# Recognizing complex patterns

Pawan Sinha

*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, USA*
*Correspondence should be addressed to P.S. (sinha@ai.mit.edu)*

Published online 28 October 2002; doi:10.1038/nn949

**How the brain recognizes complex patterns in the environment is a central, but little understood question in neuroscience. The problem is of great significance for a host of applications such as biometric-based access control, autonomous robots and content-based information management. Although some headway in these directions has been made, the current artificial systems do not match the robustness and versatility of their biological counterparts. Here I examine recognition tasks drawn from two different sensory modalities—face recognition and speaker/speech recognition. The goal is to characterize the present state of artificial recognition technologies for these tasks, the influence of neuroscience on the design of these systems and the key challenges they face.**

The ability to recognize patterns in the environment is critical for an organism's survival. It is a pre-requisite for tasks including foraging, danger avoidance, mate selection and, more generally, associating specific responses to particular events and objects. Even putatively simple animals show remarkable recognition prowess. Bees, for instance, can distinguish between complex shapes in a cue-invariant fashion[1], and pigeons seem capable of learning visual concepts from small training sets[2].

An improvement in our understanding of the processes underlying recognition has numerous potential applications. Many of the systems included in futuristic scenarios of science-fiction authors and technology pundits are predicated on recognition technology. Be it the creation of better human–machine interfaces (machines that you can talk to, or that can log you in just by looking at you), assistive devices (smart vehicles that can automatically avoid pedestrians) or autonomous agents (robots that can serve as household helpers), the key enabling technology is recognition.

Artificial recognition systems have had the greatest success in settings that correspond to a highly constrained recognition scenario—matching a new image to very similar training instances. Machine systems that implement this idea via simple variants of template matching have proven effective for numerous inspection tasks, such as monitoring quality control of silicon wafers and ensuring alignment of printed labels on medicine bottles. At these tasks, machine-based systems outperform humans, both in speed and stamina.

However, most tasks in the real world are not subject to such constraints. The patterns that need to be recognized as being the same often differ greatly from each other and may, in fact, be very similar to 'distracter' patterns that they need to be distinguished from. In such settings, humans hold a distinct edge over machines. To make machines work in unconstrained settings, it may be fruitful to complement purely engineering-based approaches with insights regarding brain mechanisms of recognition.

## Face recognition

The events of September 11, 2001, in the USA compellingly highlighted the need for systems that could identify passengers with known terrorist links. In rapid succession, three major international airports, Fresno, St. Petersburg and Logan, began testing face-recognition systems. Although such deployment raises complicated issues of privacy invasion, of even greater immediate concern is whether the technology is up to the task requirements.

The primary challenge in face recognition is that we do not know how to quantify similarity between two facial images in a perceptually meaningful manner. Images 1 and 3 in **Fig. 1** show the same individual from the front and oblique viewpoints, whereas image 2 shows a different person from the front. Conventional measures of image similarity (such as the Minkowski metrics[3]) would rate images 1 and 2 to be more similar than images 1 and 3. In other words, they fail to generalize across important and commonplace transformations. Other transforms that lead to similar difficulties include lighting variations, and aging and expression changes. Clearly, similarity needs to be computed over attributes more complex than raw pixel values.

Some computer-based recognition systems have used more sophisticated face-matching metrics comprising sets of geometric and intensity-based attributes that seem intuitively important. These features may include distances and angles between eye, nose and mouth centroids and also local intensity patches around these centroids[4–6]. However, these systems do not perform very robustly in practice. This is partly because the reliable extraction of facial features is, in itself, a major challenge.

Another popular class of face-recognition approaches is based on projecting face images into lower-dimensional spaces by representing them as linear combinations of a few prototypes[7] or of the principal component vectors. Turk and Pentland's[8] use of this idea for face recognition, a technique they christened 'Eigenfaces', popularized the use of approaches based on principal components analysis and linear discriminant analysis in the field[9–11]. This work also forms the core of the technology used by one of the major commercial face-recognition companies in the USA—Viisage Technologies Inc. Although this and related techniques are computationally elegant, they are limited in the number of image variations they can generalize across[12]. Even for simple transformations such as spatial shifts and size scaling, they have to resort to exhaustive searches through these parameter spaces.

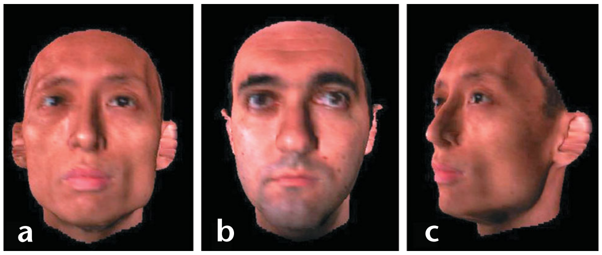

**Fig. 1.** An example that highlights the challenge inherent in the face recognition task. Most conventional measures of image similarity would declare images a and b to be more similar than images a and c, even though both members of the latter pair, but not the former, are derived from the same individual.

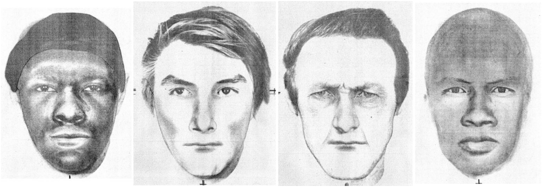

**Fig. 2.** Four facial composites generated by an IdentiKit operator at the author's request. The individuals depicted are all famous celebrities. The operator was given photographs of the celebrities and was asked to create the best likenesses using the kit of features in the system. Most observers are unable to recognize the people depicted, highlighting the problems of using a piecemeal approach in constructing and recognizing faces. The celebrities shown are, from left to right, Bill Cosby, Tom Cruise, Ronald Reagan and Michael Jordan.

Real-world tests of automated face-recognition systems have not yielded encouraging results. For instance, the Palm Beach International Airport recently evaluated face recognition software from Visionics Inc. (now Identix). Using fifteen volunteers and a database of 250 pictures, the system had a success rate of less than fifty per cent and nearly fifty false alarms per five thousand passengers (translating to two to three false alarms per hour per checkpoint). Having to respond to a terror alarm every twenty minutes would, of course, be very disruptive for airport operations. Furthermore, variations such as eyeglasses, small facial rotations and lighting changes proved problematic for the system. Tests of Viisage technology at Logan airport, Boston, yielded similar results. Clearly, a vast gulf remains between the performance of these systems and human observers.

How strong is the relationship between neuroscience and current machine systems for face-recognition? Neuroscience has influenced research on artificial systems in two ways. First, studies of the limits of human face recognition abilities have provided benchmarks against which to evaluate machine systems. Second, studies characterizing the response properties of neurons in the early stages of the visual pathway have guided strategies for image pre-processing in the front-ends of machine systems. For instance, many systems use a wavelet representation of the image that corresponds to the multi-scale gabor-like receptive fields found in the primary visual cortex[13,14]. However, beyond these early stages, it is difficult to discern any direct connections between biological and artificial systems. This is perhaps due to the difficulty in translating psychological findings into concrete computational prescriptions.

A case in point is an idea that several psychologists have emphasized—that facial configuration is important in human judgments of identity[15,16]. However, the experiments so far have not yielded a precise specification of what is meant by 'configuration' beyond the general notion that it refers to the relative placement of the different facial features. This makes it difficult to adopt this idea in the computational arena, especially when the option of using individual facial features such as eyes, noses and mouths is so much easier to describe and implement. Thus, several current systems for face recognition, and also for the related task of facial composite generation (creating a likeness from a witness description), are based on a piecemeal approach.

As an illustration of the problems associated with the piecemeal approach, consider the facial composite generation task. The dominant protocol for having a witness describe a suspect's face to a police officer involves having him/her pick out the best-matching features from a large collection of images of disembodied features. Putting these together yields a putative likeness of the suspect. The mismatch between this piecemeal strategy and the more holistic facial encoding scheme that may actually be used by the brain can lead to problems in the quality of reconstructions (**Fig. 2**).

A parts-based strategy is also limited in its ability to handle degradations that reduce feature details. For instance, **Fig. 3** shows a few famous faces at very low resolutions. Whereas human observers are able to perform well even with such impoverished images[17–22], the performance of machine vision systems breaks down dramatically. The ability to interpret such inputs is of great utility in recognizing people at large distances from the camera.

Another challenge for improving recognition performance is to incorporate information from as many facial cues as possible. Current machine-based systems focus primarily on the internal features (eyes, nose and mouth) because the external features (hair and jawline) are considered too variable and too difficult to extract reliably from images. Studies of human vision, however, point to a profound significance of the external features[23–26]. **Figure 4** shows an illusion that illustrates this point[27,28]. Understanding how to represent external features in a stable manner and how to integrate them into an analysis of the overall facial structure can help devise more robust machine-based systems.

Yet another question that deserves attention is how to use prior knowledge about faces to intelligently compensate for some of the degradations in the input image. **Figure 5** shows results from one system that attempts to use 'top-down' processing to undo image degradations[7,29]. Although it serves as a good proof of concept, this system can handle relatively modest degrees of degradation and is limited in its inability to distinguish between what is an unusual but genuine facial characteristic that should be preserved and what is a degradation that ought to be removed. However, more powerful versions of this basic idea could prove to be of great practical importance.

Face recognition is one of the most active and exciting areas in neuroscience, psychology and computer vision. Although significant progress has been made on the issue of low-level image representation, the fundamental question of how to encode overall facial structure remains largely open. Machine-based systems stand to benefit from well-designed perceptual studies that can allow precise inferences to be drawn about the encoding schemes used by the human visual system, both for single snapshots and video sequences of faces. Also of use might be an analysis of the work of minimalist portrait artists, especially caricaturists, who are able to capture vivid likenesses using very few strokes. Analyzing which facial cues are preserved or enhanced in such sim-

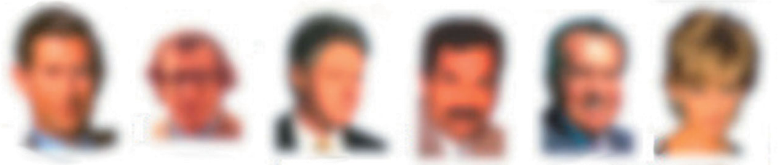

**Fig. 3.** Unlike current machine-based systems, human observers are able to handle significant degradations in face images. For instance, subjects are able to recognize more than half of all famous faces shown to them at the resolution depicted here. The individuals shown are, from left to right, Prince Charles, Woody Allen, Bill Clinton, Saddam Hussein, Richard Nixon and Princess Diana.

plified depictions can yield valuable insights about the significance of different facial attributes.

## Speaker and speech recognition

Like face recognition, the tasks of speaker identification and speech recognition have many potential practical applications. An obvious use of voiceprint analysis lies in identifying individuals. Using voiceprints as biometric cues is economically very attractive because it does not require the development of new hardware infrastructure. Given the existing telephone networks and microphones included with computers, all that is needed to deploy a voiceprint-based system is the recognition software. During recent years, several such systems have been introduced by companies including ITT, Veritel, T-NETIX and Sprint.

A speaker identification (SI) system needs to be robust against variations in the input audio signal caused by changes in a microphone's frequency response, room acoustics and background noise. Simultaneously, the system has to be capable of detecting deliberate deceit via voice impersonation. In the simplest SI scenario, a user is authenticated by having him/her speak a fixed sentence. The resulting waveform is compared with the stored one to determine if the two match. Although this idea seems straightforward, it presents two problems. First, factors such as a cold, vocal cord injury or stress can dramatically change a voiceprint, leading to mismatches with the reference. Second, the use of a fixed sentence for authentication opens up the possibility of identity theft. Someone could surreptitiously record an individual uttering the 'pass phrase' and use the recording later as a surrogate.

Secure voiceprint analysis requires the use of 'nonces'—random phrases generated at the time of user authentication. Comparisons for matching are made not at the level of the waveforms themselves, but rather between certain features abstracted from the voiceprints. The features typically used are based on modeling how specific vocal tract shapes alter the frequency content of an acoustic wave. Characteristics of the spectral shape of the acoustic signal (such as formant locations and spectral tilt) can help estimate the shape of the vocal tract and, thereby, the identity of the speaker. Several techniques have been explored for computing match scores between a novel acoustic feature vector and models of speakers' voices. These include both template models, possibly augmented with dynamic time warping[30,31], and stochastic models such as the hidden Markov model[32]. A popular scheme for representing acoustic signals is the 'mel-warped cepstrum'[32]. Motivated by perceptual experiments, this representation warps the speech spectrum to provide greater weight to the lower frequencies (less than 1.5 kHz). The frequency warping is broadly consistent with the spatial frequency organization of the auditory system.

This perceptually motivated acoustic signal representation scheme, coupled with relatively standard pattern-matching and classification back-ends, yields encouraging performance on the task of speaker identification[33]. However, these results are typically obtained on pre-recorded databases[34,35]. These corpora often do not capture the degradations and variabilities imposed on a voice signal in real settings. Therefore, more rigorous testing of current SI schemes is needed to assess their strengths and limitations and how well they mimic human abilities on this task.

We now turn to the related task of automatic speech recognition (ASR). In recent years, several companies including AT&T, IBM, Lucent, Microsoft, Philips and Speechworks have launched ASR systems. The key challenges these ASR systems face include variations in pronunciation across speakers, background acoustic noise and changes in microphone frequency characteristics. The most direct way in which studies of auditory function in biological systems have influenced ASR system design is by suggesting representation schemes that provide some stability against variations of the acoustic signal. The mel-cepstral representation mentioned above is one such scheme. Another is perceptual linear prediction (PLP)—a form of amplitude compression motivated by studies of human loudness perception[36]. Both schemes effectively smooth the spectral envelope to reflect the limited frequency resolution of the auditory system.

It seems valid to conclude that studies of biological auditory processing have influenced the 'front-ends' of ASR systems. The connections between the two fields are less direct in moving beyond the early stages of signal processing. However, some broad similarities can be discerned. ASR systems, consistent with conceptualizations of human speech processing, use both a language-model (LM) that provides top-down constraints, and 'bottom-up' phonetic classification. However, the LMs used, typically word-pair statistics, are likely to be too crude to represent the top-down influences at work in the human brain. Additionally, it is not clear how to weight the LM relative to phonetic classification. The weights are usually skewed heavily in favor of the LM. However, this seems inconsistent with human data. For instance, many

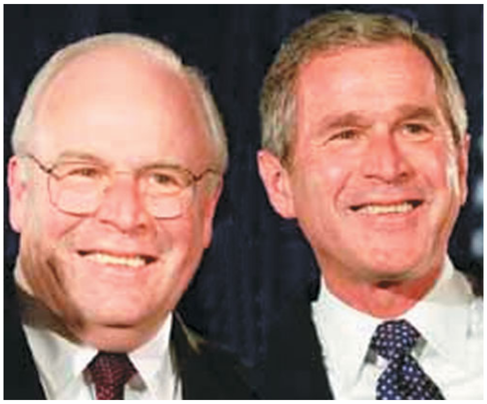

**Fig. 4.** At first glance, the image shown above seems to depict an ordinary shot of the current US president and vice-president. Closer examination reveals that the internal features of the vice-president have been supplanted by the president's features. The fact that most observers fail to notice this manipulation points to the significance of the external facial features—the hair and jawline. This illusion also highlights the importance of context. Most current machine-vision systems do not make use of either of these two sources of identity information. (From ref. 28)

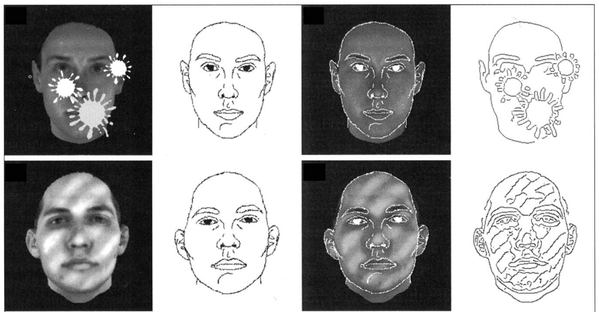

**Fig. 5.** Compensating for image degradations in a top-down manner. Knowing what faces look like (based on a training set), the system is able to ignore extraneous features and fill in missing information. From left to right, input images, the program's output line-drawing, the line-drawing overlaid on the original images for comparison, and outputs of a purely bottom-up approach to edge-detection (here the Canny edge operator). (From ref. 7).

hearing-impaired people, who experience little trouble in understanding speech in quiet conditions, are much worse in the presence of noise than would be predicted by the weights used by the ASR systems. It appears that this weighting is driven largely by the difficulty of ASR systems in reliably classifying phonetic segments due to the problems of co-articulation and lack of clean segmentation boundaries between the phonemes. Indeed, it is not entirely clear whether phonemes are the units of organization used by the brain for processing spoken language or whether other alternatives, such as syllables, might prove more appropriate[37]. These fundamental issues await further experimental investigations of human speech perception.

## Conclusion

I have broadly reviewed the issues involved in complex pattern recognition for two sensory modalities. A few common themes emerge. First, understanding the mechanisms underlying the recognition abilities of biological systems in each of the modalities has tremendous practical applicability. Second, the problems machine-based recognition systems face deal primarily with the need to robustly handle variations in the raw stimulus under different observing conditions. Current systems can function adequately only under highly controlled conditions that restrict the possible variations in input patterns. Third, the impact of studies in neuroscience has been most evident in the design of the front-ends of machine-based systems. In contrast, pattern-matching and classification stages are based largely on conventional statistical techniques without regard to neural plausibility.

The second and third themes are intimately related. With the accumulation of further insights about high-level signal representation and matching strategies used by biological systems, we can expect to see greater influence of neuroscience on machine perception endeavors. Simultaneously, these new insights will likely help alleviate the problems of artificial systems in generalizing across various stimulus transformations.

The third theme suggests that for neuroscience to have an impact on the design of machine-based systems, experimental findings have to be sufficiently specific to allow a formal implementation. Many findings about perceptual front-ends in biological systems had this characteristic—the level of description was just right to permit their translation into programs. For processes beyond the early stages, although many hypotheses and conjectures have gathered over the years, they are perhaps not defined in specific enough detail, and are not sufficiently mutually consistent, to allow a formal implementation. The challenge for neuroscientists, then, is to synthesize the body of data and hypotheses regarding recognition into realizable prescriptions for system design and also to conduct additional experiments that can lead to strong inferences about the nature of mechanisms involved in high-level perception. In this context, it is worth discussing a concern that is sometimes voiced by machine perception researchers regarding some neuroscience experiments. How do data showing a few cells or areas that respond to specific complex patterns (say, faces), help us in understanding the mechanisms underlying recognition? In other words, how does understanding the 'where' issue help us with the 'how' issue? The answer to this question lies in realizing that findings of functional localization are not merely modern-day phrenology. Rather, they open up the doors to two kinds of investigations that will eventually help answer the 'how' question. First, through a systematic probing of response properties in different parts of the brain, we may be able to infer a functional architecture that specifies how the sensory signal undergoes successive transformations leading up to recognition. Second, after identifying pattern-selective neurons, we can probe how variations in the stimulus structure change responses. Through such experiments, we can begin to infer the nature of internal representations for complex patterns in the environment.

Although experiments and hypotheses in neuroscience can profoundly facilitate progress on machine-based systems, the converse is true as well. In the process of implementing these hypotheses in artificial systems, the problems encountered and the patterns of errors obtained will suggest revisions to our conceptions regarding neural function and also ideas for experiments to help clarify previously unanticipated ambiguities.

### Acknowledgments

*For a broad ranging review such as this, one has to draw upon the expertise of several colleagues. I am grateful for generous help from Keith Kluender, Steve Greenberg, Hynek Hermansky and Paul Griffin.*

Received 12 July; Accepted 27 August 2002

1. Gould, J. L. How bees remember flower shapes. *Science* **227**, 1492–1494 (1985).
2. Herrnstein, R. J. & Loveland, D. H. Complex visual concept in the pigeon. *Science* **146**, 549–551 (1964).
3. Duda, R. & Hart, P. *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
4. Kaya, Y. & Kobayashi, K. in *Frontiers of Pattern Recognition* (ed. Watanabe, S.)265–289(Academic, New York, 1972).
5. Kanade, T. *Computer Recognition of Human Faces.* (Birkhauser, Basel and Stuttgart, 1977).
6. Campbell, R. A., Cannon, S., Jones, G. & Morgan, N. Individual face classification by computer vision. *Proc. Conf. Modeling Simulation Microcomp.* 62–63 (1987).
7. Jones, M. J., Sinha, P., Vetter, T. & Poggio, T. Top-down learning of low-level vision tasks. *Curr. Biol.* 7, 991–994 (1997).
8. Turk, M. & Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**, 71–86 (1991).
9. Belhumeur, P. N., Hespanha, J. P. & Kriegman, D. J. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 711–720 (1997).
10. Swets, D. L. & Weng, J. Discriminant analysis and eigenspace partition tree for face and object recognition from views. *Proc. Intl. Conf. Automatic Face and Gesture Recog.* 192–197 (1996).
11. Etemad, K. & Chellappa, R. Discriminant analysis for recognition of human face images. *Proc. Intl. Conf. Acoust. Speech Sign. Process.* 2148–2151 (1994).
12. Phillips, P. J., Moon, H., Rauss, P. & Rizvi, S. A. The FERET evaluation methodology for face-recognition algorithms. *IEEE Comput. Vision Pattern Recog.* 137–143 (1997).
13. Lee, T. S. Image representation using 2D Gabor wavelets. *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 959–971 (1996).

14. DeAngelis, G., Ohzawa, I. & Freeman, R. D. Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *J. Neurophysiol.* **69**, 1091–1117 (1993).
15. Bruce, V., & Young, A. *In the Eye of the Beholder: the Science of Face Perception* (Oxford Univ. Press, 1998).
16. Collishaw, S. M. & Hole, G. J. Featural and configurational processes in the recognition of faces of different familiarity. *Perception* **29**, 893–909 (2000).
17. Harmon, L. D., & Julesz, B. Masking in visual recognition: effects of two-dimensional filtered noise. *Science* **180**, 1194–1197 (1973).
18. Bachmann, T. Identification of spatially quantised tachistoscopic images of faces: how many pixels does it take to carry identity? Special issue: face recognition. *Eur. J. Cognit. Psychol.* **3**, 87–103 (1991).
19. Costen, N. P., Parker, D. M. & Craw, I. Effects of high-pass and low-pass spatial filtering on face identification. *Percept. Psychophys.* **58**, 602–612 (1996).
20. Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J. B. & Burton, M. A. Verification of face identities from images captured on video. *J. Exp. Psychol. Appl.* **5**, 339–360 (1999).
21. Yip, A. & Sinha, P. Role of color in face recognition. *Perception* **31**, 995–1003 (2002).
22. Sinha, P. Identifying perceptually significant features for recognizing faces. *Proc. SPIE Electronic Imaging Symp.* **4662**, 12–21 (2002).
23. Ellis, H. D., Shepherd, J. W. & Davies, G. M. Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition. *Perception* **8**, 431–439 (1979).
24. Shepherd, J., Davies, G. & Ellis, H. in *Perceiving and Remembering Faces* (eds. Davies, G. *et al.*) 105–132 (Academic, New York, 1981).
25. Haig, N. D. Exploring recognition with interchanged facial features. *Perception* **15**, 235–247 (1986).
26. Fraser, I. H., Craig, G. L. & Parker, D. M. Reaction time measures of feature saliency in schematic faces. *Perception* **19**, 661–673 (1990).
27. Sinha, P. & Poggio, T. I think I know that face. *Nature* **384**, 404 (1996).
28. Sinha, P. & Poggio, T. United we stand: the role of head structure in face recognition. *Perception* **31**, 133 (2002).
29. Sinha, P. & Poggio, T. in *Perceptual Learning* (ed. Fahle, M.) 273–298 (MIT Press, Cambridge, Massachusetts, 2002).
30. Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust.* **26**, 623–625 (1980).
31. Soong, F. K., Rosenberg, A. E., Rabiner, L. R. & Juang, B. H. A vector quantization approach to speaker recognition. *AT&T Technical J.* **66**, 14–26 (1987).
32. Rabiner, L. & Juang, B. H. *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, New Jersey, 1993).
33. Gish, H. & Schmidt, M. Text-independent speaker identification. *IEEE Signal Processing Mag.* **11**, 18–32 (1994).
34. Higgins, A., Bahler, L. & Porter, J. Speaker verification using randomized phrase prompting. *Digital Signal Processing* **1**, 89–106 (1991).
35. Campbell, J. P. Testing with the YOHO CD-ROM voice verification corpus. *Proc. Intl. Conf. on Acoust. Speech and Signal Processing*, 341–344 (1995).
36. Hermansky, H. Perceptual linear prediction (PLP) analysis for speech. *J. Acoust. Soc. Am.* **87**, 1738–1752 (1990).
37. Greenberg, S. Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation. *Speech Commun.* **29**, 159–176 (1999).