Proceedings of the Workshop on SPeech Recognition As pAttern Classification

Edited by
Roel Smits, John Kingston, Terrance M. Nearey and Rian Zondervan

Max-Planck-Institute for Psycholinguistics
Nijmegen, The Netherlands

(2001)

vowel main effect of with V_k; $\phi_i(C_k)$ represents the consonant main effect and $[\theta\phi]_i(V_k,C_k)$ represents a vowel-by-consonant interaction effect that relates to a specific VC diphone (rhyme) combination, (e.g., /-ɑd/ in /hɑd/).

A similar decomposition can take place for the bias elements $\{b_k\}$

$$b_k = \beta(V_k) + \delta(C_k) + [\beta\delta](V_k,C_k) \qquad (4)$$

Here, $\beta(V_k)$ represent a vowel-main effect bias, $\delta(C_k)$ a consonant-main effect biases, and $[\beta\delta](V_k,C_k)$ a diphone interaction bias.

Nearey [15,16,17] showed that excellent fits to data sets from [34,15,13] could be obtained when all stimulus-tuned diphone interaction terms, corresponding to $\{[\theta\phi]_i(V_k,C_k)]\}$ of Eq. (3), were eliminated. Thus, the $\{a_k\}$ of equation (2) could be decomposed into *additive phoneme-based* components. For the datasets of [34], diphone-bias terms $\{[\beta\delta](V_k,C_k)\}$ were clearly required. For the others, however, they could be eliminated. The models described above -- that is, those with or without diphone biases, but always without stimulus-tuned diphone effects -- will be referred to as *segmental LR models*.

These findings, together with results from the intelligibility of naturally spoken words and nonsense syllables in noise [1,16,20] provide important support for a version of Mermelstein's [14] hypothesis: that the processing of *all* syllables and words can be understood in terms of the processing of phonemes. Specifically, *stimulus information* is processed by phoneme-level elements, supplemented in some cases by *stimulus-independent* bias effects associated with higher-level phonological constructs and lexical status [15,16,17,18,20].

3. RELATIONSHIPS OF SEGMENTAL LR TO OTHER FRAMEWORKS

3.1. Segmental LR and Massaro's FLMP

Formally, segmental LR and FLMP models of Massaro and colleagues [13] have many similarities. There are a few apparent differences. One is that FLMP involves discrete-level coding of stimulus effects, while (thus far) segmental LRs have treated such cues as continuous covariates. However, LR models can also code stimulus-level differences as discrete contrasts, resulting in a version of standard log-linear modeling [7,17].

Another apparent difference is that in FLMP stimulus properties typically modulate *binary* phonological *feature* oppositions, while segmental LR deals with phoneme-level contrasts, including multi-category vowel distinctions [15]. However, as Benki [4] has shown, LR methods of [17] can be extended to study the phonological feature compositionality of phonemes. A final technical difference is that FLMPs minimize an rms error criterion between fitted and observed probabilities, while segmental LR models use a maximum (quasi-) likelihood criterion (minimum G^2), assuming multinomial (-like) distribution of response errors. Despite this, FLMP and LR models are sufficiently similar that, mathematically at least, they can often be considered notational variants.

3.2. "Secondary cues" and cue sharing

A more profound difference between LR and FLMP involves the philosophy of "cue sharing". This difference results not from the core mathematical forms, but from side conditions on how stimulus variables relate to categories. In Massaro and colleagues' approach, each stimulus dimension (cue type) is associated with one and only one phonological contrast. Thus in [13] study of /bra, bla, dra, dla/ syllable cues, FLMP model coefficients associated with /l-r/ distinction relate only to F3-cue values, while F2 cues affect only /b-d/ choices. (But see [16] for a re-analysis of MC83's experiment *with* cue sharing.)

In Nearey's segmental LR models, Mermelstein's approach [14] to cue sharing is adopted. For example, a cue such as F1 of the steady state of a vocalic microsegment of a VC syllable may contribute *simultaneously* and *independently* to both vowel choice and the following consonant choice. Roughly speaking, F1 can be considered to be a "primary" cue to vowel height while simultaneously serving as a "secondary" cue. Analyses in [14,17,15,30] show *prima facie* evidence for such cue-sharing behavior. This issue will be revisited below in connection with the accommodation of context effects in ASR models.

3.3. Segmental LR and Smits' HICAT

Smits [30,31,32] has developed a new hierarchical classification model that can cover close approximations to segmental LR models as a special case. HICAT also allows for more complex stimulus-response patterns that would require inclusion of some "contraband" (from the perspective of *segmental* LR) diphone-tuned stimulus terms in a LR framework. Smits has demonstrated that HICAT appears to provide better matches to some syllable-choice data than do segmental LRs. For the most part, these improvements are fairly small, but Smits has provided elegant and detailed theoretical discussion exploring some possible production patterns and their near-optimal decoding by variations of HICAT models. Further developments of this model should be watched carefully. An interesting question is whether cases can be found where HICAT solutions work well but where segmental LRs fail more dramatically. If such cases *cannot* be discovered, then the question can be posed why are real perceptual experiments so well approximated by segmental LR.

In any event, segmental LR models have proven to be powerful tools in investigating cue-to-response relationships. They also have the practical advantage of

over all possible stimulus values; by default, $P(k)$ is set to $1/K$, where K is the number of phonemes in the choice set); and $Gau(\bullet)$ is the multivariate Gaussian (normal) probability density function.

By plugging in values for means and covariance matrices estimated from the measured acoustic data, a researcher can predict in advance listener's responses to any stimulus, including newly synthesized ones.

This method makes many assumptions that are probably too strong and are difficult to justify psychologically. Luce choice theory is debatable in any choice situation and is certainly so if applied directly to *a posteriori* probabilities from production data. However, NAPP still has considerable appeal as an "engineering approximation" to the speech perception problem. If the distributional assumptions about the acoustic data are correct, then there is no better criterion than Eq. (1), on which an ideal observer could base category choice [29]. Furthermore, NAPP has other intuitively appealing properties. Assuming $[P(j) = P(k)$ for all $k, j]$ (equal priors), when a stimulus is much closer (in Mahalanobis distance) to the mean of one category than to any other, its predicted probability is near 100%. When a stimulus is equally close to two categories, but remote from the others, its probability will be about 50% for the nearby categories, but near zero for the others.

Despite the naiveté of NAPP, it has proven quite useful is the study of issues to vowel perception. Nearey and Assmann [22] first used the full-blown technique to compare listeners' perception of "silent center" isolated vowels to a default linear discriminant analysis model (with $P(k)$ and S_k assumed equal over all the vowels). Using the NAPP formulation as a tool to investigate alternate static and dynamic cue representations, Nearey and Assmann found that a cue-representation using formant measurements from both a nucleus and an offglide section appeared to provide a good account of listeners' perception of isolated vowels.

Subsequent research using NAPP models, in collaboration with Andruski [2], and with Hillenbrand and his colleagues, has continued to show good (though by no means perfect) correspondence with listeners' identification of natural and synthetic vowels. For example, considering the data discussed in Hillenbrand and Nearey [9], after training on a disjoint sample of some 1200 vowel tokens from males, females and children, *a posteriori probabilities* calculated with the "frozen" model for an independent test set of 300 vowel tokens (thus producing true cross-validation predictions) shows an rms prediction error of about 8.2%. Quadratic discriminant analysis, with distinct S_k per vowel, shows an even lower prediction error rate of about 6.8%. Correlations between predicted and observed confusion matrices are extremely high, above .95 in all cases, largely because both listeners' responses and the cross-validation predictions are concentrated on the original "correct " class. More detailed comparisons are reported in [9].

2.2. Logistic Discrimination and Logistic Regression

If S_k are assumed to be equal for all phonemes, it is well known [29] that the *a posteriori* probability functions of Eq. (1) can be also be generated by functions of the form

$$p(k \mid x) = \frac{\exp(x^T a_k + b_k)}{\sum_{k' \in \{1 \ldots K\}} \exp(x^T a_{k'} + b_{k'})} \qquad (2)$$

where, x^T is a row vector of cue values for a stimulus, a_k is a column vector of coefficients, and b_k is a bias (constant) for each phoneme k. If an augmented stimulus vector z is defined by the adjoining to x the squares and unique cross products of the values of he original stimulus dimensions, and z is substituted for x in Eq. (2), then these equations cover *a posteriori* probabilities arising from the general case of Eq. (1), with no restriction on the S_k

There can be advantages in optimizing coefficients in Eq. (3) directly [rather than deriving it indirectly through the class-conditional statistics of Eq. (1)], because this fits a *discriminant* function to the data [29]. A maximum (quasi-) likelihood criterion is used to minimize the discrepancy between the predicted $p(k|x)$ and observed probabilities $o(k|x)$ for each stimuli x for all categories k. In logistic discrimination, the observed probability is scored as 1.0 when k is the original category for a given stimulus and 0.0 otherwise. If, instead, $o(k|x)$, is taken as the observed response probability for a subject choosing category k for stimulus vector x, then a similar maximum likelihood procedure may be applied in a (polytomous) logistic regression (LR) of perceptual response probabilities on the stimuli. We have applied the latter technique extensively in our laboratories.

2.3. Logistic regression of phoneme strings: segmental LR models

If k in Eq. (2) is extended to range over phonological units "larger" than phonemes, then syllable probabilities can be modeled. In that case, it has proven fruitful to factor the coefficient sets $\{a_k\}$ and $\{b_k\}$ ANOVA-style, as exemplified in [7, chap. 1]. For example Nearey [15] studied listeners' categorization of a 4-dimensional stimulus set with 972 stimuli spanning 10 English /hVC/ syllables. The factorization of the stimulus coefficients $\{a_k\}$ may be represented as:

$$a_{ki} = \theta_i(V_k) + \phi_i(C_k) + [\theta\phi]_i(V_k, C_k) \qquad (3)$$

Here, a_{ki} represents the coefficient of the i-th stimulus variable of the k-th syllable. V_k indexes the vowel category and C_k, the consonant category of syllable k; thus, if the syllable in question is /hɑd/, then V_k is the vowel code number assigned to /ɑ/. The Greek letter terms represent an ANOVA-like decomposition of a_{ki} into main effects and interactions. $\theta_i(V_k)$ represents the

strategy. Second, SUMMIT provides an existence proof of a method that potentially allows integration of (pre-HMM era) expert-system "segment-and-classify" strategies with modern optimization techniques. The second point deserves elaboration.

The SUMMIT system involves a two-stage recognition process that can be described roughly as follows (although SUMMIT researchers do not typically describe it quite this way): The first stage can be viewed as providing a set of alternate acoustic segmentations of speech into microsegments, corresponding to key signal elements related to phones, such as stop closures, burst releases, vocalic regions, etc. A second stage system (roughly speaking) scores a set of phonetic hypotheses, conditional on the alternate microsegmentations, using static (fixed length input vector) statistical pattern recognizers that may use a variety of measurements. There are some other important details and refinements provided by the MIT group that should be considered carefully in application of this method to perceptual models. However, the general approach has shown itself to be very flexible.

My proposal, following some initial lines of research described by [23] for stop consonants, is that the second stage of such a process is amenable to perceptually-oriented engineering, using theoretically motivated cues and scoring them with statistical pattern-recognizers related, such as NAPP, LR, FLMP or HICAT. Research by the MIT group (e.g. [8]) has shown that a variety of heterogeneous measures can be used successfully in recognition strategies. Furthermore, although performance enhancements are possible with careful tuning, the general approach provides respectable results with respect to a variety of cue choices.

5.2. Cue sharing in ASR

A key area to explore in a second pass recognizer is the kind of cue sharing between neighboring phonemes discussed in section 3 above. There are many different methods, including stochastic segment models, triphone context models, deterministically-trended HMMs, and hybrid HMM-neural networks that attempt to better accommodate within- and/or between- phone correlation of cues. None of these deals with the issue of physical context dependency in anything like the way suggested by the results of LR segmental models.

In its second pass, SUMMIT appears to use only features extracted from microsegments that are subsumed exclusively by a single phoneme. In essence, the acoustic microsegments are being treated as "immediate constituents" of phones. In practice, as in many "standard" HMM schemes, some of the extracted cues are based on relatively long temporal regions, and a cue used in second pass scoring of a stop consonant subsuming a release-burst micro segment may contain information from portions of a waveform covered by a following vocalic microsegment subsumed by a vowel phone. (Indeed, SUMMIT's *boundary features* are designed specifically to allow spanning of microsegments). Such wide-time cues (including such entities as delta cepstra) are well known to help in speech recognition dramatically, even though they often fit at best awkwardly into the theoretical framework in which they are embedded.

A two-pass scheme like that of SUMMIT could be expanded to allow a more radical cue-sharing, like that associated with "secondary cues" in [15,17], to be exploited. For example, suppose a tentative microsegmentation includes a silence following a vocalic region, compatible with phone hypotheses of the form VC. Although the C phone may be synchronized with silent period, there is nothing to keep a second pass scoring algorithm from evaluating consonant hypotheses in light of cues read off extracted from the preceding vocoid, including its entire duration.

It is not obvious whether such contingent cue scoring can be integrated into a coherent stochastic modeling framework, but I suspect it can. Even if such statistical coherence should prove elusive, I think the strategy deserves investigation. Despite its foundations in impeccable Bayesian theory, violations of assumptions and heuristic shortcuts abound in ASR even in the most orthodox HMM frameworks, because they improve recognition scores [5]. Finding ways to incorporate what we think are important insights from speech perception into ASR seems well worth the effort with or without the benefit of a statistical imprimatur.

The ASR, lexical access and speech perception communities will all survive whether or not we find compelling ways to integrate our knowledge and methods. But life will be better for all if we do. It seems likely, unless the exigencies within the field described in [5] have changed dramatically, that champions of this cause will arise spontaneously within the ASR community. Lexical access researchers and we phoneticians will have to meet them at least half way.

6. ACKNOWLEDGEMENTS

This work was supported by SSHRC. Send correspondence to T. Nearey, Linguistics, 4-32 Assiniboia Hall, Edmonton, AB, T6G 2E7, CANADA. E-mail: t.nearey@ualberta.ca.

7 REFERENCES

[1] Allen, J. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, **2**, 567-577.

[2] Andruski, J. & Nearey, T. M. (1992). On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. *Journal of the Acoustical Society of America*, **91**, 390-410.

[3] Assmann, P. F. & Summerfield, A. Q. (1989). Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency. *Journal of the Acoustical Society of America*, **85**, 327-338.

scalability over FLMP and HICAT. LR models readily handle multiple categories and multiple cues, enabling their integration in more ambitious pattern recognition tasks.

4. BRIDGING TO LEXICAL ACCESS AND ASR: MODELING THE PERCEPTION OF PHONETIC STRINGS

4.1. Modeling lexical access

Norris and colleagues [24,25] have proposed models of the lexical access process based on phoneme-level inputs that appear to cover many aspects of human performance. However, to date all lexical access models have relied on *artificial* phonetic input, engineered by the investigators. Such input might be viewed as "an artist's conception" of the information available to a lexical access system after some initial phonetic transduction. It is unclear whether any existing lexical access models will work as advertised if they are attached to real phonetic transduction systems, i.e. to phonetic "front ends" that take waveforms as input and provide some sort of (probabilistic) phonetic transcription as an output.

A strategy that at least one group is contemplating to use front-ends from conventional ASR methods to do such a transduction. I believe that such efforts are extremely important and that they are very likely to quickly uncover any fundamental difficulties that will be encountered by interfacing any lexical access system to a waveform driven input system [18].

While many possible front ends might provide insights into the adequacy of lexical access models, at least some critical research questions rely on relatively subtle sub-phonemic differences [25]. For these problems a front end that more faithfully reflects listeners' perception will likely be required. While such a phonetically plausible front-end seems like a worthy goal in its own right, the prospect of linking speech perception research to psycholinguistic models of lexical access provides further incentive.

In my view, the path that presents the best prospects to meet this goal is for researchers in perceptual phonetics to look for guidance to our colleagues in speech technology, since they the only people on the planet who have actually achieved anything that resembles a working model of lexical access from waveforms.

4.2. Modeling the perception of variable length phonetic strings

With a few notable exceptions (e.g. [28]) almost all speech perception research has focused on simple phonetic patterns with fixed canonical form. There is no doubt that close study of well-defined problems, such as the voicing/place of initial stops in CV syllable, was the right place to start. Indeed, there no doubt remain many important details to be explored using such stimulus sets. Nonetheless, if speech perception research is to have serious impact on lexical access and on speech technology, we must find ways to expand beyond such simple cases.

Perhaps one reason there has not been more extensive research on the perception of variable length strings is that even the relatively simple cases that have been investigated have lead to very complex results. Thus, Repp [28] studied the perception of synthetic patterns that spanned response sets fitting the general template $VC_1(C_2)V$, so that the choice set could include [aba], [aga], [abga] [agba], as well as the geminates [ab.ba], [ag.ga]. Repp describes complex results that include assimilation (integration) of transition cues across the stop gap in come cases (at short gap durations when stop transitions were compatible with a single place of articulation) but dissimilation (contrast) in others. There were also complex duration and speaking rate effects for the perception of singleton versus geminate stops.

Interestingly, there has been some attempt (delayed by more than a decade) at modeling aspects of this complex behavior. Specifically, Grossberg et al. [6] have shown via simulations that their ARTPHONE neural networks can handle some of the Repp's [28] findings relating to cluster (and geminate, e.g., [bb]) versus singleton consonant choices. However, Grossberg et al.'s work does not explicitly model how place cues nor stop gap duration are assessed from the signal. Instead, much like the psycholinguists' lexical access models, the ARTPONE model uses *artificial* inputs to signal occurrence of 'b-compatible' or 'g-compatible' input. Thus, the differential assimilation and contrast effects of Repp [28] were not modeled. Additional perceptual experiments pursuing some of these issues are long overdue, as are more complete modeling efforts.

5. AN APPROACH TO INTEGRATING SPEECH PERCEPTION MODELS WITH ASR

One approach to a more complete modeling would be may be to construct a phonetically more plausible, general purpose, variable-string recognizer that can be readily customized to incorporate new hypotheses about perception. I believe the most fruitful avenue to explore is importing techniques from ASR.

5.1. The two-pass SUMMIT system

There are probably many general ASR approaches that might be modified to serve as a platform for advanced phonetic modeling. However, an approach similar to recent implementations [33,27] of the SUMMIT system strike me as the most promising way to begin to integrate perceptual modeling described earlier in this paper with speech recognition.

There are two encouraging aspects of SUMMIT. First, SUMMIT's phonetic front end appears to perform at least as well as that of any existing ASR system. This is important primarily because it shows that there is no obvious performance *disadvantage* to this kind of

TOWARDS MODELING THE PERCEPTION OF VARIABLE-LENGTH PHONETIC STRINGS

Terrance M. Nearey
University of Alberta, Edmonton, Canada

ABSTRACT

The study of speech perception and the study of lexical access are both mature fields. However, only a relatively small amount of work has even begun to bridge the gap between the two. Much speech perception work (categorization of synthetic speech continua) has focused on close study of the response probabilities of a small set of syllables in fixed canonical frames (e.g., CV). In many such experiments (e.g. [12,15,17,30,31,32]), relatively simple static pattern recognition models account well for speech perception results, including the effects of lexical status. However, much work on lexical access focuses on the time course of perception of word and non-word stimuli that vary greatly in segmental and syllabic form. Only a very limited number of experiments in the speech perception literature have dealt with stimulus sets where both the number and type of speech sounds have varied and even fewer studies have begun to deal with the modeling of response probabilities for such data. Before detailed phonetic models can be used as front ends for reasonable models of lexical access, they must address the issue of the perception of variable length phonetic strings head on. To that end, the present paper discusses initial research into a strategy for combining dynamic pattern recognition techniques from speech recognition technology with the static models used by [15].

1. INTRODUCTION

Given the relatively advanced state of psycholinguistics and speech perception, it seems remarkable that the only *working models* of lexical access from acoustic waveforms are products of the engineering technology of automatic speech recognition (henceforth ASR). Furthermore, ASR has reached its current state with little recourse to advanced knowledge from the speech or psychological sciences. Speech recognition by machines is much less accurate and robust than speech recognition by humans. A number of researchers have suggested that a better integration of rigorous scientific speech knowledge is the only likely route for substantial improvement [26,1,10].

The relative fragility of ASR together with its apparent disregard for psychophonetic speech research has led many speech perception researchers to dismiss speech technology out of hand. I will argue below that this dismissal is unwarranted. Statistical pattern recognition techniques have yielded insightful analysis of a number of speech perception experiments. The methods employed in this research so far can be described as *static* pattern recognition methods, which classify input vectors of a fixed dimension. ASR has pioneered work in extending such methods to a more general *dynamic* pattern recognition framework that is capable of dealing with variable length inputs. This framework provides the only known way to deal coherently with the phonetic transduction of signals representing words as varied as *cat* and *Saskatchewan*. It seems likely that speech perception research, if it is ever to provide a useful bridge to lexical access, has much to gain by carefully considering lessons from ASR technology. Before discussing the dynamic pattern recognition problem, I will review some applications of static pattern recognition to speech perception.

2. NAPP AND LOGISTIC REGRESSION MODELS

2.1. The Normal *a Posteriori* Probability (NAPP) Model

A standard approach to experimental phonetics is represented by [11]. In this work, researchers approached a phonological contrast of interest, voicing in stop consonants, by (i) carefully studying production patterns, (ii) hypothesizing necessary and sufficient properties specifying the contrast, (iii) synthesizing simplified speech patterns varying in those properties, and (iv) qualitatively comparing results of perception tests with the hypothesized specification of the contrasts.

In our labs, we developed a more explicit, quantitative approach to (iv) above via the normal *a posteriori* probability or NAPP model [21]. Given a hypothesis about relevant speech cues listener's performance on perception tests could be predicted *a priori*. Two key assumptions were that (a) within-phoneme cue patterns follow multivariate normal distributions, and (b) listeners' respond, using a Luce-choice rule, to outputs of phoneme-detectors that were tuned to respond with strengths proportional to the relative likelihood of each phoneme class, as characterized by normal probability density functions. Formally, the normal *a posteriori* probability function can be stated as:

$$p(k \mid x) = \frac{P(k)Gau(x, m_k, S_k)}{\sum_{k' \in \{1 \ldots K\}} P(k')Gau(x, m_{k'}, S_{k'})} \qquad (1)$$

where x is the cue vector for a given stimulus; m_k is the mean cue vector and S_k is the (variance-) covariance matrix the for the phoneme k; $P(k)$ is the prior probability for phoneme k in the language as a whole

[4] Benki, J. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics*, **29**, 1-22.
[5] Bourland, H., Hermansky, H., & Morgan, N. (1996). Towards increasing speech recognition error rates. *Speech Communication*, **18**, 205-231.
[6] Grossberg, S., Boardman, I. B., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance,* **25**, 481-503.
[7] Haberman, S. J. (1979). *Analysis of Qualitative Data, Volume 2*. New York: Academic Press.
[8] Halberstadt, A. & Glass, J. (1998). Heterogeneous Measurements and Multiple Classifiers for Speech Recognition. *Proceedings ICSLP '98*, Sydney, Australia.
[9] Hillenbrand, J. & Nearey, T. (1999). Identification of resynthesized /hVd/ utterances: Effects of Formant contour. *Journal of the Acoustical Society of America*, **105**, 3509-3523.
[10] Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Communication,* **22**, 1-15.
[11] Lisker L. & Abramson, A. (1970). The voicing dimension: some experiments in comparative phonetics. In B. Hala, M. Romportl, & P. Janota (Eds.), *Proceedings of the VIth International Congress of Phonetic Sciences* (pp. 563-567). Prague: Academia.
[12] Massaro, D. W. & Cohen, G. C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **21**, 1053-1064.
[13] Massaro, D.W. & Cohen, M.M. (1983). Phonological constraints in speech perception. *Perception & Psychophysics*, **34**, 338-348.
[14] Mermelstein, P. (1978). On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perception and Psychophysics*, **23**, 331-335.
[15] Nearey, T. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, **101**, 3241-3254.
[16] Nearey, T. (in press). On the factorability of phonological units in speech perception. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in Laboratory Phonology*. Cambridge: Cambridge University Press.
[17] Nearey, T. (1990). The segment as a unit of speech perception. *Journal of Phonetics,* **18**, 347-373.
[18] Nearey, T. (2000). Some concerns about phoneme-like inputs to MERGE. *Brain and Behavioral Sciences*, **23**, 342-343.
[19] Nearey, T. (1992). Context effects in a double-weak theory of speech perception. *Lanuage and Speech,* **35**, 153-172.
[20] Nearey, T. (in press). The factorability of phonological units in speech perception: Simulating results on speech reception in noise. In R. Smyth & A. Laubstein (Eds.), *Festschrift for Bruce L. Derwing*.
[21] Nearey, T. & Hogan, J. (1986). Phonological contrast in experimental phonetics: relating distributions of measurements in production data to perceptual categorization curves. In J. Ohala & J. Jaeger (Eds.), *Experimental Phonology* (pp. 141-161). New York: Academic Press.
[22] Nearey, T. & Assmann, P. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, **80**, 1297-1308.
[23] Nearey, T. & Kiefte, M. (1995). A multi-stage procedure for identification of acoustic micro segments in vowel+stop+vowel syllables. *Proceedings of XIII International Congress of Phonetic Sciences*, vol. 4 (pp. 304-306).
[24] Norris, D. (1994). Shortlist: a connectionist model of speech recognition. *Cognition,* **52**, 189-224.
[25] Norris, D., McQueen, J.M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, **23**, 299-325
[26] Pols, L. C. W. (1999). Flexible, robust and efficient human speech processing versus present-day speech technology. *Proceedings of the XIVth International Congress of Phonetic Sciences Sciences*, vol. 1 (pp. 9-16).
[27] Chang, J. & Glass J. (1997). Segmentation and modeling in segment-based recognition. *Proceedings Eurospeech '97* (pp. 1199-1202).
[28] Repp, B.H. (1983). Bidirectional contrast effects in the perception of VC-CV sequences. *Perception and Psychophysics*, **33**, 147-155.
[29] Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.
[30] Smits, R. (1998). A model for dependencies in phonetic categorization, *Proceedings 16th International Congress on Acoustics and 135th meeting of the Acoustical Society of America* (pp. 2005-2006). Woodbury, NY: American Institute of Physics.
[31] Smits, R. (in press). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance.*
[32] Smits, R. (in press). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. *Perception & Psychophysics.*
[33] Ström, N., Hetherington, L., Hazen, T., Sandness, E., & Glass, J. (1999). Acoustic modeling improvements in a segment-based speech recognizer. *Proceedings 1999 IEEE ASRU Workshop*, Keystone, CO.
[34] Whalen, D. (1989). Vowel and consonant judgments are not independent when cued by the same information. *Perception & Psychophysics*, **46**, 284-292.