Behav Res
DOI 10.3758/s13428-012-0222-0

Fluency Profiling System: An automated system for analyzing the temporal properties of speech

Daniel R. Little · Raoul Oehmen · John Dunn · Kathryn Hird · Kim Kirsner

Abstract The temporal characteristics of speech can be captured by examining the distributions of the durations of measurable speech components, namely speech segment durations and pause durations. However, several barriers prevent the easy analysis of pause durations: The first problem is that natural speech is noisy, and although recording contrived speech minimizes this problem, it also discards diagnostic information about cognitive processes inherent in the longer pauses associated with natural speech. The second issue concerns setting the distribution threshold, and consists of the problem of appropriately classifying pause segments as either short pauses reflecting articulation or long pauses reflecting cognitive processing, while minimizing the overall classification error rate. This article describes a fully automated system for determining the locations of speech–pause transitions and estimating the temporal parameters of both speech and pause distributions in natural speech. We use the properties of Gaussian mixture models at several stages of the analysis, in order to identify theoretical components of the data distributions, to classify speech components, to compute durations, and to calculate the relevant statistics.

Keywords Natural speech · Pause durations

Speech is a dynamic, embodied cognitive system requiring the temporal coordination of many cognitive processes, including planning, the retrieval of semantic and lexical information from memory, discourse monitoring, phonemic construction, and breathing (Davis, Zhang, Winkworth, & Bandler, 1996; Elman, 1995; Ford & Holmes, 1978; Kirsner, Dunn, & Hird, 2005; Krivokapić, 2007; Power, 1985). The temporal properties of speech, particularly pause durations, provide an interface between externally observable behavior and the underlying cognitive processes. As a result, the temporal properties of speech convey important information to the listener. For example, pause rates influence a listener's perception of a speaker's fluency (Grosjean & Lane, 1976), and pause durations mark topic shift boundaries in natural speech (Krivokapic, 2007). Consequently, an analysis of pauses can inform the evaluation of speech disorders: While Broca's aphasia is characterized by long pauses reflecting retrieval difficulty (Hird, Brown, & Kirsner, 2006), patients with Parkinson's show a decrease in the overall number of pauses as compared to controls (Skodda & Schlegel, 2008), and stuttering is associated with an overestimation of the temporal duration of speech and pauses (Barasch, Guitar, McCauley, & Absher, 2000).

Pause durations also provide insight into performance on complex tasks that tax cognitive resources such as working memory, attention, and vigilance. For example, public speaking is characterized by longer pauses than is normal speaking (Kirsner et al., 2005; Levin & Silverman, 1965). Likewise, working memory capacity is correlated with

D. R. Little (✉)
Psychological Sciences, University of Melbourne,
Parkville, Victoria 3010, Australia
e-mail: daniel.little@unimelb.edu.au

D. R. Little · R. Oehmen · K. Hird · K. Kirsner
University of Western Australia,
Crawley, Western Australia

J. Dunn
University of Adelaide,
Adelaide, South Australia

K. Hird · K. Kirsner
University of Notre Dame,
Fremantle, Western Australia

Published online: 21 June 2012

verbal fluency (Daneman, 1991), deliberate lying increases the durations of pauses in speech (Sporer & Schwandt, 2006, 2007; Vrij et al., 2008), and communicating multiple rather than single goal messages results in longer pauses at speech onset (Greene & Lindsey, 1989). Pause durations are also important markers of speech intent and are influenced by training and feedback. For example, speakers have longer pauses around topic changes than in the middle of sentences (Goldman-Eisler, 1968), but these longer pauses disappear when people have time to plan for topic changes in advance (Greene & Cappella, 1986). In addition, pauses almost completely disappear when an operant conditioning procedure is used to punish people for making pauses. Although punishment reduces pausing behavior, it increases disfluencies such as mispronunciations, repetitions, false starts, and filled pauses (Beattie & Butterworth, 1979). An analysis of pause durations would allow for quantification of these effects to facilitate formal theory building and computational modeling of speech dynamics.

A substantial barrier to temporal speech analysis is the lack of automatic, efficient, and reliable methods for segmenting natural speech into speech and pause components. Manual segmentation—that is, the separation of prerecorded speech segments into speech and pause components—has previously relied on arbitrary minimum pause durations that exclude potentially meaningful information (Goldman-Eisler, 1968; Hieke, Kowal, & O'Connell, 1983) while also being highly variable and demonstrating only low reliability across studies, across analysts, and across sessions (Oehmen, Kirsner, & Fay, 2010). The purpose of this article is to introduce a fully automated process for the segmentation and analysis of temporal speech and pause characteristics.

The temporal characteristics of speech can be captured by examining the distributions of the durations of measurable speech components (i.e., speech and pause durations). Speech and pause duration distributions are known to be log-normally distributed (Campione & Véronis, 2002, 2005; Kirsner, Dunn & Hird, 2003; Kirsner, Dunn, Hird, Parkin, & Clark, 2002; Rosen, Kent, & Duffy, 2003). Furthermore, pause distributions have been shown to be bimodal in log time (see Fig. 1; Demol, Verhelst, & Verhoeve, 2007; Kirsner et al., 2002). The two components of the bimodal pause distribution correspond to short pauses important for articulation (approximately less than 250 ms; Hieke et al., 1983) and longer pauses related to breathing, planning, retrieval, discourse coordination, and a variety of other cognitive functions (Kirsner et al., 2005). Both short and long pause duration distributions provide valuable information about fluency across different clinical populations and different social contexts. For instance, the mean of the long pause duration distribution, but not of the short pause distribution, is increased relative to controls in individuals with Broca's aphasia (Hird & Kirsner, 2010; Kirsner et al., 2005). Consequently, the parameters of the pause distributions can be used to assess multiple factors pertaining to cognitive workload, aphasia, breathing difficulties, arousal, and emotion that affect speech fluency.

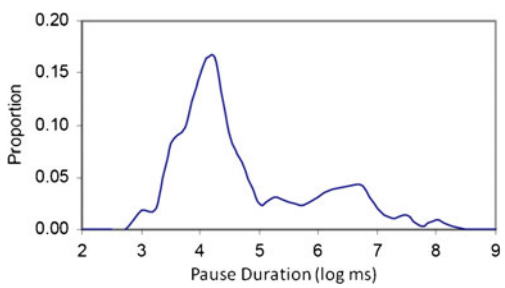

Fig. 1 Distribution of pause durations (log milliseconds) showing bimodality

There are several barriers to conducting an analysis of pause durations. The first is that natural speech is often noisy. Previous evidence (Jantvik, Gustafsson & Paplinski, 2011; Schulze & Langner, 1997; Song, Skoe, Banai, & Krauss, 2011) has identified a dynamic and interactive network of processes required for the perception of speech in noise. Consequently, any speech analysis undertaken in soundproof conditions employs a set of processes that is not representative of those that are used in natural speech processing. In developing our system, we have been careful to ensure that it is capable of providing reliable output for spoken monologues recorded in natural conditions.

An additional barrier to pause analysis is finding the minimum duration that counts as a true pause segment. Short articulatory pauses are not related to cognitive processes and should be treated separately; however, determining the criterion between articulation and cognition is difficult (see, e.g., Hieke et al., 1983). If we assume that the log-normal pause distribution is bimodal, how do we appropriately classify the pause segments as either short or long in order to minimize the classification error rate in a principled manner? In order to conduct an analysis of pauses, we require a method for determining these thresholds. Traditionally, the pause durations are segmented from speech using a visualization program (e.g., Praat; Boersma, 2001) and by manually entering boundaries at speech–pause transitions. The threshold between the short- and long-pause duration distributions is then either estimated by hand or derived by using an ad hoc criterion (Goldman-Eisler, 1968; Hieke et al., 1983; Jaffe & Feldstein, 1970; Kirsner et al., 2005); however, the latter method typically ignores the distinction between short and long pauses (e.g., thresholds range from 100 to 1,000 ms), and the former can be error prone and unreliable. Additionally, both approaches fail to consider individual differences between speakers, such that

while the 250-ms threshold employed by Goldman-Eisler appears to be a reasonable approximation on average, it results in substantial errors of classification for most individual speakers. Because the automatic analysis can be applied to each speaker individually, the specification of a fixed boundary is unnecessary, but instead boundaries can be estimated from the properties of each individual's speech. Following presentation of the automatic system, we will provide a demonstration of the system's reliability and validity.

Automatic fluency measurement

This article describes a fully automated system for determining speech–pause transitions and estimating the temporal parameters of those distributions for individual speakers. We use the properties of Gaussian mixture models (GMM) and the expectancy-maximization (EM; McLachlan & Peel, 2000) algorithm at several stages of the analysis in order to identify theoretical components of the data distributions, classify speech components, compute durations, and calculate relevant statistics for different speech components.

The Fluency Profiling System (FPS) shown in Fig. 2 is composed of two main processing steps: (a) *pause identification* and (b) *pause analysis*. The goal of the first stage is to identify segments of the recording that correspond to low-energy periods of the speech signal, which we interpret as pauses. The goal of the second stage is to partition the low-energy segments identified in the first stage into "short" and "long" pauses and then to define speech segments as connected high-energy segments interrupted only by sequences of short pauses.

The input to the system is a prerecorded segment of natural speech. The speech segment is digitally sampled and filtered, resulting in a smoothed speech segment. In the pause identification stage, the speech amplitude data are converted to *energy*, and a Dirichlet process mixture of Gaussians (DP-GMM; Navarro, Griffiths, Steyvers, & Lee, 2006; Rasmussen, 2000) is used to identify a low-energy component, interpreted in the first instance as a pause. Having partitioned low and high energies using DP-GMM, breaths are then recovered from the high-energy component using pretrained Gaussian mixture models. Finally, a set of classification rules are applied to convert punctate samples of low energy (i.e., nominal pauses), high-energy nonspeech (i.e., breaths), and high-energy speech into a more homogeneous set of segments satisfying certain minimal criteria (e.g., speech segments must exceed a minimum duration; see Table 1). In the pause analysis stage, an EM algorithm is used to classify the pause durations as either short or long pauses; the distributional parameters of the short- and long-pause distributions are computed; and the final speech segments, now defined by the positions of the long pauses, are available for postprocessing. In the current version of the FPS, we incorporate manual transcription of the final speech segments and posttranscription analysis. Each of these steps is further illustrated below.

Pause identification

A prerecorded .wav file is sampled at a standard of 44.1 KHz, providing roughly one sample every 0.02 ms. Depending on the duration of the total speech sample, the resulting list of amplitude values can be very large. Hence, subsequent processing serves two purposes: first, to remove high- and low-frequency noise, and second, to reduce the overall size of the amplitude data to increase computational efficiency.

FPS follows standard procedures for auditory speech recognition (Morgan, Bourlard, & Hermansky, 2004). First, the system uses a low-pass filter to remove high-frequency noise by applying an anti-aliasing Kaiser window to the sampled data. (A windowing function outputs frequencies within a certain range and sets all frequencies outside of this range to zero.) A polyphase finite impulse response filter is applied to down-sample the data from 44.1 to 16 KHz. The effect of downsampling the amplitude data is to implement a moving average that smoothes the overall speech segment and reduces the number of sampled points. The final filtering process involves the application of a high-pass filter designed with a Hamming window to filter out low-frequency content. The amplitude data are then

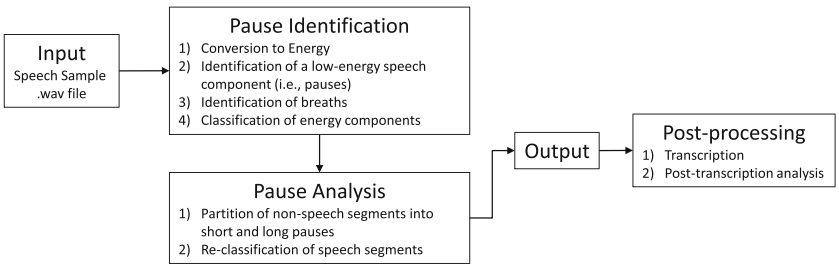

Fig. 2 Overview of the Fluency Profiling System. The analysis is conducted in two stages: a pause identification stage and a pause analysis stage. The output is then available for postprocessing

Table 1 Classification table for entire speech segment

A. High Energy	B. Low Energy	C. Breath	D. Low Energy	E. Speech
IF ≥ 50 ms, classify as Speech	IF ≤ 19 ms, classify as Speech and combine with A	Classify as Speech and combine with A and B	IF ≤ 19 ms, classify as Speech and combine with A, B, C, and E	IF ≥ 50 ms, classify as Speech
IF ≥ 50 ms, classify as Speech	IF ≥ 20 ms, classify as Pause	Classify as Pause and combine with B	IF ≥ 20 ms, classify as Pause and combine with B and C	IF ≥ 50 ms, classify as Speech
IF ≥ 50 ms, classify as Speech	IF ≥ 20 ms, classify as Pause	Classify as Pause and combine with B	IF ≤ 19 ms, classify as Speech and combine with E	IF ≥ 50 ms, classify as Speech
IF ≥ 50 ms, classify as Speech	IF ≤ 19 ms, classify as Speech and combine with A	Classify as Pause and combine with D	IF ≥ 20 ms, classify as Speech and combine with C	IF ≥ 50 ms, classify as Speech
IF commencement of discourse		Classify as Speech and combine with D and E	IF ≤ 19 ms, classify as Speech and combine with E	IF ≥ 50 ms, classify as Speech
IF commencement of discourse		Classify as Speech	IF ≥ 20 ms, classify as Nonspeech	IF ≥ 50 ms, classify as Speech

A, B, C, D, and E comprise a temporal sequence of events

converted to instantaneous energy according to the following formula:

$$\text{Energy} = (\text{Amplitude})^2. \quad (1)$$

Converting the amplitudes to energy has the effect of increasing the difference between the high-amplitude speech segments and the low-amplitude nonspeech segments, improving the efficiency of the threshold estimation process (see Fig. 3).

Threshold estimation

We assume that the *log* energy data are distributed according to an unknown mixture of Gaussians (see Fig. 4a). We assume that periods of relative silence (i.e., pauses) are likely to correspond to low-energy Gaussian noise. Consequently, our goal is to first estimate the parameters of a mixture of Gaussians and then to find the threshold that separates the low-energy component from the remaining components. We use the DP-GMM algorithm to determine the number of component groups in the mixture and the best parameters for the component groups. We then apply signal detection theory to find the threshold between the speech and nonspeech distributions. The location of this threshold

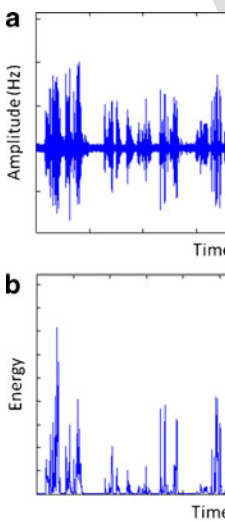

Fig. 3 Examples of (**a**) raw amplitude data and (**b**) amplitude data after conversion to energy

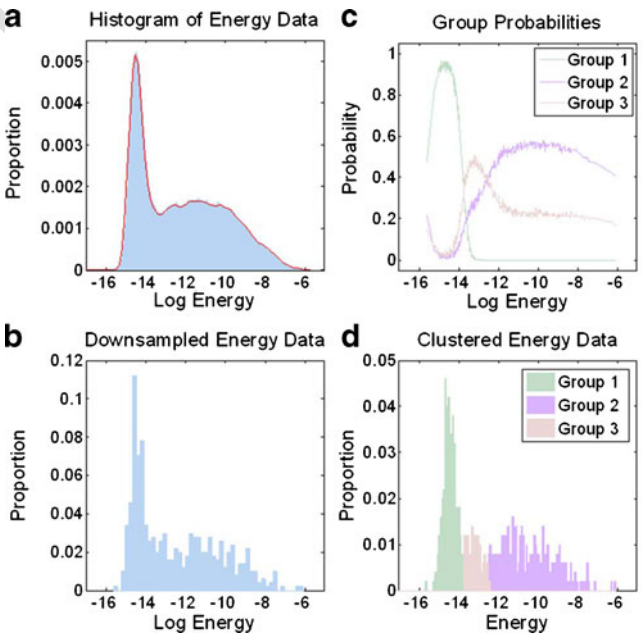

Fig. 4 Example of input and output from the DP-GMM procedure. **a** Histogram of the energy data. The red line is a cubic spline fit to the energy data that is used to sample the data shown in panel B. **b** Down-sampled energy data. **c** Posterior probability that each energy point belongs to each group. **d** Down-sampled energy data clustered into three groups based on the posterior probabilities from the DP-GMM procedure

between the low- and high-energy components is the cutoff value used to classify each energy value.

The DP-GMM algorithm is a nonparametric Bayesian method for fitting a mixture of Gaussians with a potentially infinite number of components. More specifically, each data point, x_i, is assumed to have been generated by a Gaussian distribution with a specific mean and variance. If the number of component Gaussian distributions is known to be J, then the finite Gaussian mixture model of the data is as follows:

$$p\left(x_i|\mu_1,\ldots,\mu_j,s_1^2,\ldots,s_j^2,w_1,\ldots,w_j\right)=\sum_{j=1}^{J}w_jN\left(\mu_j,s_j^2\right),$$

where μ_j and σ_j^2 are the mean and variance of the jth Gaussian distribution (specified by N) and w_j is the proportion of data from the entire data set generated by distribution j. In an infinite mixture of Gaussians, the number of components is not fixed at J, but is instead allowed to be any number. The goal of the nonparametric Bayesian approach is to find the posterior probability over the number of components (and the parameters of those components). This is achieved by setting a Dirichlet process prior on the number of components. The details of Dirichlet process models are described in detail elsewhere (Navarro et al., 2006; Rasmussen, 2000; Teh, 2011); it suffices to note that the DP-GMM model assigns each data point to a component distribution according to (1) the likelihood that the data point was generated from that component and (2) the probability of adding another data point to that distribution, which in the Dirichlet process prior is proportional to the number of data points in the component distribution, versus the probability of forming a new component, which in the Dirichlet process prior is proportional to a parameter, α (e.g., higher values of α increase the probability of forming a new component; Teh, 2011). The component distributions could be components that also generated other data or a component that contains solely the data point in question. Each of these possibilities has some probability in the posterior. In our application, we summarize the posterior by simply choosing the mixture that has the highest posterior probability (see Fig. 4).

Ultimately, the parameters of the DP-GMM are used to define the data upon which the rest of the analysis is based. Once the parameters are known, the next step is to classify each amplitude value as either high-energy or low-energy. As is shown in Fig. 5 (final panel), the threshold between the distributions is identified by finding the minimum of the function formed by summing the survivor function, $S(x)$, of the first component [the survivor function of x is $1 - F(x)$] with the cumulative distribution function [cdf; $F(x)$] of the second component. The minimum of this function is the point at which the two components of the mixture distributions can be separated with the least error. Using this threshold, each data point is classified as either high- or low-energy. The length of consecutive data points in the same category multiplied by the sampling rate gives the duration of each high- and low-energy sample. This procedure allows us to define a period of alternating high- and low-energy segments of different durations. The durations of these segments are the target for the remainder of the analysis, and in the next pause identification stage, the identified high-energy segments are passed to the breath classifier.

One benefit of using the DP-GMM algorithm to find the parameters of a Gaussian mixture model is that the number of component distributions does not need to be specified in advance. However, like other machine-learning algorithms (e.g., Gaussian process classification, Rasmussen & Williams, 2006; or support vector machines, Schölkopf & Smola, 2002), the DP-GMM is inefficient with large data sets. This is a problem for the present task, as we sample the speech file using a high sampling rate, resulting in a large data set for analysis. We deal with this by down-sampling the data, but this potential reduction in information could result in a misestimation of the mixture parameters. We address the trade-off between the speed and accuracy of the algorithm in the section on reliability below.

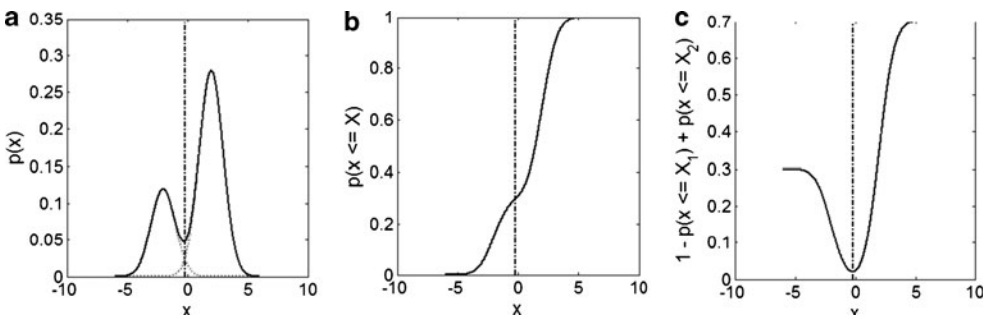

Fig. 5 Method for finding the threshold between the Gaussian components of a finite mixture model. The means and variances of the models were [–2, 1] and [2, 1] for each component, respectively. Component 1 had a weight of .3; Component 2 had a weight of .7. (**a**) Probability density function for the mixture model. (**b**) Cumulative density function (cdf) for the model. (**c**) Sum of the cdf of the second component and one minus the cdf of the first component

Breath classification

Communicative speech is not the only type of utterance that might have energy that exceeds the classification threshold; other types of vocal sounds, such as clicks or pops occurring due to physical properties of the mouth and tongue, as well as breathing, might also have high energy levels. The aim of this stage is to identify and remove high-energy segments that are likely to be breaths and not communicative speech. In principle, the same method could be used for other high-energy noncommunicative sounds (e.g., clicks, pops, or ums and ahs), but in the current version of the system, we opted to address these sounds in a later manual transcription stage. To identify breaths, we again use the properties of GMMs; however, because we know in advance the number of components, we apply the EM algorithm (McLachlan & Peel, 2000) to find the parameters of the component distributions.

The EM algorithm for estimating the parameters of a finite mixture model is described succinctly in Martinez and Martinez (2002, p. 296), and we simply repeat the relevant equations here. The idea is to use a priori considerations to specify the number of mixture components in the model; an initial guess at the component parameters (e.g., means and variances for a mixture of Gaussians), as well as at weights, w, for each of the component distributions, is used to start a cycle of EM updating steps. For each step, we determine the probability, t_{ij}, that each data point, i, is a member of component distribution j:

$$t_{ij} = \frac{w_j p\left(x_i|\mu_j, \sigma_j^2\right)}{\sum_{j=1}^{J} w_j p\left(x_i|\mu_j, \sigma_j^2\right)} \tag{2}$$

These quantities are used to compute new parameters given the initial set of parameters. The new weights are found by averaging the component probabilities over all of the data points.

$$w_j^{'} = \frac{\sum_{i=1}^{N} t_{ij}}{N} \tag{3}$$

The new means and variances are given by

$$\widehat{\mu}_j = \frac{1}{N} \sum_{i=1}^{N} \frac{t_{ij} x_i}{w_i^{'}} \tag{4}$$

and

$$\widehat{\sigma}_j^2 = \frac{1}{N} \sum_{j=1}^{N} \frac{t_{ij}\left(x_i - \mu_j\right)^2}{w_i^{'}} \tag{5}$$

The updating algorithm is initiated by specifying random starting points for mixture parameters (i.e., the means, variances, and mixing weights). On each iteration, we compute the membership probabilities and update the parameters, repeating the process until the change in the likelihood of the data given the mixture parameters is reduced to some cutoff. For the present application, we repeated the entire EM process from a number of different random starting points (i.e., 20) to avoid local minima and ensure convergence to the global minimum.

For breath classification, rather than finding a threshold to separate components of a bimodal distribution, the parameters of the GMMs are found for a prerecorded set of breath and communicative speech samples and then used to classify new segments according to the likelihood that a sample was generated by each distribution.

To train the GMMs, we used a set of prerecorded breath samples as training data. To briefly summarize, the breath samples are converted from a time series of amplitude data into a discrete set of 16 Mel-frequency-scale cepstral coefficient features. Mel frequency bands provide a good approximation to the human auditory response (Stevens, Volkman & Newman, 1937). Specifically, for each auditory segment, a fast Fourier transform converts the amplitude data into spectral data. The spectral data are converted to Mel-scale cepstral coefficients by first taking the log of the spectral data and then taking a discrete cosine transformation (Reynolds, Quatieri, & Dunn, 2000). Each sample provides the same number of features, and each of the feature distributions (across the entire set of samples) is associated with a different Gaussian distribution. The purpose of training the GMMs is to identify the parameters (means and variances) of the Gaussian distributions for each feature, and the mixture probabilities or weights across all of the features. The parameters of the GMM (means, variances, and weights) are tuned, using the EM algorithm, to the cepstral features of the breath segments. This process is faster than the initial speech segmentation process because the total number of data points is much smaller, and the number of component distributions is known. Each high-energy segment identified in the previous stage is then converted into the same feature set, and the likelihood that a high-energy segment was generated by the GMM trained on the breath samples is determined by calculating a likelihood of observing the features of the high-energy segment, given the parameters of the GMM trained on each of the prerecorded breath samples. If a high-energy segment has a much higher probability of having been generated by a GMM trained on breath segments (i.e., 95 %) than by a GMM trained on speech segments, the high-energy segment is reclassified as a breath.

Energy classification

Following the breath classification, each high- and low-energy temporal sequence is classified according to Table 1. The classifications are recursively applied until no further classifications are necessary. The energy classification acts to smooth over categories to eliminate short-duration events that can be thought of as noise or error arising from the initial classification. The logic behind the classification table is as follows: (1) Only low-energy segments with durations of 20 ms or longer and high-energy segments of 50 ms or longer are considered to be true pause or true speech segments, respectively. These assumptions are based on relevant precedents (e.g., Hird & Kirsner, 2010; Kirsner et al., 2002; Rosen et al., 2003) and on neurophysiological evidence identifying temporal constraints on speech perception and production (Schulze & Langner, 1997; Song et al., 2011). Any segments less than this cutoff length are combined with the adjacent segments. (2) Breath segments are concatenated with the nearest true segment. An example of the classification process is illustrated in Fig. 6. The figure shows a temporal sequence containing a breath segment. Consider first the 49-ms high-energy (speech) segment preceding the breath (location A in Fig. 6). Because this segment is less than 50 ms, it would be reclassified as nonspeech and concatenated with the surrounding nonspeech segments (i.e., the preceding nonspeech segment of 63 ms, and the following nonspeech segment of 141 ms, location B, both of which are shown in green). This classification yields an initial speech segment of 756 ms, followed by a nonspeech segment of 253 ms, a breath of 94 ms, nonspeech of 1,084 ms, and a speech segment of 106 ms. Following Table 1, the breath is classified as a pause and concatenated with the preceding and following nonspeech segments, yielding a final sequence of speech (756 ms), pause (1,431 ms), and speech (106 ms).

The consequence of the classification stage is that high-energy segments are classified as periods of speech and low-energy segments (including any reclassified breaths) are classified as periods of nonspeech or pauses. The durations of these final sequences are the target of the pause analysis stage.

Pause analysis

The aim of the pause analysis is to find the parameters of the GMMs that define the short- and long-pause distributions. Following the classification of energy and breath into speech and pause, the durations of all speech and pause segments are known. As is shown in Fig. 1, the pause durations have a bimodal distribution (in log milliseconds); hence, the first aim is to find the threshold that separates the short-pause distribution from the long-pause distribution.

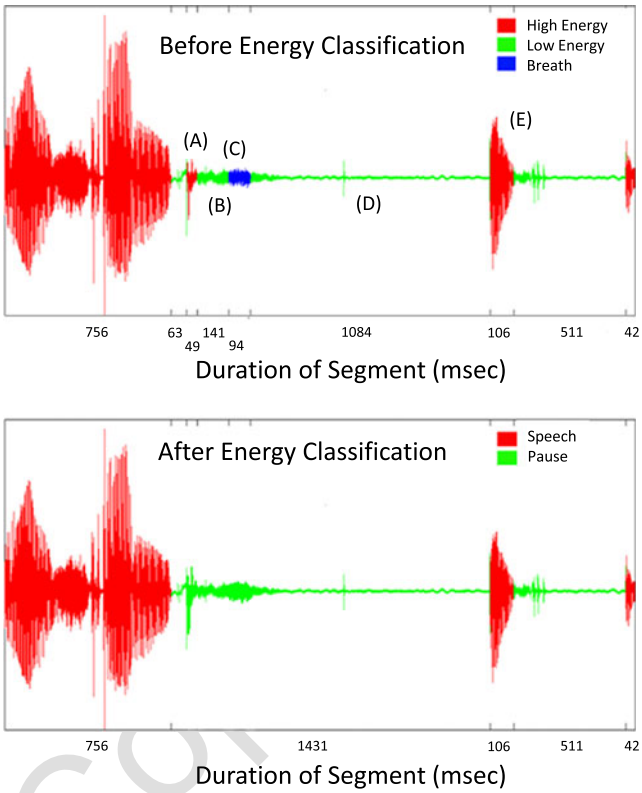

Fig. 6 Example of speech segment after breath classification (see the text for details). The letters correspond to the components in Table 1

The pause analysis again applies the EM algorithm to identify the parameters of the two-component mixture model that best fits the pause duration data. The pause analysis then applies the threshold estimation procedure that was used to separate high energy from low energy (see Fig. 5). For this classification, we use the EM algorithm rather than the DP-GMM because the size of the to-be-classified data set is smaller, making the EM algorithm more efficient than the DP-GMM, and because there is a clear theoretical justification for using a two-component Gaussian mixture model for classifying the log pause durations.

Several statistics are returned from the analysis, including the total durations of each of the components, descriptive statistics (means, standard deviations, and minimum and maximum values) for the speech and pause components on real and log scales, classification thresholds and, perhaps most importantly, the log-likelihood values giving the log-likelihood of the pause durations under a single Gaussian distribution and under a Gaussian mixture model comprising two Gaussian distributions (see Table 2). The log-likelihood values are useful not only for providing optimal parameters but also for conducting model selection between the single and mixture models (see, e.g., Myung, 2003; Myung & Pitt, 2004).

At this stage of the analysis, concurrent speech segments may be separated by short pause segments. The goal of the pause analysis is to identify these short pauses using the

Table 2 Pause analysis output for an example .wav file

Parameter	Value
Number of speech–pause pairs	35
Duration of pause component (ms)	12,542.24
Duration of speech component (ms)	19,029.18
Duration of sample (i.e., speech+pause) (ms)	31,571.42
Speech to pause ratio	1.05
High vs. low energy classification threshold (log energy)	-13.73
Pause minimum (log ms)	2.86
Pause maximum (log ms)	7.67
Pause mean (log ms)	5.20
Pause standard deviation (log ms)	1.24
Pause minimum real (ms)	17.44
Pause maximum real (ms)	2,151.00
Pause mean real (ms)	358.35
Pause standard deviation real (ms)	461.30
Speech segment minimum (log ms)	4.79
Speech segment maximum (log ms)	8.130
Speech segment mean (log ms)	6.04
Speech segment std. dev. (log ms)	0.72
Speech segment minimum real (ms)	120.57
Speech segment maximum real (ms)	3,390.50
Speech segment mean real (ms)	559.68
Speech segment std. dev. real (ms)	583.74
Short vs. long pause classification threshold (log ms)	3.54
Short vs. long pause classification threshold real (ms)	34.41
Proportion (misclassification): Short as long	0.01
Proportion (misclassification): Long as short	0.04
Proportion of data contained in short pause distribution	0.10
Predicted number of short pauses	3
Actual number of short pauses	5
Short pause mean (log ms)	3.21
Short pause standard deviation (log ms)	0.25
Short pause mean real (ms)	24.83
Short pause standard deviation real (ms)	1.28
Proportion of data contained in long pause distribution	0.90
Predicted number of long pauses	32
Actual number of long pauses	30
Long pause minimum (log ms)	3.96
Long pause maximum (log ms)	7.67
Long pause mean (log ms)	5.53
Long pause standard deviation (log ms)	1.01
Long pause minimum real (ms)	52.69
Long pause maximum real (ms)	2,151.00
Long pause mean real (ms)	413.71
Long pause standard deviation real (ms)	476.75
Long speech segment minimum (log ms)	4.88
Long speech segment maximum (log ms)	8.13
Long speech segment mean (log ms)	6.23
Long speech segment std. dev. (log ms)	0.71
Long speech segment minimum real (ms)	131.50
Long speech segment maximum real (ms)	3,390.50
Long speech segment mean real (ms)	660.70
Long speech segment std dev. real (ms)	617.26
Fit of single distribution to pause data (–lnL)	56.67
Fit of two-mixture distribution to pause data (–lnL)	55.48
Fit of single distribution to speech data (–lnL)	36.54

–lnL = negative log likelihood

threshold between the short- and long-pause distributions. Short pauses, which we assume are noncommunicative, are then "folded" back into the adjacent speech segments. Once the pause durations are classified as short (articulatory) or long (cognitive processing) pauses (Goldman-Eisler, 1968), the short pauses are reclassified as part of the surrounding speech segments. At this stage, the speech segments separated by the long pauses (*long speech segments*) and the long pause segments are available for a number of postprocessing analyses, including transcription, syllable counts, and correct information unit analysis (Nicholas & Brookshire, 1993). Before turning to these final processes, we next address the efficiency, reliability, and validity of the FPS analysis.

Efficiency and reliability

The main bottleneck in computation time is the initial segmentation of high and low energy. Several methods are available to reduce the computation time of the initial segmentation, including (1) reducing the number of posterior samples in the DP-GMM and (2) reducing the proportion of data included in the analysis.

We conducted a number of simulations using a sample prerecorded .wav file to test the veracity of the high-versus-low-energy threshold returned by the DP-GMM algorithm under a number of conditions. The .wav file contained a segment of a natural speech that was approximately 32 s in length. In the simulations, we varied the number of posterior samples that were used to estimate the high-/low-energy threshold from 500 to 250, 100, and 50. Orthogonally, we also varied the size of the down-sampled data set from 1,000 samples, to 500, 250, 100, and 50 samples. Smaller data sets were created by sampling each data point according to its proportion in the entire data set, determined by fitting a cubic spline to a histogram of the energy data (see Fig. 4a). For each number of DP-GMM samples and data set size, we ran the FPS 100 times, each time recording the time taken to estimate the threshold and the threshold estimate. Figure 7 shows the average estimated threshold and the average time to estimate the threshold in each of the conditions. Using the

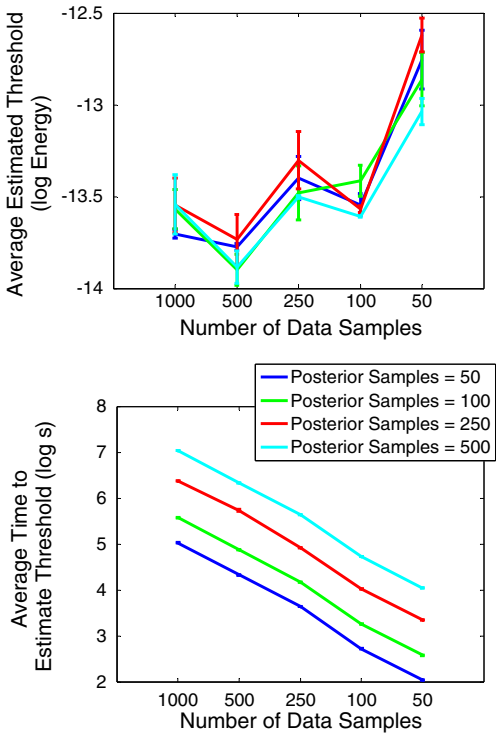

Fig. 7 (Top) Average estimated high-/low-energy threshold using the DP-GMM procedure posterior. (Bottom) Average time to estimate the high-/low-energy threshold using the DP-GMM procedure. Error bars are standards errors

default FPS settings (500 DP-GMM samples and 500 data samples), the high-/low-energy threshold analysis takes approximately 562 s to complete. (All simulations were conducted on an Intel Core i5 CPU 2.67 GHz with 3 GB of RAM.) To summarize these simulations, decreasing the number of data samples improved the speed more than did decreasing the number of samples in the posterior included in the analysis; however, reducing the amount of data included in the analysis resulted in large increases in the threshold estimate. By contrast, reducing the number of posterior samples has a small effect on the estimated threshold. Overall, the procedure produces reliable estimates of the high-/low-energy threshold.

Validity

To assess validity, we compared the automated analysis with four manual segmentations of four different recordings covering a range of conversational topics and recording styles. All of the files involved participants engaging in monologue speech with topics ranging from discussion of their favorite holiday (Files 2 and 4) to the analysis of various fictional scenarios (Files 1 and 3). The lengths of the files were 126, 130, 118, and 80 s, and the signal-to-noise ratios (SNRs) of the recordings were 53.86, 16.36, 5.90, and 20.36 for Files 1–4, respectively. Due to the different durations and SNRs, the sound files can be thought of as providing a test of the FPS with recordings that vary in quality.

We compared the manual segmentation boundary locations (i.e., junctures between speech and pause, in decibels) with the locations found by applying the automatic high-/low-energy threshold. Note that these locations are variable because the threshold is estimated from the smoothed and down-sampled speech waveform. As is shown in Fig. 8, the variability in the speech/nonspeech boundary locations is comparable with manual segmentation.

Figure 9 shows a comparison of the output statistics (log short-pause mean duration and standard error and log long-pause mean duration and standard error) estimated by the four manual observers for each file and the automatic system. In all cases, the automatic system predicted durations in the regions found by the manual analyses. These results confirm that the automatic analysis performed as well as the manual segmentation.

Transcription and final statistics

Along with the FPS, we have also developed a transcription program, TRANSCRIBE, that allows each of the long speech segments to be played separately while a manual transcription is made of the speech content (see Fig. 10). At this stage, a manual operator can also flag segments as nonspeech. This may be necessary to eliminate clicks and pops that have passed through the initial analysis. During the transcription stage, if any speech segments are flagged as nonspeech, those segments are

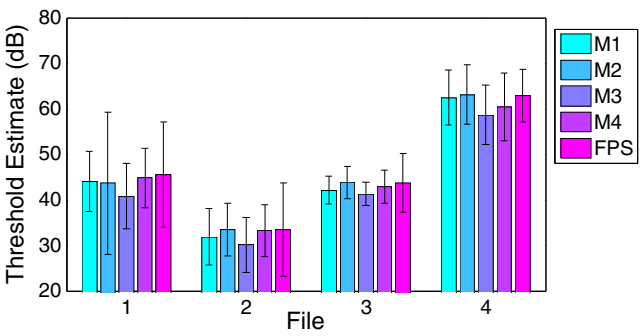

Fig. 8 Comparison of averaged threshold locations (in decibels) between four manual segmenters (M1–M4) and the Fluency Profiling System (FPS) for four different speech samples (see the text for details). Error bars are standards errors

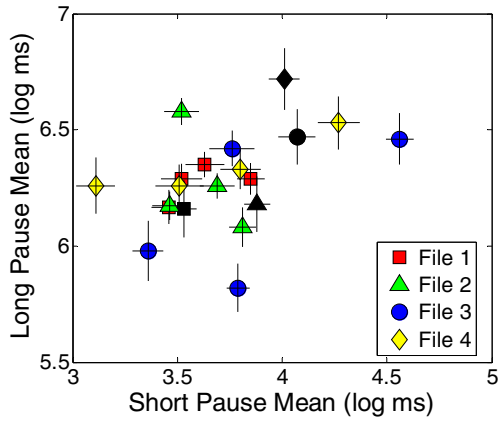

Fig. 9 Comparison of the short-pause distribution mean (log milliseconds) and standard error with the long-pause distribution mean and standard error for four manual segmenters and the FPS for four different speech samples (see the text for details). The black symbols are the predicted output from the FPS system

reclassified as part of the surrounding pause segments and the pause distribution analysis is run a second time to update the relevant statistics (see Table 2). Manual entry of syllable counts and correct information unit counts can also be entered at this stage. The addition of these measures allows for an analysis of the speech "content" within a speech segment, allowing for the calculation of the amount of information transmitted within a given period of time. These analyses would provide further information about the cognitive construction of language and potentially allow for the differentiation of clinical groups (e.g., in aphasia; Hird & Kirsner, 2010).

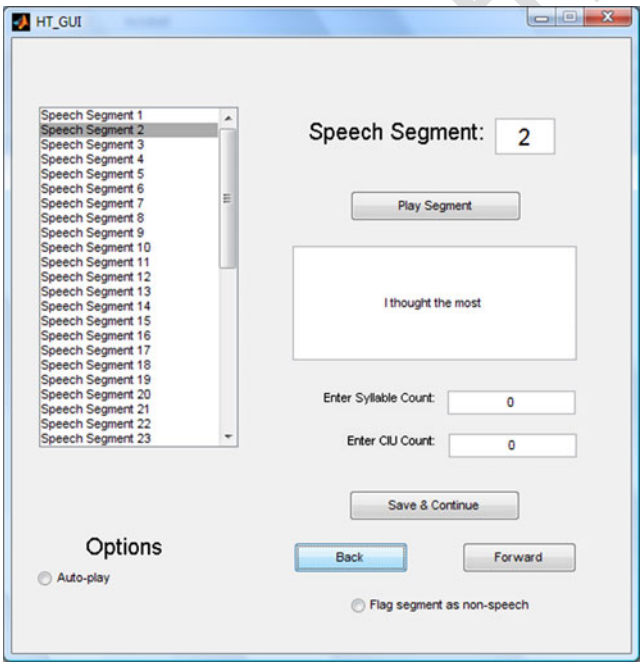

Fig. 10 Screen shot of the TRANSCRIBE interface

Summary

This project responds to a growing interest in developing automatic measures of speech production (Preston, Ramsdell, Oller, Edwards, & Tobin, 2011). The FPS offers an automated way of reliably summarizing important temporal characteristics of natural speech. While manually analyzed studies have been able to demonstrate differences between groups on variables as diverse as reading speed (Demol et al., 2007), aphasia and neurogenic brain disorders (Hird & Kirsner, 2010; Rosen et al., 2003), and language type (Camione & Véronis, 2002, 2005), questions remain over the accuracy and reliability of manual analysis (see, e.g., Oehmen et al., 2010). A reliable and valid automatic method can provide an efficient means of studying differences between populations that vary in fluencies.

Although there are time costs associated with the automatic analysis, several procedures have been put in place to offset these costs. For instance, files may be batch-processed without the presence of an analyst. In short, an automated speech analysis system allows for a reduction in a set of costs associated with the time taken to manually analyze speech samples and the time needed to train and develop expertise in speech analysis. Processing speech segments by hand can take a substantial amount of time, even for well-trained analysts (up to 45 min to identify speech and pause segments within a 1-min sample in our lab); instantiating the pause identification and analyses in an automated program can reduce this time substantially. Furthermore, the expertise necessary to effectively identify pauses in speech is not easily acquired without dedicated training and repetition. Even with expertise, humans may be biased toward detecting boundaries that follow specific types of articulation sounds (e.g., sounds associated with preparatory speech, such as opening the mouth and fricatives; Oehmen et al., 2010). In addition to ameliorating time costs, an automated analysis provides substantial benefits in standardization, objectivity, and reliability. Previous work from different laboratories has been characterized by a high level of idiosyncrasy (e.g., in the short-pause identification threshold; Goldman-Eisler, 1968; Hieke et al., 1983; Jaffe & Feldstein, 1970; Kirsner et al., 2005). Likewise, the subjective nature of manual segmentation places an upper limit on the reliability even for the best trained operators. The FPS, however, allows for consistent results to be obtained.

There are several areas in which the system might be further developed and improved. For example, the FPS still requires the manual transcription of data. Methods based on hidden Markov models of speech transitions might allow for the segmentation and transcription

processes to be automated (Sjölander, 2003; see also Haeb-Umbach, Beyerlein, & Thelen, 1995). A second future direction would be to supplement the breath classification stage with further classifiers to capture well-known transitory phrases, such as 'um' or 'ah' (O'Connell & Kowal, 2005). Finally, some clinical disorders might have particular reliable fluency characteristics beyond the distributions of short and long pauses (e.g., the repetition of starting syllables in patients with a stutter; Riley & Ingham, 2000). In the future, filters to identify these characteristics could be incorporated into the FPS to identify particularly diagnostic features of clinical fluency disorders.

Author note The Fluency Profiling System (FPSv6) and TRANSCRIBE may be downloaded from www.psych.unimelb.edu.au/research/labs/knowlab/fps.htm. Instructions for installation are provided in the downloadable .rar file. The first author was supported by Discovery Project Grants DP120103120 and DP120103888.

References

Barasch, C. T., Guitar, B., McCauley, R. J., & Absher, R. B. (2000). Disfluency and time perception. *Journal of Speech, Language, and Hearing Research, 43,* 1429–1439.

Beattie, G. W., & Butterworth, B. L. (1979). Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech, 22,* 201–211.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5,* 341–345.

Campione, E., & Véronis , J. (2002). A large-scale multilingual study of silent pause duration. In B. Bel & I. Marlien (Eds.), Proceedings of the Speech Prosody 2002 conference (pp. 199–202). Aix-en-Provence: Laboratoire Parole et Langage.

Campione, E., & Véronis, J. (2005). Pauses and hesitations in French spontaneous speech. Proceedings of DiSS'05, Disfluency in Spontaneous Speech Conference, 10-12 september 2005, pp. 43–46. Aix-en-Provence (France).

Daneman, M. (1991). Working memory as a predictor of verbal fluency. *Journal of Psycholinguistic Research, 20*, 445–464.

Davis, P. J., Zhang, S. P., Winkworth, A., & Bandler, R. (1996). Neural control of vocalization: Respiratory and emotional influences. *Journal of Voice, 10,* 23–38.

Demol, M., Verhelst, W., & Verhoeve, P. (2007). The duration of speech pauses in a multilingual environment. In *Proceedings of INTERSPEECH 2007* (pp. 990–993). Antwerp, Belgium: International Speech Communication Association.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 195–226). Cambridge, MA: MIT Press.

Ford, M., & Holmes, V. M. (1978). Planning units and syntax in sentence production. *Cognition, 6,* 35–53.

Goldman-Eisler, F. (1968). *Psycho-linguistics: Experiments in spontaneous speech*. New York, NY: Academic Press.

Greene, J. O., & Cappella, J. N. (1986). Cognition and talk: The relationship of semantic units to temporal patterns of fluency in spontaneous speech. *Language and Speech, 29,* 141–157.

Greene, J. O., & Lindsey, A. E. (1989). Encoding processes in the production of multiple-goal messages. *Human Communication Research, 16,* 120–140.

Grosjean, F., & Lane, H. (1976). How the listener integrates the components of speaking rate. *Journal of Experimental Psychology. Human Perception and Performance, 2,* 538–543.

Haeb-Umbach, R., Beyerlein, P., & Thelen, E. (1995). Automatic transcription of unknown words in a speech recognition system. In *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing* (pp. 840–843). Piscataway, NJ: IEEE Press. doi:10.1109/ICASSP.1995.479825

Hieke, A. E., Kowal, S., & O'Connell, D. C. (1983). The trouble with "articulatory" pauses. *Language and Speech, 26,* 203–215.

Hird, K., Brown, R., & Kirsner, K. (2006). Stability of lexical deficits in primary progressive aphasia: Evidence from natural language. *Brain and Language, 99,* 218–219.

Hird, K., & Kirsner, K. (2010). Objective measurement of fluency in natural language production: A dynamic systems approach. *Journal of Neurolinguistics, 23,* 518–530.

Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue*. New York, NY: Academic Press.

Jantvik, T., Gustafsson, L., & Paplinski, A. (2011). A self-organized artificial neural network architecture for sensory integration with applications to letter-phoneme integration. *Neural Computation, 23,* 2101–2139.

Kirsner, K., Dunn, J., & Hird, K. (2003). Fluency: Time for a paradigm shift. In R. Eklund (Ed.), *Gothenburg papers in theoretical linguistics* (Vol. 90, pp. 13–16). Gothenburg, Sweden: University of Gothenburg, Department of Linguistics.

Kirsner, K., Dunn, J., & Hird, K. (2005, May). *Language productions: A complex dynamic system with a chronometric footprint*. Paper presented at the 2005 International Conference on Computational Science, Atlanta, GA.

Kirsner, K., Dunn, J., Hird, K., Parkin, T., & Clark, C. (2002). Time for a pause. In *Proceedings of the 9th Australasian International Conference on Speech Sciences and Technology* (pp. 52–57). Melbourne, Australia: Australian Speech Sciences & Technology Association Inc.

Krivokapić, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics. 35*, 162–179

Levin, H., & Silverman, I. (1965). Hesitation phenomena in children's speech. *Language and Speech, 8,* 67–85.

Martinez, W. L., & Martinez, A. R. (2002). *Computational statistics handbook using MATLAB*. New York, NY: Chapman & Hall/CRC.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Morgan, N., Bourlard, H., & Hermansky, H. (2004). Automatic speech recognition: An auditory perspective. In S. Greenberg, W. A. Ainsworth, A. N. Popper, & R. R. Fay (Eds.), *Speech processing in the auditory system* (pp. 309–338). New York, NY: Springer.

Myung, I.-J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47,* 90–100.

Myung, I.-J., & Pitt, M. A. (2004). Model comparison methods. *Methods in Enzymology, 383,* 351–366.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology, 50,* 101–122. doi:10.1016/j.jmp.2005.11.006

Nicholas, L. E., & Brookshire, R. H. (1993). Quantifying connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36,* 338–350.

O'Connell, D. C., & Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research, 34,* 555–576.

Oehmen, R., Kirsner, K., & Fay, N. (2010). Reliability of the manual segmentation of pauses in natural speech. *Advances in Natural Language Processing, 6253,* 263–268.

Power, M. J. (1985). Sentence production and working memory. *Quarterly Journal of Experimental Psychology, 37A,* 367–385.

Preston, J. L., Ramsdell, H. L., Oller, D. K., Edwards, M. L., & Tobin, S. J. (2011). Developing a weighted measure of speech sound accuracy. *Journal of Speed, Language and Hearing Research, 54,* 1–18.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems, 12,* 554–560.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10,* 19–41.

Riley, G. D. & Ingham, J. C. (2000). Acoustic duration changes associated with two types of treatment for children who stutter. *Journal of Speech, Language & Hearing Research, 43*, 965–978.

Rosen, K. M., Kent, R. D., & Duffy, J. R. (2003). Lognormal distribution of pause length in ataxic dysarthia. *Clinical Linguistics & Phonetics, 17,* 469–486.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

Schulze, H., & Langner, G. (1997). Periodicity coding in the primary auditory cortex of the Mongolian gerbil (*Meriones unguiculatus*): Two different coding strategies for pitch and rhythm? *Journal of Comparative Physiology. A, 181,* 651–663.

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. *PHONUM, 9,* 93–96.

Skodda, S., & Schlegel, U. (2008). Speech rate and rhythm in Parkinson's disease. *Movement Disorders, 23,* 985–992.

Song, J., Skoe, E., Banai, K., & Krauss, N. (2011). Perception of speech in noise: Neural correlates. *Journal of Cognitive Neuroscience, 23,* 2270–2274.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology, 20,* 421–446.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law, 13,* 1–34.

Stevens, S. S., Volkman, J., & Newman, E. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 3,* 185–190.

Teh, Y. W. (2011). Dirichlet process. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 280–287). New York, NY: Springer.

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior, 32,* 253–265.