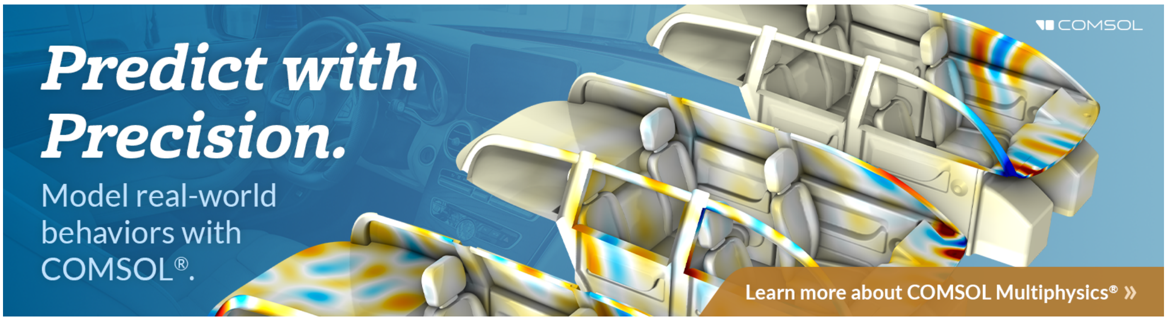
APRIL 08 2005

Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition ✓

Katrin Kirchhoff; Steven Schimmel

J. Acoust. Soc. Am. 117, 2238–2246 (2005)
https://doi.org/10.1121/1.1869172



Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition[a)]

Katrin Kirchhoff[b)] and Steven Schimmel
Department of Electrical Engineering, Box 352500, University of Washington, Seattle, Washington 98195

(Received 18 April 2004; revised 20 January 2005; accepted 21 January 2005)

Previous studies have shown that infant-directed speech ('motherese') exhibits overemphasized acoustic properties which may facilitate the acquisition of phonetic categories by infant learners. It has been suggested that the use of infant-directed data for training automatic speech recognition systems might also enhance the automatic learning and discrimination of phonetic categories. This study investigates the properties of infant-directed vs. adult-directed speech from the point of view of the statistical pattern recognition paradigm underlying automatic speech recognition. Isolated-word speech recognizers were trained on adult-directed vs. infant-directed data sets and were tested on both matched and mismatched data. Results show that recognizers trained on infant-directed speech did not always exhibit better recognition performance; however, their relative loss in performance on mismatched data was significantly less severe than that of recognizers trained on adult-directed speech and presented with infant-directed test data. An analysis of the statistical distributions of a subset of phonetic classes in both data sets showed that this pattern is caused by larger class overlaps in infant-directed speech. This finding has implications for both automatic speech recognition and theories of infant speech perception. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1869172]

PACS numbers: 43.72.Ne, 43.71.−k, 43.71.Ft [DOS] Pages: 2238–2246

I. INTRODUCTION

Many studies on infant speech perception have analyzed the acoustic-phonetic properties of infant-directed speech (also called ''motherese'' or ''parentese''). It has been observed that infant-directed speech is generally slower in tempo and exhibits increased segmental duration, longer pauses, and more pronounced pitch contours (Grieser and Kuhl, 1988; Fernald *et al.*, 1989; Kuhl *et al.*, 1997). Most importantly, it has been shown that motherese is characterized by overemphasized acoustic-phonetic contrasts: compared to adult-directed (AD) speech, vowels in infant-directed (ID) speech have mean formant values closer to the outermost points of the vowel triangle and are thus acoustically more distinct (Andruski and Kuhl, 1996; Kuhl *et al.*, 1997; Burnham *et al.*, 2002). Infants have been shown to prefer infant-directed over adult-directed speech (Cooper and Aslin, 1994; Fernald and Kuhl, 1987), and a recent study (Liu *et al.*, 2003) has found a correlation (though not necessarily a causal link) between the clarity of mothers' speech as measured by vowel space expansion and infants' phonetic discrimination abilities. It has also been demonstrated that auditory input with overemphasized acoustic-phonetic contrasts is helpful in second-language learning (Protopapas and Calhoun, 2000; Hazan and Simpson, 2000) and that it is more intelligible to individuals with hearing impairments (Picheny *et al.*, 1986; Payton *et al.*, 1994). For these reasons, it has often been suggested that ID-style speech might also be useful for training automatic speech recognition (ASR) systems. It is generally recognized that the amount of training data is of primary importance in developing good ASR systems; however, a recent study (Moore, 2003) which extrapolates state-of-the-art ASR results to larger training sets suggests that ASR performance would still be far from human performance even if training sets were increased by orders of magnitude. An alternative to simply using more data might be to use ''better'' data, e.g., more clearly articulated acoustic data. The present study explores this hypothesis by training an isolated word recognizer on adult-directed and infant-directed data sets, respectively, and analyzing its performance under identical versus mismatched test conditions.

In addition to assessing the benefit of ID-style speech for ASR we are interested in exploring the ASR framework as a computational modeling technique for infant speech perception, in particular phonetic category learning. The development of a computational model whose predictions match data gathered from human speech perception experiments could prove valuable in predicting the outcome of future experiments through simulations. A good computational model allows one to observe the effect of adjusting individual experimental variables without having to conduct expensive perceptual studies. The framework of ASR can potentially make a significant contribution towards this goal. First, ASR relies on a suite of statistical modeling techniques that have been fine-tuned to speech (e.g., perceptually inspired signal processing and temporal modeling algorithms) and are more advanced than many of the computational models of speech perception proposed in the past. Second, the ASR community has developed automated methods for collecting, preprocessing and ''cleaning'' speech data (e.g., high-accuracy automatic segmentation) which enable speech

[a)] Portions of this work were presented at the 146th Meeting of the Acoustical Society of America, Austin, Texas, November 2003.

[b)] Electronic mail: katrin@ee.washington.edu researchers to analyze large speech data sets more rapidly and efficiently.

Previous computational or numerical models of speech perception have mainly focused on predicting experimental data using techniques such as logistic regression and discriminant analysis (e.g., Hillenbrand *et al.*, 1995; Nearey, 1997) or connectionist models (Protopapas, 1999; Damper and Hanard, 2000; Guenther and Bohland, 2002). Recently, the idea of using ASR techniques for modeling human speech perception and language acquisition has attracted increased attention. Scharenborg *et al.*, (2002, 2003), for instance, have investigated the problem of word segmentation in human speech perception using an ASR decoder as a modeling tool. They showed, in a small word discrimination experiment, that the segmentations selected by an ASR decoder matched the outcome of human word segmentation experiments. In de Boer and Kuhl (2003) the learnability of infant-directed and adult-directed speech datasets was investigated using a computer model. A mixture of Gaussians was fitted to samples of the vowels /iː/, /uː/ and /ɒ/ obtained from ID and AD speech, respectively. The resulting means of the learned categories were then compared to the expected means for those vowels. Since the means learned from ID speech matched the reference values more closely, it was concluded that the ID speech has better learnability. However, this study did not evaluate the generalizability of the models by applying them to a separate test set. A fundamental concept of statistical learning, however, is the capability of the trained models to classify unseen samples that are not present in the training set.

The present study investigates the benefit of ID versus AD training data for the formation of phonetic category models which are subsequently applied to unseen instances of those categories. The ability of the trained models to generalize to novel data is measured in terms of their classification error rate. Unlike most previous work, which has focused on the analysis of individual phonetic categories (e.g., vowels), this study also investigates the recognition of entire words.

The remainder of this paper is structured as follows: in Sec. II we give a more detailed explanation of the rationale for this work. The data is described in Sec. III, and experiments and results are presented in Sec. IV. Section V concludes.

II. RATIONALE

In order to assess the potential benefit of ID-style speech for training ASR systems, it is necessary to take a closer look at its properties and view them in light of the standard phonetic classification procedures employed in state-of-the-art ASR systems.

As mentioned above, ID speech is characterized by a greater distance between vowel class means measured in formant space. In addition, the variances associated with individual classes are often greater than in adult-directed speech (Kuhl *et al.*, 1997; de Boer and Kuhl, 2003). These properties have been observed in a number of studies of vowel patterns in ID speech, in American English as well as other languages (Chinese, Swedish and Russian) (Kuhl *et al.*, 1997; Liu *et al.*, 2003). Phonetic overspecification has also been observed for stop consonants in the form of longer voice onset times in ID speech addressed to infants aged 11–14 months (Sundberg, 2001). Most of these studies have focused on monosyllabic content words with stressed vowels. In a study of both content and function words in ID speech, van de Weijer (2001) confirmed the observation of an enlarged vowel space for content words but found the opposite pattern in function words.

Studies of infant speech perception, in particular word segmentation, have shown that infants respond to statistical regularities in the speech input (e.g., Jusczyk *et al.*, 1994; Saffran *et al.*, 1996). The assumption that some form of statistical learning underlies the acquisition of phonetic categories as well (Holt *et al.*, 1998) has recently been demonstrated empirically: Maye *et al.* (2002) showed that infants exposed to speech samples from a bimodal distribution of stops (with the two peaks representing voicedness and voicelessness, respectively) were capable of discriminating new samples from the endpoints of those distributions, whereas infants exposed to a unimodal training distribution were not. The precise nature of such a learning mechanism remains as yet obscure, but it is apparent that spectral discrimination as well as temporal information integration must be involved. However, computational models of infant speech perception that explicitly incorporate temporal information processing are rare. In order to better assess the potential contributions of the ASR framework to this field, we briefly review the statistical models used in present-day ASR systems.

A. Acoustic modeling in automatic speech recognition

Current ASR systems are based on a statistical pattern recognition framework. Recognizers attempt to find the best word sequence W^* given a sequence of acoustic observations O:

$$W^* = \arg\max_W P(W|O) \tag{1}$$

The probability $P(W|O)$ is computed using Bayes' rule:

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \propto P(O|W)P(W). \tag{2}$$

The *language model* $P(W)$ gives the prior probability of a word sequence W. Since our focus is on acoustic classification and isolated word recognition, properties of the language model are of no further concern in this study. The *acoustic model* $P(O|W)$ maps acoustic observations first to intermediate phonetic classes (phones/phonemes) and eventually to words. The normalization by the probability of the observation sequence, $P(O)$, can be omitted since it is constant for all W.

The acoustic observations take the form of multidimensional acoustic feature vectors that are extracted from the speech signal at equidistant time intervals, e.g., every 10 ms. Various feature extraction schemes have been devised; the most widely used representation is based on mel-frequency cepstral coefficients (MFCCs), which mimic the nonlinear frequency resolution characteristics of human auditory perception.

The classification of acoustic feature vectors is typically performed by hidden Markov models (HMMs). An HMM λ consists of a set of states $S = s_1, s_2, \ldots, s_N$, a set of observation symbols $O = o_1, o_2, \ldots, o_M$, and probability distributions governing the transitions between states and the emission of observation symbol from states. Given these parameters, the probability of an observation sequence $O = o_1, \ldots, o_T$ given a model λ is computed as

$$P(O|\lambda) = \sum_S \prod_{t=1}^{T} a(s_t|s_{t-1}) b(o_t|s_t) \tag{3}$$

where $a(s_t|s_{t-1})$ denotes the transition probability from state s_{t-1} to s_t, and $b(o_t|s_t)$ is the probability of the tth observation at state t. Thus, the global probability of an observation sequence given an HMM is defined as the product of all transition and observation probabilities over time points $1-T$, summed over all possible state sequences.

Of particular importance for the problem studied in this paper is the way in which the state-conditional observation probabilities, $b(o_t|s_t)$ in Eq. (3), are computed. In practice, the observation probabilities (as opposed to the transition probabilities) contribute most to the final probability score and are thus primarily responsible for acoustic-phonetic classification performance.

The most widespread model used in this context is a Gaussian mixture model, which computes the probability of the continuous acoustic observation vector $\mathbf{o}$ at time t, $P(\mathbf{o}_t|s_j)$, as a weighted sum of M individual Gaussian probability density functions (mixture components):

$$p(\mathbf{o}_t|s_i(t)) = \sum_{m=1}^{M} c_{mi} \mathcal{N}(\mathbf{o}_t; \mu_{mi}, \Sigma_{mi}) \tag{4}$$

where μ_{mi} and Σ_{mi} are the mean vector and covariance matrix, respectively, of the mth mixture component of state i and c_{mi} is the mixture weight for that component. Each mixture component has the form of a Gaussian or Normal distribution:

$$\mathcal{N}(\mathbf{o}; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-(1/2)(\mathbf{o}-\mu)' \Sigma^{-1} (\mathbf{o}-\mu)}. \tag{5}$$

where μ is the mean vector and Σ the covariance matrix of the distribution and d is the dimensionality of the feature space. The Gaussian mixture parameters, as well as the transition probabilities in the HMM, are estimated from training data using the expectation-maximization algorithm (Dempster *et al.*, 1977). Typically, HMMs are constructed for individual speech units such as phones, and are concatenated to form words, according to constraints specified in a pronunciation dictionary. The Viterbi algorithm is then used to find the best path through the concatenation of models.

B. Use of infant-directed training data in speech recognition

Given this background, two hypotheses can be made regarding the use of ID-style training data in speech recognition:

(1) Phonetic classification using Gaussian mixture models is easier and more accurate when individual classes are well separated in the input space, i.e., when the class-conditional distributions overlap as little as possible. In previous experimental studies, the means of acoustic classes have been shown to be better separated in infant-directed than in adult-directed speech. While some researchers have also observed enlarged class variances in ID speech, all studies emphasize the strongly separated class means as the predominant property of ID speech. This suggests that the use of infant-directed training data might indeed be beneficial during training.

(2) On the other hand, it is well known that matched training and test conditions are of utmost importance in ASR. When test data differ from training data, e.g., due to different recording conditions, presence of noise, or unknown accents, speech recognition performance deteriorates. For this reason, it might be hypothesized that training on infant-directed speech will not be useful for developing a speech recognizer for adult-directed speech, since the training and test conditions will be too different.

In order to test these hypotheses, separate speech recognizers are trained on two different data sets consisting either only of AD speech or only ID speech. Each of them is then tested on an AD data set and an ID data set, reflecting matched versus mismatched test conditions. In order to focus entirely on the contributions of the acoustic modeling component, we perform only isolated word recognition experiments—this serves to eliminate differences in performance that could be attributed to the language model or the vocabulary size (in ID speech, for instance, speakers tend to use a much smaller vocabulary).

TABLE I. Cue words representing the vowels /iː/, /uː/, /ɒ/.

Vowel	Cue words		
/iː/	key	sheep	bead
/uː/	boot	shoe	spoon
/ɒ/	pot	top	sock

III. DATA

The data used for the experiments described in this paper were drawn from a corpus of infant-directed speech provided by the Institute for Learning and Brain Sciences (formerly the Center for Mind, Brain and Learning) at the University of Washington. The corpus consists of 64 conversations by 32 different mothers. Each mother had two conversations, one with an adult experiment facilitator, the other with her infant. The data collection was designed to elicit certain cue words during both conversations, viz. nine monosyllabic words (shown in Table I) containing the vowels /iː/, /uː/, and /ɒ/ (the most distant points in the vowel triangle). Only these words were used for the present study since they had a sufficient number of samples and allow the comparison with previous studies focusing on the same vowels.

TABLE II. Distribution of training and test samples over words in AD and ID speech data sets.

	AD		ID	
Word	train	test	train	test
bead	136	26	196	56
boot	161	43	148	34
key	148	44	125	25
pot	121	25	136	29
sheep	138	41	148	45
shoe	216	54	154	35
sock	148	37	150	44
spoon	135	24	157	41
top	139	42	128	27

TABLE III. Word accuracy (%) for speech recognizers trained on AD versus ID speech and tested on matched versus mismatched conditions.

		Test	
		AD	ID
train	AD	95.5	81.6
	ID	90.2	93.5

During the adult-infant conversations, toys of the same name (sheep, keys, etc.) were provided, which mothers used in interacting with their infants. In the adult-adult conversations, cue words were elicited simply by talking about related topics. Conversations were recorded using a far-field microphone (suspended from the ceiling near speakers' heads). For this reason, the recorded speech is far from studio-quality; rather, it reflects natural listening conditions. First, the use of a far-field rather than a close-talking microphone results in energy fluctuations in the speech signal due to the varying distance of the speaker's mouth to the microphone. Second, much of the speech is overlapped by background noise, such as infants crying or banging toys on the table, people entering and leaving the room, etc. Two infant-directed conversations were discarded *a priori* due to the presence of constant background noise. In addition, the adult-adult conversations contain many instances of speaker overlap or speech overlapped with laughter.

All audio files were time-segmented and transcribed orthographically by phonetically experienced transcribers. The time boundaries of noise events, speaker overlaps, and all instances of cue words were marked. These time marks were double-checked in a second transcription pass, and all instances of cue words were extracted from the audio data. Each instance was categorized by a trained listener with respect to the degree of background noise (on a scale from 1 to 10, with 1 being the lowest and 10 being the highest noise level), and with respect to the presence or absence of pitch-accent on the word (on a scale from 1 to 5, with 1 being the lowest confidence and 5 being the highest). The latter was determined by listening to the word in its context in the original speech file. All samples of cue words that had a noise level higher than three were discarded. Furthermore, only data from those speakers with sufficient instances in both the ID and the AD conversations were selected. As a result, the data from eight speakers could not be used at all, such that the final data set used for the experiments reported below consists of the cue words from 22 different speakers. The remaining samples were randomly assigned to training and test sets. For each speaking mode (ID and AD), there were 1342 training and 336 test samples; every speaker had the same number of training and test samples in both the AD and ID data sets. Table II shows the distributions of samples over training and test sets. Further balancing of the data in terms of speaker and word identities (e.g., enforcing exactly the same number of samples for each word) would have resulted in sets with too few training samples. Although the data sets are very small by ASR standards, they are larger than the samples used by most other phonetic studies on infant-directed speech.

IV. EXPERIMENTS AND RESULTS

The original recorded data files were sampled at 16 kHz and preprocessed using MFCC analysis. The acoustic front-end consisted of 39 coefficients (12 MFCCs, normalized log energy, and their first and second derivatives), extracted every 10 ms with a window size of 25 ms. Cepstral mean and variance normalization were applied on a per-conversation basis to reduce the effects of recording conditions and speaker variation. Whole-word acoustic HMMs were then constructed for each cue word, with eight emitting states and two Gaussian mixture components per state. In accordance with standard practice in small-vocabulary ASR, whole-word- rather than subword-unit-based HMMs were used in order to achieve better coarticulation modeling. The models had a left-to-right topology without skip transitions, e.g., each state could transition only to itself or the immediately following state; states could not be skipped. Given that most ASR systems use three-state models for phone units (which are on average 30 ms long), the number of states reflects the expected duration of words consisting of three phones. Diagonal covariance matrices were used in each mixture component. The models were trained using three iterations of EM; recognition was performed by one-best Viterbi search. Due to the small size of the data set, more advanced modeling techniques which increase the number of parameters (such as context-dependent models or full covariance matrices) were not used.

A. Experiment I: Baseline study

The performance of the recognizers is measured as word accuracy, i.e., the percentage of correctly recognized samples in the test set. The results (Table III) show that absolute recognition performance is highest on the AD data set when using an AD-trained recognizer (upper left corner of Table III). In both conditions, the recognizer trained and tested on matched sets (the diagonal of the table) performs better than the corresponding recognizer trained on mismatched data. However, systems behave differently with respect to the relative degradation of recognition performance under mismatched conditions. The relative degradation of the AD-trained system on ID speech is 16.5% whereas the ID-trained system applied to AD speech shows a relative loss of 3.3%

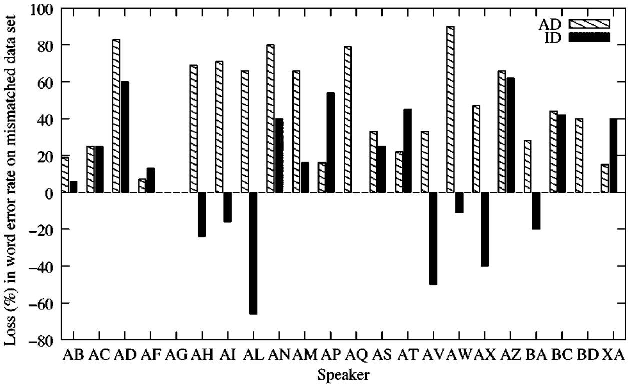


FIG. 1. Per-speaker relative degradation (in % word error rate) of AD and ID recognizers on mismatched test data.

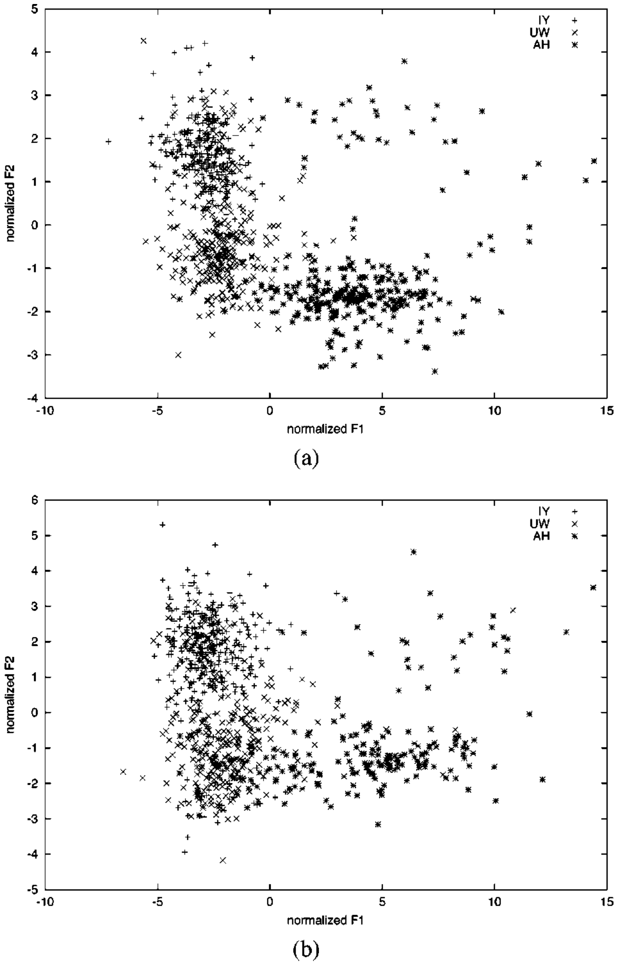
accuracy. This difference was statistically significant at the 0.0001 level, using a difference of proportions significance test. The absolute differences in accuracy are on the same scale (loss of 13.9% absolute of AD recognizer on ID speech and 3.3% of the ID recognizer on AD speech).

The relative degradation was also computed on a per-speaker basis. Results are displayed in Fig. 1. For 16 out of 22 speakers, the AD-trained recognizer showed a stronger degradation on ID speech (indicated by the striped bars) than vice versa (black bars). For two speakers, the performance did not change, and four speakers showed a stronger relative degradation on AD than on ID speech. The ID-trained recognizers also showed a better performance on the AD data set than on the ID test set for several speakers (the bars extending into the negative region in Fig. 1). From this we can conclude that the ID-trained recognizers perform more robustly under mismatched training and test conditions. There was no correlation with the amount of training or test data used per speaker.

B. Analysis

An analysis of the recognition errors showed that confusions occurred mainly between words sharing the same vowel (e.g., confusions between *key* and *bead* or *top* and *sock*) and words of the /uː/ and /ɒ/ categories (such as *spoon* and *sock*). Confusions between /iː/ words and the other categories were rare, although some examples of *shoe–sheep* confusions did occur, possibly due to strong coarticulation of the /uː/ vowel with the palatal fricative /sh/.

For a more detailed analysis it was necessary to look at individual phonetic segments. In order to automatically extract signal portions corresponding to subword units, we trained individual phone models, using a transcription of words in terms of their constituent phones. The preprocessing of the signals was identical to the recognition experiments described above. The phone models consisted of three emitting states with two Gaussian mixture components each; the topology was left-to-right without any skip transcriptions. The initial model parameters were set to the global mean and variance of the data, i.e., the parameters were computed from all feature vectors, regardless of phonetic class. The models were trained in three iterations of EM and were then used in a forced alignment procedure in order to obtain time boundaries of the constituent phones. During this procedure, the reference transcription was provided to the recognizer, and the Viterbi algorithm was used to identify the globally best time alignment of the model sequence defined by the transcription. The resulting time alignments were then checked for accuracy and the corresponding time segments were extracted. This data was used to analyze the distribution and separability of phonetic classes under different acoustic representations and prosodic conditions, as described in the following sections.



(a)

(b)

FIG. 2. Distribution of vowel classes in normalized formant space in AD speech (a) and ID speech (b).

1. Separability of vowel classes

Standard formant-based analysis was applied to the vowel spaces in AD and ID speech. For each vowel segment (as identified by the forced alignment), formant values were obtained using the formant program of the Entropics ESPS package, and they were averaged over the entire length of the vowel. Cross-speaker normalization in the form of mean and variance normalization was applied.

The class distributions in the space defined by the normalized first two formants are shown in Fig. 2. Similar to previous studies which have investigated phonetic discriminability in either speech perception or automatic classification scenarios (e.g., Nossair and Zahorian, 1991; Zahorian and Jagharghi, 1993; Hillenbrand *et al.*, 1995), we computed

TABLE IV. Comparison of within-class distance (V^2), within-class distance (D^2), Fisher ratio ($\mathcal{F}$) (the ratio of within-class to between-class distance), and average Euclidean distance (E) between class means for AD and ID vowel data. The upper row shows results based on speaker-normalized formant values; the lower row shows results for PCA-mapped mel-frequency cepstral coefficients.

	AD				ID			
	V^2	D^2	$\mathcal{F}$	E	V^2	D^2	$\mathcal{F}$	E
Formants	4.98	24.32	0.17	3.96	8.73	20.94	0.29	3.84
MFCCs	1.08	12.72	0.08	3.24	2.70	8.82	0.23	2.89

various numerical measures of class separation. One of these is the Fisher ratio $\mathcal{F}$ (e.g., Schuermann 1996) defined as

$$\mathcal{F}=\frac{V^2}{V^2+D^2}, \tag{6}$$

where

$$V^2=\sum_{k=1}^{K} P_k \mathbf{trace}[\Sigma_k] \tag{7}$$

and

$$D^2=\frac{1}{1-\Sigma_{k=1}^{K} P_k^2}\sum_{k=1}^{K}\sum_{j=1}^{K} P_k P_j(\mu_{\mathbf{k}}-\mu_{\mathbf{j}})^2 \tag{8}$$

where K is the number of classes, Σ_k is the covariance matrix of the kth class, μ_k is the mean of the kth class, and P_k is the prior probability of the kth class. V^2 measures the within-class variance of the features with respect to the class mean. D^2 denotes the interclass distance, i.e., the distance between class means, weighted by the class priors. $\mathcal{F}$ thus expresses the ratio of within-class distance to the between-class distance and ranges between 0 and 1. A lower $\mathcal{F}$ value indicates better separability. We also computed the average Euclidean distance between the three class means. These measures are shown in Table IV.

The results indicate a slightly better separation of class means in AD speech than in ID speech. Overall class separability is noticeably poorer in ID speech than in AD speech; this is due to a stronger degree of class overlap caused primarily by larger class variances.

In order to eliminate possible effects caused by the automatic segmentation, the analysis was repeated for average formant values computed only over the center third of each vowel. However, no significant difference was found between the results.

Formant values are useful for comparing the present analysis to other phonetic studies of vowel spaces; however, they are not used in the speech recognizers. It is not immediately obvious whether class distributions in formant space can be equated with class distributions in the 39-dimensional MFCC space that defines the recognizer front-end. The above analysis was therefore repeated for the MFCC data. Principal components analysis (PCA) (see, e.g., Duda *et al.*, 2001) was applied to the data and feature vectors, and the feature space was mapped to the coordinate system defined by the first two principal components, in order to be able to display the result graphically. Table IV lists the Fisher ratio and distance measures computed on this data, and Fig. 3 shows the corresponding distributions. Similar results can be observed: the average distance of class means is slightly larger and class separation is more pronounced in AD speech. Class separability measures computed on the full 39-dimensional feature space showed the same pattern.

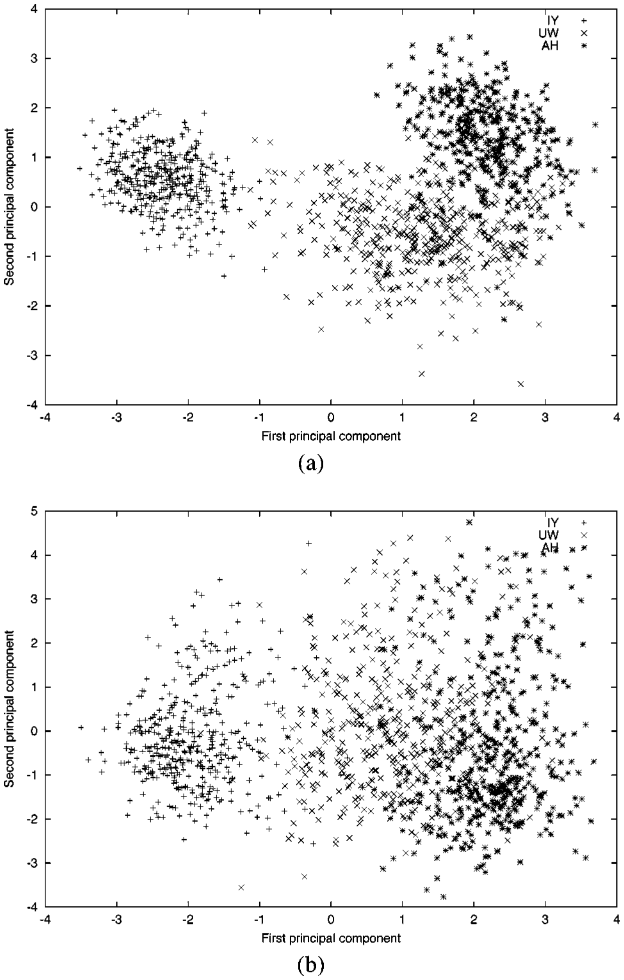


FIG. 3. Distribution of vowel classes in MFCC space mapped to the first two principal components, in AD speech (a) and ID speech (b).

One factor contributing to the variability in ID speech may be that no distinction was made between vowels that are typical examples of ''motherese,'' i.e., vowels that exhibit strong hyperarticulation, and those that are not. Our definition of ''infant-directed speech'' includes the ensemble of speech used in infant-directed conversations. Although this definition reflects natural listening conditions more faithfully, ID speech is a continuum, and a further distinction can be made between different points within this continuum. To this end, pitch-accented versus non-pitch-accented vowels were considered separately. Vowels bearing linguistic stress or pitch-accent are often articulated more clearly, whereas unstressed or unaccented vowels show a stronger degree of coarticulatory reduction, with a concomitant decrease in the size of the vowel space. This has been verified in a number of experimental and corpus-based studies (e.g., Lindblom, 1963; Sluijter and van Heuven, 1996; Kirchhoff and Bilmes, 1999). Vowels that are more representative of ID speech can be expected to be perceived as pitch-accented. We extracted

TABLE V. Within-class distance (V^2), between-class distance (D^2), Fisher ratio ($\mathcal{F}$) (the ratio of within-class to between-class distance), and average Euclidean distance (E) between class means for pitch-accented vowels in AD and ID speech.

AD				ID			
V^2	D^2	$\mathcal{F}$	E	V^2	D^2	$\mathcal{F}$	E
1.91	8.39	0.19	2.86	3.01	8.11	0.27	2.91

TABLE VI. Between-class distance (D^2), within-class distance (V^2), Fisher ratio ($\mathcal{F}$), and average Euclidean distance (E) between class means for syllable-initial consonant pairs /b/-/p/ and /sh/-/i/.

	AD				ID			
	V^2	D^2	$\mathcal{F}$	E	V^2	D^2	$\mathcal{F}$	E
/b/-/p/	1.83	6.83	0.21	1.06	2.05	4.99	0.29	1.04
/sh/-/s/	0.77	3.24	0.19	0.83	1.21	1.63	0.43	0.82

pitch-accented samples according to the manual annotations described above (Sec. II). For each condition, we extracted the 220 vowel samples with the highest pitch-accent values. The analysis measures were then recomputed for these data sets based on PCA-mapped MFCCs. The results are shown in Table V.

In spite of the elimination of non-pitch-accented vowel samples, the pattern observed before still holds: class overlap is greater in ID speech than in AD speech. However, the average mean Euclidean distance is now slightly greater than in ID than in AD speech. These observations confirm results obtained in previous studies (see above Sec. I). Although studies focusing on content words mostly report an expansion of the vowel space, van de Weijer (2001) explicitly compared vowels in content versus function words and found that expansion of the vowel space in ID speech occurred in content words but was reversed in function words, which are presumably not pitch-accented.

2. Separability of consonant classes

Most previous studies of infant-directed speech have focused on the acoustic properties of vowels. Consonants, on the other hand, are less well studied. An exception is Sundberg and Lacerda (1999) and Sundberg (2001), where the effects of infant-directed speech on the voice onset time (VOT) of stop consonants was studied. It was shown that VOT is significantly shorter in infant-directed speech than AD speech for infants in younger infants (three months) but changed to longer, overspecified patterns in older infants (11–14 months).

The error patterns in the recognizer output indicate that word confusions were caused not only by poor separation of vowels, but also by consonant substitutions. Therefore, the class separability measures described above were also applied to consonant distinctions, viz. the stops /b/ vs. /p/ in syllable-initial position, and the fricatives /sh/ vs. /s/ in syllable-initial position. These pairs were chosen because they represent the same broad phonetic class in comparable prosodic contexts; in particular, syllable-initial consonants were chosen because they tend to be less strongly coarticulated than syllable-final consonants. As before, the MFCC space was mapped to two dimensions using PCA. Table VI shows the resulting measurements. Whereas the average class mean distance is fairly similar, the Fisher ratio is again lower in AD speech.

C. Experiment II: Linear discriminant analysis

In pattern recognition, a common technique to enhance acoustic class separability is linear discriminant analysis (LDA) (see, e.g., Duda *et al.*, 2001). LDA is a transformation designed to maximize the between-class distance while minimizing within-class distance; it is thus related to the Fisher criterion discussed above. Applying LDA to the infant-directed speech data should reduce the class overlap and produce better recognition results, possibly with the effect of smoothing out differences in performance between the AD and ID recognizer. We used LDA on both the AD and ID data sets. In each case, the transformation parameters were estimated on the training set and were then applied to both the training and the test set. The transformed data was used to train the speech recognizers, using exactly the same modeling procedures as before (see above Sec. IV A). The recognition results, listed in Table VII,, show that recognition performance improves for both recognizers under both conditions. The relative improvement compared to the baseline results (Table III) is greatest for the AD-trained recognizer on ID test conditions. As before, however, the relative loss in word accuracy on the mismatched condition is more severe for the AD-trained recognizer than for the ID-trained recognizer; the difference is statistically significant at the 0.001 level.

V. DISCUSSION

Automatic speech recognizers trained on ID and AD speech, respectively, were applied to both matched and mismatched test sets. It was found that, on average, matched conditions produced better results than mismatched conditions, confirming the second of the hypotheses stated in Sec. II B. ID-trained recognizers performed better on AD speech test sets for some but not for most most speakers. However, the relative degradation of ID-trained recognizers on AD speech was significantly less severe than in the reverse case.

An analysis of a subset of phonetic class distributions in ID and AD speech showed that ID speech was characterized by a stronger class overlap, which provides an explanation of the recognition results: models trained on strongly overlapped classes can accommodate well-separated test data, but models trained on well-separated data will fail to correctly

TABLE VII. Word accuracy (%) of AD- and ID-trained speech recognizers under identical and mismatched test conditions, after application of LDA transformations.

		test	
		AD	ID
train	AD	97.4	91.8
	ID	92.9	94.7

classify test samples whose distributions show a stronger overlap. Samples in the intersection of the class decision boundaries may receive equally likely scores from the acoustic models, thus causing more confusions. The use of linear discriminant analysis to enhance class separation improved recognition results in all test conditions.

The automatic classification procedure used here does not claim to be a cognitively adequate model of infant perceptual learning. First, automatic speech recognizers are trained on a limited, well-defined training set, whereas it is, in general, impossible to quantify how much training data infants have previously been exposed to. Data collections where the amount of training data for a particular phonetic contrast can be controlled precisely [as in Kuhl *et al.*, (2003)] might turn out to be an interesting testbed for pattern recognition algorithms. Second, the acoustic representation and/or the statistical modeling techniques used in ASR systems certainly have deficiencies compared to human speech perception, since ASR performance still falls short of human performance. Nevertheless, the analyses presented above highlight interesting research questions for the study of infant speech perception. Previous studies on ID speech have emphasized the greater distance between vowel class means and have concluded that ‘‘motherese’’ might facilitate phonetic category learning. In contrast, it was found here (using comparatively large sets of samples and speakers) that class separability is actually poorer in ID speech. A possible reason is that *all* samples produced from infant-directed conversations were included, whereas previous studies may have focused on samples that are most ‘‘motherese-like’’ and thus occupy extreme positions in the vowel space. The slightly greater separation of the class means of pitch-accented vowels observed above confirms this interpretation. Nevertheless, the stronger class overlap is still present in those sets, suggesting that ID speech is poor training data. This result is in line with earlier studies comparing the intelligibility of speech directed to adults versus speech directed to children of 1–3 years (Bard and Anderson, 1983, 1994). In those studies, isolated words excised from either adult-adult or adult-child conversations had to be identified by adult listeners and young children. In Bard and Anderson (1994) additional context was provided for the tokens. It was found that word intelligibility was inversely related to word predictability: predictable words without their context were actually less intelligible in child-directed than in adult-directed speech. The conclusion drawn from this study was that adults might reduce predictable words more when talking to their infants/children than when talking to adults in order to draw the child’s attention to referents that are new in the discourse or in the extralinguistic environment. The observation made by van de Weijer (2001) that ID speech exhibits an enlarged vowel space in content words but a reduced vowel space in function words (see Sec. I) supports this analysis. It seems plausible that words which are predictable or already given in the discourse are underarticulated in favor of words that are new, some of which may then exhibit the typical motherese effect.

Our observations also pose the question of why infants are able to acquire phonetic categories and generalize to new test samples in spite of strongly overlapped training data—we may assume that the majority of everyday speech that infants are exposed to consists of predictable function words [see also Cutler (1993)]. Several explanations may be advanced, e.g., that auditory representations in human speech perception are more invariant, that different (distribution-free) classification mechanisms are involved, or that listeners perform some form of data normalization, selection, or variance reduction, similar to the LDA transformation. Not all of these explanations are equally likely—it has been shown experimentally that infants do respond to distributions in auditory training data (Maye *et al.*, 2002). Moreover, although the acoustic representations and normalization techniques used in this study may not model human auditory representations perfectly, it is safe to assume that they are correlated. It seems most plausible that infants make use of perceptual variance reduction techniques. One possible candidate for such a technique is the perceptual magnet effect (Kuhl, 1991; Kuhl *et al.*, 1992). This effect is often described as a perceptual warping in the sense that auditory stimuli close to a ‘‘prototypical’’ representation of the category cannot be distinguished perceptually, thus decreasing within-class variance. Although the existence of this effect as well as the precise nature of the magnet theory have been under much discussion (Lively and Pisoni, 1997; Lotto *et al.*, 1998; Guenther, 2000; Lotto, 2000), it is striking that phonetic category perception does not occur until fairly late in the first year of life, after the emergence of the magnet effect as dated by proponents of this theory. It is unlikely that only the *amount* of speech data encountered by infants is responsible for the onset of categorical perception, or that categories are formed simply by passively processing all available input without some active contribution by the perceptual system: naturally occurring speech data is extremely variable in its distribution, and, as the present study shows, may even be more variable in ID speech than in AD speech. It may be assumed that infants selectively filter the available input, or perform a compaction of stored class representations [see recent evidence from neural imaging (Guenther and Bohland, 2002)]. Filtering might be done by focusing on the most motherese-like items in the speech input, which then become the prototypes of phonetic classes.

In summary, two conclusions can be drawn from this study. First, from an ASR point of view, ID-style speech seems to be of no immediate benefit to present-day ASR systems: matched training data still leads to superior results on AD test data. Furthermore, filtering of the training data in the way suggested above would not be advisable for ASR systems because of the resulting reduction in training material. The second conclusion is that parameters encoding contextual information (either linguistic or extralinguistic) should be explicitly integrated into data selection methods in acoustic-phonetic studies of ID speech, and into actual theories of infant speech perception. It should be noted that the ASR framework provides a tool for quantifying linguistic predictability (in the form of language model perplexity), which can in turn be used in data selection. It thus seems that the main benefits of combining ASR technology and the study of speech perception lie in the use of ASR techniques for speech data selection and statistical representation as well as for building and testing perceptual models.

ACKNOWLEDGMENTS

This work was funded by the Institute for Learning and Brain Sciences, University of Washington. We would like to thank Patricia Kuhl and Bart de Boer for providing the speech data and three anonymous reviewers for their comments.

Andruski, J. E., and Kuhl, P. K. (**1996**). ‘‘The acoustic structure of vowels in mothers' speech to infants and adults,'' in *Proceedings of ICSLP*, pp. 1541–1544.

Bard, E. G., and Anderson, A. H. (**1983**). ‘‘The unintelligibility of speech to children,'' J. Child Lang **10**, 265–292.

Bard, E. G., and Anderson, A. H. (**1994**). ‘‘The unintelligibility of speech to children: effects of referent availabiity,'' J. Child Lang **21**, 623–648.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (**2002**). ‘‘What's new, pussycat? on talking to babies and animals,'' Science **296**, 1435.

Cooper, R. P., and Aslin, R. N. (**1994**). ‘‘Development differences in infant attention to the spectral properties of infant-directed speech,'' Child Dev. **65**, 1663–1677.

Cutler, A. (**1993**). ‘‘Phonological cues to open- and closed-class words in the processing of spoken sentences,'' J. Psycholinguistic Res. **22**, 109–131.

Damper, R. I., and Hanard, S. R. (**2000**). ‘‘Neural network models of categorical perception,'' Percept. Psychophys. **62**(4), 843–867.

de Boer, B., and Kuhl, P. K. (**2003**). ‘‘Investigating the role of infant-directed speech with a computer model,'' ARLO **4**, 129–134.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). ‘‘Maximum-likelihood from incomplete data via EM algorithm,'' J. R. Stat. Soc. Ser. B. Methodol. **39**(1), 1–38.

Duda, R. O., Hart, P. E., and Stork, D. G. (**2001**). *Pattern Classification*, 2nd edition (Wiley, New York).

Fernald, A., and Kuhl, P. K. (**1987**). ‘‘Acoustic determinants of infant preference for motherese speech,'' Inf. Behav. Dev. **10**, 279–293.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B., and Fukui, I. (**1989**). ‘‘A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants,'' J. Child Lang **16**, 477–501.

Grieser, D. L., and Kuhl, P. K. (**1988**). ‘‘Maternal speech to infants in a tonal language: support for universal prosodic features in motherese,'' Dev. Psychol. **24**, 14–20.

Guenther, F. H. (**2000**). ‘‘An analytical error invalidates the ‘depolarization' of the perceptual magnet effect,'' J. Acoust. Soc. Am. **107**, 3576–3580.

Guenther, F. H., and Bohland, J. W. (**2002**). ‘‘Learning sound categories: a neural model and supporting experiments,'' Acoust. Sci. Technol. **23**(4), 213–220.

Hazan, V., and Simpson, A. (**2000**). ‘‘The effect of cue-enhancement on consonant intelligibility in noise: speaker and listener effects,'' Lang Speech **43**, 273–294.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). ‘‘Acoustic characteristics of American English vowels,'' J. Acoust. Soc. Am. **97**, 3099–3111.

Holt, L. L., Lotto, A. J., and Kluender, K. R. (**1998**). ‘‘Incorporating principles of general learning in theories of language acquisition,'' M. Gruber, C. D. Higgins, K. S. Olson, and T. Wysocki (eds.) in *Chicago Linguistic Society: Papers from the Panels*, 34(2), pp. 253–268, Chicago: Chicago Linguistic Society.

Jusczyk, P. W., Luce, P. A., and Charles-Luce, J. (**1994**). ‘‘Infants' sensitivity to phonotactic patterns in the native language,'' J. Mem. Lang. **33**, 630–645.

Kirchhoff, K., and Bilmes, J. (**1999**). ‘‘Statistical acoustic effects of coarticulation,'' in *Proceedings of the 14th Int. Congress on Phonetic Sciences*, San Francisco, CA.

Kuhl, P. K. (**1991**). ‘‘Human adults and human infants show a ‘perceptual magnet effect' for the prototypes of speech categories, monkeys do not,'' Percept. Psychophys. **50**, 93–107.

Kuhl, P. K., Tsao, F. M., and Liu, H. M. (**2003**). ‘‘Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning,'' Proc. Natl. Acad. Sci. U.S.A. **100**, 9096–9101.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (**1992**). ‘‘Linguistic experience alters phonetic perception in infants by 6 months of age,'' Science **255**, 606–608.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (**1997**). ‘‘Cross-language analysis of phonetic units in language addressed to infants,'' Science **277**, 684–686.

Lindblom, B. (**1963**). ‘‘Spectrographic study of vowel reduction,'' J. Acoust. Soc. Am. **35**, 1773–1781.

Liu, H. M., Kuhl, P. K., and Tsao, F. M. (**2003**). ‘‘An association between mothers' speech clarity and infants' speech discrimination skills,'' Deve. Sci. **6**(3), F1–F10.

Lively, S. E., and Pisoni, D. B. (**1997**). ‘‘On prototypes and phonetic categories: a critical assessment of the perceptual magnet effect in speech perception,'' J. Exp. Psychol. Hum. Percept. Perform. **23**(6), 1665–1679.

Lotto, A. J. (**2000**). ‘‘Reply to An analytical error invalidates the ‘‘depolarization'' of the perceptual magnet effect,'' J. Acoust. Soc. Am. **107**, 3578–3581.

Lotto, A., Kluender, K., and Holt, L. L. (**1998**). ‘‘Depolarizing the perceptual magnet effect,'' J. Acoust. Soc. Am. **103**, 3648–3655.

Maye, J., Werker, J. F., and Gerken, L. A. (**2002**). ‘‘Infant sensitivity to distributional information can affect phonetic discrimination,'' Cognition **82**, B101–B111.

Moore, R. K. (**2003**). ‘‘A comparison of the data requirements of automatic speech recognition systems and human listeners,'' in *Proceedings of Eurospeech*, pp. 2581–2584.

Nearey, T. M. (**1997**). ‘‘Speech perception as pattern recognition,'' J. Acoust. Soc. Am. **101**, 3241–3254.

Nossair, Z. B., and Zahorian, S. A. (**1991**). ‘‘Dynamic spectral shape features as acoustic correlates for initial stop consonants,'' J. Acoust. Soc. Am. **89**, 2978–2990.

Payton, K., Uchanski, R., and Braida, L. (**1994**). ‘‘Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing,'' J. Acoust. Soc. Am. **95**, 1581–1592.

Picheny, M., Durlach, N., and Braida, L. (**1986**). ‘‘Speaking clearly for the hard of hearing II: acoustic characteristics of clean and conversational speech,'' J. Speech Hear. Res. **32**, 93–103.

Protopapas, A. (**1999**). ‘‘Connectionist modeling of speech perception,'' Psychol. Bull. **125**(4), 410–436.

Protopapas, A., and Calhoun, B. (**2000**). ‘‘Adaptive phonetic training for second language learners,'' in *Proceedings of InSTIL (2nd International Workshop on Integrating Speech Technology in Language Learning)*, Dundee, UK, pp. 31–38.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (**1996**). ‘‘Statistical learning by 8-month-olds,'' Science **274**, 1926–1928.

Scharenborg, O., Boves, L., and de Veth, J. (**2002**). ‘‘ASR in a human word recognition model: generating phonemic input for Shortlist,'' in *Proceedings of ICSLP*, pp. 633–636.

Scharenborg, O., McQueen, J. M., ten Bosch, L., and Norris, D. (**2003**). ‘‘Modelling human speech recognition using automatic speech recognition paradigms in SpeM,'' in *Proceedings of Eurospeech*, pp. 2097–2100.

Schuermann, J. (**1996**). *Pattern Classification: A Unified View of Statistical and Neural Approaches* (Wiley, New York).

Sluijter, A. M. C., and van Heuven, V. J. (**1996**). ‘‘Acoustic correlates of linguistic stress and accent in Dutch and American English,'' in *Proceedings of ICSLP*, pp. 630–633.

Sundberg, U. (**2001**). ‘‘Consonant specification in infant-directed speech. Some preliminary results from a study of voice onset time in speech to one-year-olds,'' Lund University Department of Linguistics Working Papers **49**, 148–151.

Sundberg, U., and Lacerda, F. (**1999**). ‘‘Voice onset time in speech to infants and adults,'' Phonetica **56**, 186–199.

van de Weijer, J. (**2001**). ‘‘Vowels in infant- and adult-directed speech,'' Lund University Department of Linguistics Working Papers **49**, 172–175.

Zahorian, S. A., and Jagharghi, A. J. (**1993**). ‘‘Spectral-shape features versus formants as acoustic correlates for vowels,'' J. Acoust. Soc. Am. **94**, 1966–1982.