word correct rate of recognition at iteration ***t*** can be interpreted as

$$Cor_t = \sum_{w_i \in \Gamma,\ v_j \in \Omega_i} P(W = w_i) P(V = v_j \mid W = w_i) \frac{P_t(\hat{V} = v_j \mid \hat{W} = w_i) P(\hat{W} = w_i)}{P_t(\hat{V} = v_j)}, \quad (4)$$

where $P(W=w_i)$ is the probability of producing a word $\boldsymbol{w_i}$, which is acquired from a unigram language model trained on the training corpus. $P(\hat{W} = w_i)$ is the probability of recognizing a word as $\boldsymbol{w_i}$, and is modeled by the same unigram language model. Besides, $P(V = v_i \mid W = w_i)$ stands for the probability of a word $\boldsymbol{w_i}$ being realized as a pronunciation $\boldsymbol{v_j}$, and is acquired from the data-driven word-level confusion table. $P_t(\hat{V} = v_i \mid \hat{W} = w_i)$ stands for the probability of a pronunciation $\boldsymbol{v_j}$ allowed to be realized for a word $\boldsymbol{w_i}$ in the pronunciation dictionary, and it is going to be determined at each iteration ***t*** such that the overall simulated word correct rate $\boldsymbol{Cor_t}$ at iteration ***t*** can be maximized. Furthermore, $P_t(\hat{V} = v_i)$ is the probability of a pronunciation $\boldsymbol{v_j}$ allowed to be realized in the pronunciation dictionary at iteration ***t***.

3.3.2. Simulating the Word Correct Rate in a Context-dependent manner

From ***eq.* (3)**, the word correct rate of recognition at iteration ***t*** is simulated with the available context-dependent statistics and interpreted as

$$\begin{aligned} Cor &= \sum_{\substack{w_l, w_m \in \Gamma, \\ v_j \in \Omega_i}} P(W = w_l, W_{-1} = w_m) \cdot P(V = v_j \mid W = w_l, W_{-1} = w_m) \cdot P(\hat{W} = w_l, \hat{W}_{-1} = w_n \mid \hat{V} = v_j, W_{-1} = w_m) \\ &= \sum_{w_i, w_m \in \Gamma,\ v_j \in \Omega_i} P(W = w_i, W_{-1} = w_m) \cdot P(V = v_j \mid W = w_i, W_{-1} = w_m) \cdot \\ &\qquad \sum_{w_n \in \Gamma} P(\hat{W}_{-1} = w_n \mid \hat{V} = v_j, W_{-1} = w_m) \cdot P(\hat{W} = w_i \mid \hat{W}_{-1} = w_n, \hat{V} = v_j, W_{-1} = w_m), \end{aligned} \quad (5)$$

where $\boldsymbol{W_{-1}}$ stands for the previous produced word before the currently produced word $\boldsymbol{W}$, and $\hat{W}_{-1}$ stands for the previous recognized word before currently recognized word $\hat{W}$.

Generally, the recognizer makes decision without knowledge about the future. Therefore, recognizing the previous word is irrelevant to the current recognized pronunciation, i.e.

$$P(\hat{W}_{-1} = w_n \mid \hat{V} = v_j, W_{-1} = w_m) = (\hat{W}_{-1} = w_n \mid W_{-1} = w_m) \cdot$$

As before, the probability of recognizing the word $\hat{W}$ is assumed to be independent of the produced word $\boldsymbol{W}$, i.e.

$$P(\hat{W} = w_i \mid \hat{W}_{-1} = w_n, \hat{V} = v_j, W_{-1} = w_m) = P(\hat{W} = w_i \mid \hat{W}_{-1} = w_n, \hat{V} = v_j) \cdot$$

Consequently, the simulated word correct rate at iteration ***t*** becomes

$$\begin{aligned} Cor_t &= \sum_{w_i, w_m \in \Gamma,\ v_j \in \Omega_i} P(W = w_i, W_{-1} = w_m) \cdot P(V = v_j \mid W = w_i, W_{-1} = w_m) \cdot \\ &\qquad \sum_{w_n \in \Gamma} P_t(\hat{W}_{-1} = w_n \mid W_{-1} = w_m) \cdot P_t(\hat{W} = w_i \mid \hat{W}_{-1} = w_n, \hat{V} = v_j) \\ &= \sum_{w_i, w_m \in \Gamma,\ v_j \in \Omega_i} P(W = w_i, W_{-1} = w_m) \cdot P(V = v_j \mid W = w_i, W_{-1} = w_m) \cdot \\ &\qquad \Big[P_t(\hat{W}_{-1} = w_m \mid W_{-1} = w_m) \cdot P_t(\hat{W} = w_i \mid \hat{W}_{-1} = w_m, \hat{V} = v_j) + \\ &\qquad \sum_{w_n \in \Gamma, w_n \neq w_m} P_t(\hat{W}_{-1} = w_n \mid W_{-1} = w_m) \cdot P_t(\hat{W} = w_i \mid \hat{W}_{-1} = w_n, \hat{V} = v_j) \Big] \cdot \end{aligned} \quad (6)$$

Furthermore, it can be interpreted with the available statistics as

$$\begin{aligned} Cor_t &= \sum_{w_i, w_m \in \Gamma,\ v_j \in \Omega_i} P(W = w_i, W_{-1} = w_m) \cdot P(V = v_j \mid W = w_i, W_{-1} = w_m) \cdot \\ &\qquad \Big[Cor_{t-1} \cdot P_t(\hat{W} = w_i \mid \hat{W}_{-1} = w_m, \hat{V} = v_j) + \\ &\qquad (1 - Cor_{t-1}) \sum_{\substack{w_n \in \Gamma, \\ w_n \neq w_m}} P_{\overline{m}}(\hat{W}_{-1} = w_n) \cdot P_t(\hat{W} = w_i \mid \hat{W}_{-1} = w_n, \hat{V} = v_j) \Big], \end{aligned} \quad (7)$$

where $P_{\overline{m}}(\hat{W}_{-1} = w_n) = \frac{N_n}{N - N_m}$, $\boldsymbol{N_m}$ and $\boldsymbol{N_n}$ represents the times of occurrence of the word $\boldsymbol{w_m}$ and $\boldsymbol{w_n}$ in the training corpus respectively, $\boldsymbol{N}$ is the total number of word tokens in the training corpus, and

$$\begin{aligned} P_t(\hat{W} = w_i \mid \hat{W}_{-1} = w_n, \hat{V} = v_j) &= \frac{P_t(\hat{V} = v_j \mid \hat{W} = w_i, \hat{W}_{-1} = w_n) P(\hat{W} = w_i \mid \hat{W}_{-1} = w_n)}{P_t(\hat{V} = v_j \mid \hat{W}_{-1} = w_n)} \\ &= \frac{P_t(\hat{V} = v_j \mid \hat{W} = w_i) P(\hat{W} = w_i \mid \hat{W}_{-1} = w_n)}{\sum_{w_r \in \Gamma} P_t(\hat{V} = v_j \mid \hat{W} = w_r) P(\hat{W} = w_r \mid \hat{W}_{-1} = w_n)}, \end{aligned}$$

since a context-independent pronunciation dictionary is commonly used for recognition.

4. THE ALGORITHM FOR MAXIMIZING THE SIMULATED WORD CORRECT RATE

An algorithm is developed for selecting the pronunciations to be included in the recognition dictionary based on a criterion of maximizing the simulated word correct rate during human-machine communication process.

As defined in section 4, $\boldsymbol{\Gamma}$ is a finite set of words and $\boldsymbol{\Omega_i}$ represents the set consisting of all pronunciations realized for the word $\boldsymbol{w_i \in \Gamma}$, according to the word-level confusion table. Furthermore, let $y_i = \{b_{ij} = 1 \text{ or } 0 \ ; \ 1 \le j \le |\Omega_i|\}$ be a set consisting of $|\boldsymbol{\Omega_i}|$ binary variables for the word $\boldsymbol{w_i}$, in which $\boldsymbol{b_{ij}}$=1 indicates that the pronunciation $\boldsymbol{v_j \in \Omega_i}$ is allowed for word $\boldsymbol{w_i}$ in the compiled pronunciation dictionary and otherwise $\boldsymbol{b_{ij}}$=0. Thus, our goal is to determine an appropriate $Y = \{y_1, ..., y_{|\Gamma|}\}$, such that the simulated word correct rate, as formulated as **eq. (4)** or **eq.(7)**, can be maximized by the resulting recognition dictionary.

The algorithm consists of the following steps:

1. Initialize $\boldsymbol{Y_0}$ ($\boldsymbol{Y}$ at iteration ***t***=0) by setting all $\boldsymbol{b_{ij}}$=1 for each set $\boldsymbol{y_i}$.
2. For iteration ***t=1,…,T,***
   For ***i=1,…,|Γ|***,
   (a) Calculate the overall simulated word correct rate $\boldsymbol{Cor_t}$ with all words excluding the word $\boldsymbol{w_i}$, which is referred as $\boldsymbol{Cor_t(i,0)}$.
   (b) For ***j=1,…,|Ω_i|***,
   (i.) Calculate the overall simulated word correct rate $\boldsymbol{Cor_t}$ with all words including the word $\boldsymbol{w_i}$, but only the corresponding pronunciation $\boldsymbol{v_j}$ is allowed in the dictionary for the word $\boldsymbol{w_i}$, which is referred as $\boldsymbol{Cor_t(i,j)}$.

(ii.) Add/allow the corresponding pronunciation v_j for the word w_i in the dictionary only if $Cor_t(i,j) - Cor_t(i,0) > TH$

T is the number of times of iteration. The chosen pronunciation set and the corresponding simulated word correct rate have no significant change after about 5 iterations. In addition, TH is the threshold of the incremental correct rate which is empirically tuned for controlling the number of pronunciations to be included into the pronunciation dictionary.

5. EXPERIMENTS AND DISCUSSION

To evaluate the effectiveness of our proposed pronunciation modeling approach for ASR, the preliminary experiments were performed with the HTK tools on spontaneous conversational speech of LDC Mandarin Call Home and Call Friend corpora. After removing the laughters, filled pauses, corruptive background and channel noise, and those words in other languages, about 18 hours of speech is used to train the gender-dependent and context-dependent Mandarin Initial/Final models. The test set is another 30 minutes of data in Mandarin Call Home corpus with Putonghua accent and annotation of reasonably good level of quality. It was tested with a bigram language model based on a lexicon of roughly 10K words, trained on the training set of about 400K word tokens. With a dictionary containing only canonical pronunciations, the character accuracy of baseline system is **38.19%**. When using a pronunciation dictionary compiled by selecting pronunciations concerning their frequency of occurrence [7], the character accuracy is improved to **39.48%**. While using our new pronunciation modeling approach with the word correct rate simulated in context-independent manner, the character accuracy is further improved to **40.68%**.

We also use Hub4NE 1997 Broadcast News database provided by LDC to evaluate the effectiveness of this modeling approach. About 30 hours of Hub4NE data is used to train the acoustic models and acquire the word confusion table. The test data is one hour of Hub4NE Test Material provided by LDC. It was tested with a bigram language model based on a lexicon of roughly 24K words, trained on the Hub4NE training set. The character accuracy of baseline system is **62.99%**. When using a pronunciation dictionary compiled by selecting pronunciations concerning their frequency of occurrence [7], the character accuracy is improved to **63.79%**. The character accuracy is further improved to **64.51%** by our new pronunciation modeling approach with the word correct rate simulated in context-independent manner.

These experimental results indicate that our approach can achieve further improvement by modeling pronunciation variation with explicit correlation to the recognition performance. Moreover, it reveals that although the human-machine communication model is derived on some assumptions and may not so fit in the truth, it performs well as simulating the human-machine communication process.

6. COUNCLUDSION

The framework and its underlying models proposed in this paper provide us a new direction and good opportunity to model the pronunciation variation or even other components in ASR systems concerning with enhancing the recognition performance directly.

So far, we have implemented the new pronunciation modeling approach with the word correct rate simulated in context-independent manner and the preliminary experimental results showed that significant improvement can be achieved. We will continue to implement the context-dependent version of this model for better modeling pronunciation variation in automatic speech recognition by taking context dependent information into account.

7. REFERENCES

[1] M. Ravishankar, et al., "Automatic Generation of Context-Dependent Pronunciations," Proc. of European Conference on Speech Communication and Technology, 1997.

[2] H. J. Nock, et al., "Detecting and Correcting Poor Pronunciations for Multiword Units," Proc. of ESCA Workshop: Modeling Pronunciation Variation for Automatic Speech Recognition, 1998.

[3] K. L. Markey, et al., "Lexical Tuning Based on Triphone Confidence Estimation," Proc. of European Conference on Speech Communication and Technology, 1997.

[4] D. A. G. Williams, "Knowing What You Don't Know: Roles for Confidence Measures in Automatic Speech Recognition," in *Department of Computer Science*: University of Sheffield, 1999, pp. 123.

[5] T. Holter, et al., "Maximum Likelihood Modelling of Pronunciation Variation," *Speech Communication*, vol. 29, pp. 177-191, 1999.

[6] M. Riley, et al., "Stochastic Pronunciation Modelling from Hand-Labelled Phonetic Corpora," *Speech Communication*, vol. 29, pp. 209-224, 1999.

[7] M.-y. Tsai, et al., "Improved Pronunciation Modeling by Properly Integrating Better Approaches for Baseform Generation, Ranking and Pruning," Proc. of ISCA Workshop: Pronunciation Modeling and Lexicon Adaptation for Spoken Language, 2002.

[8] E. Fosler-Lussier, et al., "On the Road to Improved Lexical Confusability Metrics," Proc. of ISCA Workshop: Pronunciation Modeling and Lexicon Adaptation for Spoken Language, 2002.

[9] L. T. Bosch, et al., "Pronunciation Modeling and Lexical Adaptation Using Small Training Sets," Proc. of ISCA Workshop: Pronunciation Modeling and Lexicon Adaptation for Spoken Language, 2002.

[10] H. Schramm, et al., "Discriminative Optimization of the Lexical Model," Proc. of ISCA Workshop: Pronunciation Modeling and Lexicon Adaptation for Spoken Language, 2002.

[11] M. Wester, et al., "A Comparison of Data-Derived and Knowledge-Based Modeling of Pronunciation Variation," Proc. of International Conference on Spoken Language Processing, 2000.

[12] W. O'Grady, et al., *Contemporary Linguistics*, 4 ed: Bedford/St. Martin's, 2001.

[13] W. J. M. Levelt, "Spoken Word Production: A Theory of Lexical Access," Proc. of National Academy of Sciences, 2001.