Chapter #

Is phonetic knowledge of any use
for speech technology?

Helmer Strik
A²RT, Dept. of Language and Speech, University of Nijmegen, the Netherlands

NOTE: This is the penultimate version of the paper, as it has been submitted. The final
version might be (slightly) different.

Abstract: Although it has often been advocated that more phonetic knowledge should be
incorporated in speech technology, the amount of phonetic knowledge used in
speech technology has decreased over the years. In order to get a better
understanding of why this is the case, some examples of attempts to transfer
phonetic knowledge to speech technology are presented. These examples make
clear that there are several reasons why this transfer is problematic: different
approaches are used in the fields of phonetics and speech technology, phonetic
knowledge is based on small amounts of 'lab speech' and therefore does not
generalize to 'real speech', the knowledge is not complete, and the knowledge
is not quantified in the right format.

Key words: Phonetic knowledge, speech technology, ASR, TTS

1

1.   INTRODUCTION

"Is phonetic knowledge any use?" was the title of the panel discussion that took place at Eurospeech 2001 on Friday September 7, 2001 in Aalborg. This panel discussion was the second part of the Eurospeech special event entitled "Integration of Phonetic Knowledge in Speech Technology". In this paper we will take a look at the integration of phonetic knowledge in speech technology. We start with some notes on the two terms: 'phonetic knowledge' and 'speech technology'.

Speech technology is a term that covers many fields, like speech coding, speech-to-speech translation, text-to-speech (TTS), concept-to-speech, speaker identification, speaker verification, speaker tracking, automatic speech recognition (ASR), speech understanding, etc. Of all these fields, only ASR and TTS are addressed in the current paper, while the main focus is on ASR.

Giving a short and clear definition of phonetic knowledge is not straightforward. In fact, what exactly constitutes phonetic knowledge has been the topic of many discussions. In the context of the current article, we will not attempt to establish what the exact nature of phonetic knowledge is (e.g. to specify what exactly is phonetic and what is phonological knowledge). We believe that more important questions concern the role of phonetic/linguistic knowledge in speech technology: to what extent is it used, should this increase or diminish, etc. Consequently, in the current paper, the focus is on phonetic knowledge in a broad sense, which sometimes may even mean more general linguistic knowledge.

It has often been advocated that more phonetic knowledge should be used in ASR and TTS (Stevens, 1960; Zue, 1983; Zue, 1991; Pols, 1999). However, in many ASR and TTS systems it is not straightforward how phonetic knowledge should be integrated into these systems. One way of doing this is by using articulatory(-based) features, which has been tried in various research projects. In general, the goal of integrating phonetic knowledge into ASR and TTS systems was to increase the performance of these systems. However, ASR and TTS have also been regarded as means to test existing phonetic knowledge, to find out whether gaps and/or errors in the existing phonetic knowledge were present, and where. Furthermore, it has been suggested that ASR and TTS should be (partly) integrated, because human speech production and perception are not independent (Stevens, 1960).

Although many seem to be in favor of integrating (more) phonetic knowledge in speech technology, in the last decades we have witnessed a decrease in the amount of phonetic knowledge used in ASR and TTS (e.g. Zue, 1983; Zue, 1991). However, it is certainly not the case that the use of phonetic/linguistic knowledge has been abandoned in current systems. For instance, ASR and TTS systems make use of the knowledge that speech consists of words, that these words do not occur in a random order, that these words are made up of syllables and phonemes, that these phonemes do not occur in a random order, and much more knowledge on speech production, acoustics, and perception. More specifically, when we develop ASR and TTS systems, we often make use of e.g. the phoneme inventory of a language, a lexicon, grapheme-to-phoneme conversion, phonetic transcriptions, segmentations, and phonetic features, which are often derived using knowledge about speech perception.

In section 2 we will argue that phonetics and speech technology are essentially two different worlds, which hinders the transfer of phonetic knowledge to speech technology. Some examples of (not) using phonetic knowledge in speech technology are given in section 3. First, a description is given of three examples of attempts to use phonetic knowledge in ASR which were pursued in research carried out at our department. They are presented in chronological order. Two other examples are discussed at the end of section 3. Finally, the discussion is presented in section 4.

2. PHONETICS AND SPEECH TECHNOLOGY: TWO DIFFERENT WORLDS

Why has the amount of phonetic knowledge used in speech technology decreased over the years? An obvious answer would be: Because systems in which less phonetic knowledge is used perform better. For many people (researchers, developers, retailers and users) this is indeed the most important aspect of a system: it should perform well. Therefore, if a system that uses less phonetic knowledge performs better than one using more phonetic knowledge, the former is preferred. However, this answer does not provide any insight into why the transfer of phonetic knowledge to speech technology is so difficult.

Part of the answer is certainly related to the fact that phonetics and speech technology are essentially two different worlds. This should not be underestimated. At the universities of most countries, research and education in phonetics and speech technology are conducted by different people in different faculties, i.e. those of linguistics and engineering. Consequently, many differences exist between these two groups of researchers: they study

different theories, acquire different practical skills, and use different jargons. To a large extent they even have different frames of reference, carry out experiments differently, etc.

These differences between the two worlds are certainly a problem, and hinder the transfer of knowledge to some extent. Interestingly, the situation in The Netherlands is quite different from that of most other countries. In some Dutch universities, research and education in phonetics and speech technology take place in the same faculty, i.e. the faculty of Arts. However, although the gap between the two worlds should thus be smaller in The Netherlands, the role of phonetic knowledge in speech technology is not noticeably larger than in other countries. So there must be other reasons that hinder the transfer of knowledge.

These reasons might be found in the different approaches used in phonetics and speech technology. Let us take a closer look at those differences. In order to make the differences clearer, a somewhat caricatured overview is presented of a classical phonetic versus a speech technology approach (see Table 1). Although in most cases the differences will not be so extreme, this comparison is useful to get an idea of the what hampers transfer of phonetic knowledge to speech technology.

```
Table 1. The classical phonetic vs. a speech technology
approach.

approach        classical phonetic  speech technology
-----------------------------------------------------------
condition       controlled          less controlled
setting         studio, lab         many places
sound quality   high                varied: noise, etc.
speech style    formal              informal, spontaneous
articulation    careful             varied: hypo- to hyperart.
database        small, balanced     large, less balanced
subjects        few                 many
processing      manual              automatic
analysis        deterministic       statistical
features        formants, LPC, etc. cepstra, (rasta-)PLP, etc.
approach        linguistic          information-theoretical
goal            knowledge, theory   applications
```

Table 1 is based on a table from a presentation I gave in Nijmegen in 1996 at a meeting of the 'Dutch Organization of Phonetic Sciences' (see http://fonsg3.let.uva.nl/FonetischeVereniging/). The presentation was entitled 'Two methods of speech research: The classical phonetic and the speech technological approach' (the original Dutch title was: "Twee

methodes van spraakonderzoek: klassiek fonetische & spraak-technologische").

Some clarification is in order here. In a prototypical classical phonetic experiment, a factorial design is used to make proper statistical analysis possible. Preferably, all cells in the factorial design are filled with the same number of observations. Care is taken to control other (known) factors, to reduce their (disturbing) effect as much as possible. Therefore, high quality sound is often used in a controlled setting (a studio), instead of e.g. spontaneous speech in a train station. For instance, in investigating lexical stress, subjects are asked to carefully pronounce contrastive pairs like "SUBject" and "subJECT" in a very controlled way (see also section 3.2).

With such (classical) phonetic experiments a great deal of phonetic knowledge has been acquired over the years. The question is whether this phonetic knowledge can be used in speech technology, and of course how.

3. USING PHONETIC KNOWLEDGE IN SPEECH TECHNOLOGY: SOME EXAMPLES

3.1 Duration model

Within the European ESPRIT project POLYGLOT, an isolated word recognition (IWR) system that had been originally developed for Italian (Billi et al., 1989), had to be localized to a number of other European languages, including Dutch. This system made use of some phonetic knowledge, among others a duration model. This duration model contained statistics on the duration of phonetic units, which essentially were classes of phones with similar properties (Strik and Konst, 1992). What was needed for the IWR system were the conditional probabilities of a certain duration given the class of phones: P(duration | class of phones).

In order to obtain this duration model for Dutch, we first had a look at the literature. We found that research on this topic had indeed been carried out (e.g. Nooteboom, 1972; Nooteboom and Slis, 1972; Koopmans van Beinum, 1980). Although part of the phonetic knowledge in these publications was quantitative, it was not sufficient to derive the duration model needed, mainly for the following two reasons. [1] The phonetic knowledge was not complete: data on vowels were present, but not on consonants. [2] The knowledge was not in the correct format: It was specified in terms of means (and, sometimes, standard deviations), while for the duration model a probability density function was needed. Since the required duration model could not be derived from existing phonetic

knowledge, we decided to use a data-driven method to obtain it (Strik and Konst, 1992). Isolated words were recorded, labeled, segmented and on the basis of these data a duration model was calculated.

Table 2: Duration of short vowels (SAMPA notation is used in this article).

phone	Strik and Konst		Koopmans van Beinum	
	Mean	SD	Mean	SD
I	91	24	124	24
U	98	29	108	22
O	99	25	108	23
A	103	23	120	19
i	105	37	140	12
E	107	25	124	15
u	111	25	150	17
y	140	55	136	23

In table 2 mean and standard deviation values of the durations of some short vowels are given. These values are compared to the measurements of Koopmans van Beinum (1980): mean and standard deviation values of five measurements of the duration of vowels in isolated monosyllabic words spoken by an untrained male speaker. Of the various conditions for which vowel lengths were measured by Koopmans van Beinum (1980), this was the condition that most closely matched the isolated word condition of this ASR system. The average values found by Koopmans van Beinum (1980) are larger (except for /y/), which is not surprising since she only used monosyllabic words and the database in Strik and Konst (1992) contains both monosyllabic and polysyllabic words.

These findings make clear that, besides the two reasons already mentioned above in this section, there is another reason why it is problematic to transfer existing knowledge to a speech technology application: the existing phonetic knowledge is based on data that is not representative of the speech that will be used in the application. In this case: only monosyllabic words (in the phonetic experiment) versus monosyllabic and polysyllabic words (for the ASR). Furthermore, it is questionable whether the five measurements of a single male subject are representative of the whole population.

Although duration has been studied a great deal, and thus substantial knowledge about duration should be present, duration is hardly used in current ASR systems. However, it is often used in TTS systems, in which case also data-driven methods are used to derive the duration models (see e.g. Lopez and Hernandez, 1995; van Santen, Sproat, Olive, and Hirschberg, 1996).

3.2   Lexical stress

Phonetic research has shown that there are systematic acoustic differences between (the vowels in) syllables with and without lexical stress (see e.g. van Bergem, 1993; Sluijter and van Heuven, 1996). It has been observed that stressed syllables have a longer duration, higher energy, less spectral tilt, and a different vowel quality (i.e. more like a full vowel than like a reduced vowel). Given these systematic differences, one would expect that this knowledge could be used to improve the performance of ASR systems. A couple of years ago, this issue was investigated at our department. The procedure that was followed is described below.

First, different models for vowels in stressed and unstressed syllables were trained (Kuijk, Heuvel and Boves, 1996). The recognition results on independent test-sets showed no clear improvements in the performance of the ASR system. Nevertheless, the resulting models for the vowels in stressed and unstressed condition were different, since swapping the models (i.e. using models trained on stressed vowels to recognize unstressed ones, and vice versa) led to higher error rates (Kuijk, Heuvel and Boves, 1996).

Other attempts at making use of lexical stress in ASR have led to varying results. Adda-Decker and Adda (1992) found improvements for a French corpus, but not for the American English DARPA-RM corpus. Hieronymus et al. (1992) reported a 65% reduction in word error rate and a 45% reduction in sentence error rate. And recently, Wang and Seneff (2001) obtained a small but significant relative improvement of 5.5% in word error rate.

In order to get a better understanding of why knowledge on lexical stress cannot easily be applied to obtain substantial improvements in the performance of ASR systems, a more detailed study was carried out (Kuijk and Boves, 1999). Measurements of various phonetic features were made for 5000 phonetically rich sentences from the Dutch POLYPHONE corpus. A comparison was made of the phonetic feature values in stressed and unstressed condition. For instance, the distributions of the durations of the vowels /9y/ and /a:/ are shown in Figure 1. These distributions clearly illustrate the two extremes that were observed in comparing the distributions of the phonetic features: from almost no difference to large differences. Significant differences were found in the majority of cases, reflecting that systematic acoustic differences are present. Such differences might be useful to classify vowels as either stressed or unstressed. This possibility was verified in a number of tests, using both raw and normalized phonetic features. The results for correct classification of stress varied from 57.16% to 76.05%, for the various vowels. In conclusion, although there are systematic

and significant differences between vowels in stressed and unstressed syllables, the resulting classification scores are not very high.

Insert Figure 1 about here.

In trying to understand these results one should keep in mind that even if the differences are significant the overlap can be considerable. This is the case for almost all distributions of the phonetic features in this experiment. Other (classical) phonetic experiments have generally yielded smaller overlaps, because in these experiments the effects of other factors were reduced as much as possible by using a controlled setting: e.g. stress-minimal pairs (like "SUBject" versus "subJECT") were carefully pronounced in identical phonetic contexts. However, in real life the effects of other factors are present and cannot be ruled out. The consequences are that effects of lexical stress which are present in 'lab speech' are blurred by the effects of other factors in 'real speech'.

In this case the main reason why phonetic knowledge fails to improve ASR performance is that this knowledge is based on carefully controlled speech that is not representative of the speech encountered in everyday life. In addition, one should realize that knowledge about lexical stress is rather qualitative in nature: although in many publications on this topic measurement data are presented, it is obvious that there is no ready-made 'lexical stress model' that can be plugged directly into an ASR system.

3.3 Pronunciation variation modeling for ASR

A well-known problem in ASR is pronunciation variation. Various methods to model pronunciation variation at the lexical level have been investigated, in order to enhance the performance of ASR systems (for an overview see Strik & Cucchiarini, 1999; and Strik, 2001).

Since knowledge about pronunciation variation is available in the literature, it seems logical to employ this knowledge in ASR systems. In general, knowledge on pronunciation variation is qualitative and is often expressed in the form of rewrite rules. With these rewrite rules, pronunciation variants can be generated and subsequently added to the lexicon. In this way the performance of an ASR system can be improved (Kessens, Wester and Strik, 1999). This can, for instance, be observed in Figure 2 (taken from Kessens, to appear; and Kessens, Cucchiarini and Strik, to appear). The upper curve (labeled 'Lexicon') shows the word error rates (WERs) when pronunciation variants are added to the lexicon. When going from 1 to about 1.5 variants, the WER becomes lower. However, when more variants are added the WER goes up again and even reaches levels that are

much higher than that of the baseline system. Probably this is because the confusability in the lexicon becomes too large if too many variants are added to the lexicon.

Insert Figure 2 about here.

Somewhat better results can be obtained if the acoustic models are retrained (see curve 'HMMs' in Figure 2). The best results are obtained if the probabilities of the variants are taken into account in the language model (LM) of the ASR (see curve 'LM' in Figure 2), but these probabilities are not readily available in the literature. A possibility then is to use a knowledge-based approach: start with the known rules, and calculate the probabilities of these rules, or the pronunciation variants generated with these rules, on the basis of a speech corpus. Such a knowledge-based approach has often been resorted to (see references in Strik & Cucchiarini, 1999; Strik, 2001). Although in this approach knowledge is used (i.e. rules), it is important to notice that the probabilities (which are essential) have to be derived from data, preferably substantial amounts of representative data.

Another possibility is to use a data-driven approach in which both the rules and their probabilities are derived from data (see references in Strik & Cucchiarini, 1999; Strik, 2001). The data-driven method generally comes up with known rules, which provide a description of the connected speech processes that are present in the speech corpus under investigation, plus many new rules which were not yet known (Wester, Kessens and Strik, 1998; Kessens, Wester and Strik, 2000; Kessens, Strik and Cucchiarini, 2000). Consequently, error rates obtained with data-driven approaches are usually lower (Kessens, Strik and Cucchiarini, 2000; Wester and Fosler-Lussier, 2000; Wester, 2002).

To sum up, linguistic knowledge can be used to model pronunciation variation for ASR. Simply using the knowledge as is (i.e. in the form of rewrite rules) can enhance the performance of an ASR system. However, even lower error rates can be obtained if probabilities of the rules (or variants) are derived from recorded and labeled data. And, if the data are available, one can probably best resort to data-driven methods, since they generally yield the best results (and in this way new rules can be learned). In this case the main obstacle to using existing knowledge is that the knowledge is not complete, and that it is not quantitative in nature.

3.4   **Prosodic models and language models**

Besides the three examples taken from our own research, which were presented above, many more examples can be found in the literature, of which two are mentioned here.

The first example concerns prosodic models. Despite the enormous amount of phonetic/linguistic research on prosody that has been carried out, prosodic models are rarely used in ASR systems. Some reasons why this is the case are presented in Batliner et al. (2001). An important reason is that in most prosodic models too much emphasis is put on intonation (pitch, F0), and thus these models are not complete since prosody does not manifest itself in terms of F0 alone. In fact, F0 cannot even be varied in isolation without affecting other acoustic properties of the speech signal like spectral tilt and intensity (Strik, 1994).

The last example we want to mention is that of language models used in ASR. Generally, N-grams are used, which are simple stochastic models that can easily be integrated into ASR systems. Although syntax has been studied extensively, and many grammars have been proposed and developed over the years, so far (classical) linguistics has not provided a viable alternative to the N-grams (see e.g. Rosenberg, 2000). To a large extent this is due to the fact that this branch of linguistics has mainly been engaged with written language and not with spoken language. When we speak we often produce utterances that are not grammatically correct. And, to make things even more difficult, we also produce many disfluencies. Some studies have focused explicitly on spoken language, and tried to incorporate linguistically motivated language models in ASR systems (see the many references in Brill, Florian, Henderson, and Mangu, 1998). However, none of them succeeded in achieving substantial improvements over the N-gram.

4.   **DISCUSSION**

What impedes the transfer of phonetic knowledge to speech technology? First of all, it is clear that in order to be used in speech technology, phonetic knowledge has to be incorporated into the computational framework of a speech technology system. There are several factors that make this incorporation problematic for much of the existing phonetic knowledge, which mainly has been obtained through controlled (classical) phonetic experiments. Some of these problems were illustrated in the examples in the previous section. To summarize, the main problems that emerged from the examples in the previous section are that the knowledge is based on small amounts of 'lab speech' and therefore does not generalize to

'more realistic speech', that the knowledge is not complete, and finally that it is not quantified at all or not quantified in the right format. In other words, phonetics does not provide ready-made quantitative models that can be plugged directly into a system.

These quantitative models can be derived on the basis of the large speech corpora that are available nowadays, with knowledge-based or data-driven methods, or combinations of these two types of methods. If the existing knowledge is not complete, as is often the case, then it is probably best to use data-driven approaches. Initial ideas about phonetic phenomena could come from (controlled) phonetic experiments. Subsequently, these ideas should be tested and quantified using large speech corpora. In this way knowledge can be acquired which can more easily be integrated in speech technology.

Of course, one could wonder whether more phonetic knowledge should be used in speech technology at all. A reason for doing so, which is often mentioned in this context, is that humans perform better than machines on many tasks. However, should an ASR system have ears and a basilar membrane, or should a TTS system have a larynx and a tongue (see also Hermansky, 1998)? No! We do not need to make replicas of (parts of) humans. Another extreme is not using phonetic knowledge at all. In this case, e.g., a speech corpus is seen as just a bunch of CDs, files or signals for which the word error rate or another error criterion should be minimized. It is obvious that to improve speech technology systems using phonetic/linguistic knowledge can be useful. A good example to illustrate this is knowledge about human auditory perception, which was applied to improve ASR performance (Hermansky, 1998). Nowadays, perception-based features (such as Mel, Bark or PLP) are used in most ASR systems.

Another reason why phonetic/linguistic knowledge could be useful is the following. Progress with current ASR and TTS techniques has been steady but slow. This could indicate that the ceiling of the performance for current techniques has almost been reached. Therefore, the best way to proceed is probably not to put only a lot of extra effort into fine-tuning the existing techniques, but instead to study some innovative approaches too. And although the complete solution cannot be found in current phonetic/linguistic knowledge, this knowledge should certainly be taken into consideration while searching for new techniques for better systems.

Speech production is a process that is constrained at various levels: acoustic, phonetic, phonological, lexical, syntactic, and semantic. Knowledge about these constraints could be of benefit to speech technology. To this end, these constraints have to be identified, described (in a certain formalism), and quantified in such a way that they can be incorporated in a

complete computational framework. The best results in ASR so far have been obtained with a stochastic computational framework, so it is likely that the description and the quantification of the constraints should be of a stochastic nature. These constraints can be described at different levels (multiple tiers). Information missing on one level can then be derived from, or complemented with, information from other levels.

So far, the emphasis in ASR and TTS has been on word recognition and synthesis. Since speech is mainly used for communication, the focus of research should shift more towards understanding and expressing messages, i.e. speech-to-concept and concept-to-speech (see also Zue, 1991, and Furui, 2000). This does not only require phonetic knowledge, it also requires knowledge from many other disciplines. Between these worlds even more gaps will exist. For instance, psycholinguistic models often use the correct phone(me) sequences as input, and many natural language processing models take the correct word sequences as input. In ASR the correct phone(me) and word sequences are not readily available. Therefore, these models cannot be directly integrated with ASR systems. In order to integrate models from different disciplines, a lot of gaps still have to be bridged.

5.   ACKNOWLEDGEMENTS

6.   REFERENCES

Adda-Decker, M. and Adda, G. (1992)
Experiments on stress-dependent phone modeling for continuous speech recognition.
In: Proceedings of ICASSP-92, San Francisco, USA, pp. 561-564.

Bergem, D.R. van (1993)
Acoustic vowel reduction as a function of sentence accent, word stress, and word class.
Speech Communication 12, pp. 1-23.

Batliner, A., Möbius, B., Möhler, G., Schweitzer, A., Nöth, E. (2001)
Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground.
In: Proceedings of Eurospeech-2001, Aalborg, Denmark, pp. 2285-2288.

Billi, R., Arman, G., Cericola D., Massia, G., Mollo, M., Tafini, F., Varese, G., Vittorelli, V. (1989)
A PC-based large vocabulary isolated word speech recognition system.
In: Proceedings of Eurospeech-89, Paris, France, pp. 157-160.

Brill, E., Florian, R., Henderson, J., and Mangu, L. (1998)
Beyond N-Grams: Can Linguistic Sophistication Improve Language Modeling?
In: Proceedings of COLING/ACL 1998 Conference, Montreal, Canada, pp. 186-190.

Furui, S. (2000)
Steps towards natural human-machine communication in the 21st century.
In: Proceedings of COST249 Workshop on Voice Operated Telecom Services, Ghent, Belgium, pp. 17-24.

Hermansky, H. (1998)
Should recognizers have ears?
Speech Communication, 25, 3-28.

Hieronymus, J.L., McKelvie, D., McInness, F.R. (1992)
Use of acoustic sentence level and lexical stress in HMM speech recognition.
In: Proceedings of ICASSP-92, San Francisco, USA, pp. 225-229.

Kessens, J.M. (to appear)
Making a difference: On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition.
Ph.D. dissertation, University of Nijmegen, the Netherlands.

Kessens, J.M., Cucchiarini, C., Strik, H. (to appear)
A data-driven method for modeling pronunciation variation.
Submitted to Speech Communication

Kessens, J.M., Strik, H., Cucchiarini, C. (2000)
A bottom-up method for obtaining information about pronunciation variation.
In: Proceedings of ICSLP-2000, Beijing, China, pp. 274-277.

Kessens, J.M., Wester, M., and Strik, H. (1999)
Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation.
Speech Communication 29, pp. 193-207.

Kessens, J.M., Wester, M., and Strik, H. (2000).
Automatic Detection and Verification of Dutch Phonological Rules.
In: PHONUS5, Proceedings of the "Workshop on Phonetics and Phonology in ASR", Saarbrücken, Germany, pp. 117-128.

Koopmans van Beinum, F.J. (1980)
Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions.
Ph.D. thesis, University of Amsterdam, the Netherlands.

Kuijk, D. van, Boves, L. (1999)
Acoustic characteristics of lexical stress in continuous telephone speech.
Speech Communication, 27, pp. 95-111.

Kuijk, D. van, Heuvel, H. van den, Boves, L. (1996)
Using lexical stress in continuous speech recognition for Dutch.
Proceedings ICSLP-96, Philadelphia, USA, pp. 1736-1739.

Lopez-Gonzalo, E. and Hernandez-Gomez, L.A. (1995)
Automatic Data-Driven Prosodic for Text to Speech.
In: Proc. Eurospeech-95, Madrid, Spain, pp. 585-588.

Nooteboom, S.G. (1972)
Production and perception of vowel duration: a study of durational properties of vowels in
    Dutch.
Ph.D. thesis, University of Utrecht, the Netherlands.

Nooteboom, S.G. and Slis, I.H. (1972)
The phonetic feature of vowel length in Dutch.
Language and Speech, Vol. 15, pp. 301-316.

Pols, L.C.W. (1999)
Flexible, robust, and efficient human speech processing versus
present-day speech technology.
In: Proceedings of ICPhS-99, San Fransisco, USA, pp. 9-16.

Rosenfeld, R. (2000)
Two decades of Statistical Language Modeling: Where Do We Go From Here?
Proceedings of the IEEE, 88(8), August 2000.

Santen, J. van, Sproat R., Olive J., and Hirschberg, J. (1996)
Progress in speech synthesis.
Springer Verlag, New York, 1996.

Sluijter, A.M.C. and Heuven, V.J. van (1996)
Spectral balance as an acoustic correlate of linguistic stress.
J. Acoust. Soc. Amer. 100, pp. 2471-2485

Stevens, K.N. (1960)
Toward a model for speech recognition.
J. Acoust. Soc. Amer. 32, pp. 47-55.

Strik, H. (1994)
Physiological control and behaviour of the voice source in the production of prosody.
Ph.D. dissertation, University of Nijmegen, the Netherlands.

Strik, H. (2001).
Pronunciation adaptation at the lexical level.
In: Proceedings of the ITRW 'Adaptation Methods for Speech Recognition', Sophia-
    Antopolis, France, pp. 123-130.

Strik, H., Cucchiarini, C. (1999)
Modeling pronunciation variation for ASR: a survey of the literature.
Speech Communication 29, pp. 225-246.

Strik, H. and E. Konst (1992)
A Duration Model for Phonetic Units in Isolated Dutch Words.
In: AFN-Proceedings, Universiteit of Nijmegen, Vol. 15, pp. 71-78.

Wang, C. and Seneff, S. (2001)
Lexical stress modeling for improved speech recognition of spontaneous telephone speech in
    the JUPITER domain.
In: Proceedings of Eurospeech-2001, Aalborg, Denmark, pp. 2761-2764.

Wester, M. (2002)
Pronunciation variation modeling for Dutch automatic speech recognition.
Ph.D. dissertation, University of Nijmegen, the Netherlands.

Wester, M., Fosler-Lussier, E. (2000)
A comparison of data-derived and knowlegde-based modeling of pronunciation variation.
In: Proceedings of ICSLP-2000, Beijing, China, pp. 270-273.

Wester, M., Kessens, J.M., Strik, H. (1998)
Modeling pronunciation variation for a Dutch CSR: testing three methods.
In: Proceedings of ICSLP-98, Sydney, Australia, pp. 2535-2538.

Zue, V. (1983)
The Use of Phonetic Rules in Automatic Speech Recognition.
Speech Communication, Vol. 2, pp. 181-186.

Zue, V. (1991)
From Signals to Symbols to Meaning: On Machine Understanding of Spoken Language.
In: Proc. of ICPhS-99, Aix-en-Provence, France, pp. 74-83.

Figure 1. Distributions of the durations of the vowels
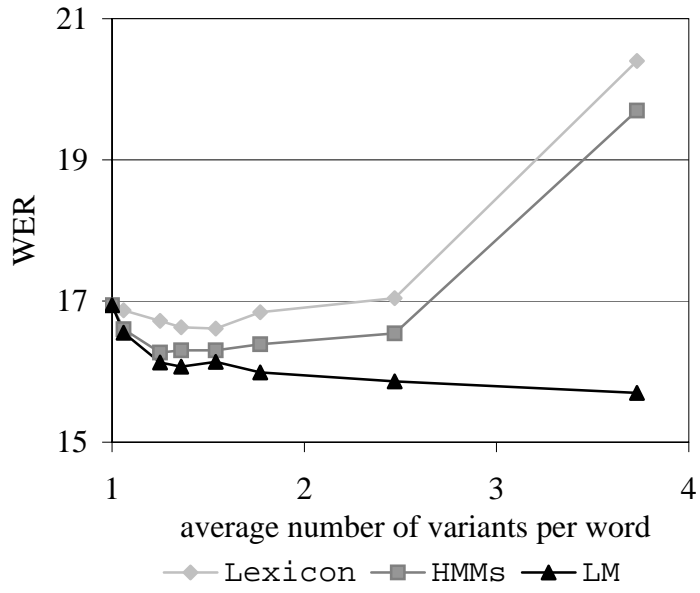/9y/ and /a:/ (solid line: stressed condition, dotted
line: unstressed condition).

Figure 2: WERs for the different testing conditions