# Multimodal Person Recognition for Human–Vehicle Interaction

**Engin Erzin, Yücel Yemez, and A. Murat Tekalp**
*Koç University, Turkey*

**Aytül Erçil, Hakan Erdogan, and Hüseyin Abut**
*Sabanci University, Istanbul*

**Next-generation vehicles will undoubtedly feature biometric person recognition as part of an effort to improve the driving experience. Today's technology prevents such systems from operating satisfactorily under adverse conditions. A proposed framework for achieving person recognition successfully combines different biometric modalities, borne out in two case studies.**

Over the past 30 years, the field of biometric person recognition—recognizing individuals according to their physical and behavioral characteristics—has undergone significant progress. Next-generation human–vehicle interfaces will likely incorporate biometric person recognition, using speech, video, images, and analog driver behavior signals to provide more efficient and safer vehicle operation, as well as pervasive and secure in-vehicle communication. Yet, technical and deployment limits hamper these systems' ability to perform satisfactorily in real-world settings under adverse conditions. For instance, environmental noise and changes in acoustic and microphone conditions can significantly degrade speaker recognition performance. Similarly, factors such as illumination and background variation, camera resolution and angle, and facial expressions contribute to performance loss in visually identifying a person. Biometric person recognition in vehicles is especially likely to challenge researchers because of difficulties posed by the vehicle's interior compartment as well as by economics.

In this article, we present an overview of multimodal in-vehicle person recognition technologies. We demonstrate, through a discussion of our proposed framework, that the levels of accuracy required for person recognition can be achieved by fusing multiple modalities. We discuss techniques and prominent research efforts, and we present the results of two case studies we conducted. The sidebar, "Solutions for In-Vehicle Person Recognition," discusses related work.

## Why in-vehicle person recognition?

To improve the driving experience, making it better and safer, manufacturers are making vehicles smarter. As vehicles become smarter, information processing from vehicle sensors will become more complex, as will vehicle personalization, which involves adapting the driver and passenger compartment for safer driving and travel.

Personalization features will, for example, increase vehicle safety by determining whether the person behind the wheel is an authorized driver (say, the vehicle's legal owner). If so, the driver will be able to operate the vehicle. If the individual isn't authorized, the vehicle will prevent operation and could communicate with authorities to report the incident and initiate an enforcement procedure, such as preventing the driver from starting the ignition.

Personalization will promote safe driving by monitoring a driver's behavior. Once the human–vehicle interface has identified a driver and determined road and traffic conditions, the vehicle can monitor the driver's behavioral signals from braking, accelerating, or swerving. With these signals, the human–vehicle interface can verify if the driver is alert, sleepy, or drunk.

We also envision that vehicle personalization will enable secure transactions. In today's increasingly mobile and connected society, we'll need—or want—to do more transactions anywhere we can, including inside a vehicle. Secure transactions might involve travel planning and arrangements, and mobile banking, database access, and shopping, all of which require varying levels of personal authentication. Unfortunately, a vehicle compartment isn't communication-friendly and pervasive; thus, it poses major challenges to secure communication.

## The person recognition problem

Biometric person recognition can be formulated as either an authentication or identification problem. In an authentication situation, an

1070-986X/06/$20.00 © 2006 IEEE

## Solutions for In-Vehicle Person Recognition

Researchers, who've extensively studied biometric person recognition for more than 20 years, have developed technologies with varying degrees of success.[1-3] Most promising systems, despite performing well in controlled environments, suffered significantly when deployed in challenging environments such as an airplane cockpit or a moving vehicle. (A note about terminology: We use the term *speaker recognition* if the modality is only speech or an audio signal, otherwise we use the term *person recognition*. Recognition refers to both authentication and open or closed set identification.)

Potential solutions have included

- audio (speech)-only person recognition with speech enhancement or with robust feature extraction from noisy speech;
- video-only person recognition with enhancement at the image or feature level;
- recognition based on signals related to a person's driving behavior; and
- person recognition based on combinations of audio, video, and signal information in a multimodal framework.

Of particular interest is a roadmap that researchers have identified for fusing, or combining, physical and behavioral sensor information.[4] In their experiments, conducted in a controlled laboratory setting, researchers worked with three sets of physical features: faces, hand geometry, and fingerprints. Although hand geometry isn't practical as a means of recognizing someone in a vehicle, similar studies could be used as benchmarks for applications under more adverse conditions.

Audio is probably the most natural nonintrusive modality to identify a person for in-vehicle applications, although video also contains important biometric information—still frames of faces and temporal lip motion information tightly correlate with audio. But, at present, most speaker recognition systems use only audio data.[2] Under noisy conditions, of course, such systems are far from perfect for high-security applications, an observation that's equally valid for systems using only visual data. Poor picture quality, changes in pose and lighting conditions, inclement weather conditions, or varying facial expressions may significantly degrade person recognition performance.[1,5]

To overcome the limitations of audio and video data, and in response to the increased global awareness for personal, institutional, and information security, researchers have initiated several large-scale programs in academia and in industrial R&D. One difficulty facing researchers is the lack of generally accepted databases for in-vehicle person recognition applications. As a result, researchers have been using other databases or they've been mimicking the vehicle compartment, artificially adding noise to "clean" speech or video. Their findings, therefore, don't truly reflect real-life in-vehicle scenarios.

With funding from the Japanese government and industry, Itakura et al. at Nagoya University's Center for Acoustic Information Research (CIAIR) have embarked on a megaproject called "Construction and Analysis of the Multi-Layered In-Car Spoken Dialogue Corpus."[6,7] This project uses 12 audio and 3 video channels of data; the researchers have also collected analog driver behavior signals from five different sensors and location information from 812 male and female drivers, resulting in a databank measured in terabytes. Although it's primarily in Japanese, a number of groups, including the authors of this article, actively use this database under an international research framework called the International Alliance for Advanced Studies on In-Car Human Behavioral Signals.[8,9]

Other initiatives include iCU-Move at Colorado University at Boulder,[10] In-Car Interaction System at the Center for Scientific and Technological Research (ITC-IRST) in Trento, Italy (see http://www.itc.it/irst), and Avicar: An Audiovisual Speech Corpus in a Car Environment, in Illinois.[11] Yet another initiative is a European Union Network of Excellence called Similar (see http://www.similar.cc), whose name essentially derives from "The European taskforce creating human–machine interfaces similar to human–human communication." Our work described in this article is one of the application areas in that project.

### References

1. W. Zhao and R. Chellappa, Eds., Face Processing: Advanced Modeling and Methods, Academic Press, 2005
2. J. Campbell, "Speaker Recognition: A Tutorial," *Proc. IEEE*, vol. 85, no. 9, 1997, pp. 1437-1462.
3. R. Snelick et al., "Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, 2005, pp. 450-455.
4. A. Ross and A. Jain, "Information Fusion in Biometrics," *Pattern Recognition Letters*, vol. 24, issue 13, 2003, pp. 2115-2125.
5. M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, 1991, pp. 586-591.
6. K. Igarashi et al., "Biometric Identification Using Driving Behavior," *Proc. IEEE Int'l Conf. Multimedia & Expo* (ICME 04), vol. 1, IEEE Press, 2004, pp. 65-68.
7. N. Kawaguchi et al., "Construction and Analysis of the Multi-Layered In-Car Spoken Dialogue Corpus," *DSP in Vehicular and Mobile Systems*, H. Abut, J.H.L. Hansen, and K. Takeda, eds., Springer, 2005,pp. 1-18.
8. H. Abut, J.H.L. Hansen, and K. Takeda, eds., *DSP in Vehicular and Mobile Systems*, Springer, 2005.
9. H. Erdogan et al., "Multimodal Person Recognition for Vehicular Applications," *Proc. 6th Int'l Workshop on Multiple Classifier Systems,* LNCS 3541, Springer, 2005, pp. 366-375.
10. H. Abut, J.H.L. Hansen, and K. Takeda, eds., *DSP in Vehicular and Mobile Systems*, Springer, 2005, pp. 19-46.
11. B. Lee et al., "AVICAR: An Audiovisual Speech Corpus in a Car Environment," *Proc. 8th Int'l Conf. Spoken Language Processing* (ICSLP 04), Int'l Speech Communication Assoc., 2004, pp. 2489-2492.

unknown person claims an identity and requests access to a service. The claimed identity is verified using his or her model in a known pool of subjects. If the person's biometric features match those of the claimed identity in the database, access is granted, otherwise it's rejected.

In an identification situation, we can further classify biometric person recognition as either closed- or open-set:

- *Closed-set identification*: An unknown person requests access to a service without an explicit identity claim. The subject is classified as one of *N* registered people with the most similar biometric features in the database. Access is granted with the personalized profile of the best match. A reject scenario isn't defined and impostors aren't handled.
- *Open-set identification*: An unknown person requests access to a service without an explicit identity claim. Unlike closed-set identification, this situation includes a reject scenario—the system identifies the person if there's a match with one of *N* registered people, and rejects otherwise. Hence, the problem becomes an $N + 1$ class identification problem, including a reject class. Note that the authentication problem can be considered as a special case of the open-set identification problem with $N = 1$.

The open-set identification scenario is well suited for a vehicle access application in which several drivers might be authorized to access a single car. In the authentication scenario, on the other hand, we're more interested in controlling the driving behavior to assure safe driving for a specific, already identified, driver. We've conducted case studies on these two applications, which we discuss later.

## Multimodal feature extraction and classification

To design an in-vehicle multimodal person recognition system we must do two things:

- Decide which modalities to employ and choose the best feature set representing each.
- Determine how to fuse multiple feature sets extracted from different modalities.

In multimodal person recognition, the word *modality* typically refers to information that can be deduced from biometric signals. For instance, the video signal can be split into different modalities, such as face texture and motion. A speaker's characteristic motion is lip movement, which is tightly correlated with speech. Gesture (or gait) motion, on the other hand, could also be characteristic but probably less significant for speaker recognition. Iris, hand geometry, and fingerprints can also be used to identify a person but they require a cooperative user, and their acquisition is more intrusive. Finally, more-application-specific biometrics, such as analog signals associated with driver behavior, can provide valuable modalities for in-vehicle applications. These include acceleration and brake pedal pressures, steering wheel dynamics, vehicle speed variations, and engine speed.

Person recognition applications require that we represent each modality's raw biometric data with a discriminative, low-dimensional set of features, together with the best matching metric to classify each modality. This step usually includes a training phase through which we represent each person (or class) with a statistical model or a representative feature set. The most important criteria in selecting the feature set and classification methodology for each modality are dimensionality, computational efficiency, robustness, invariance, and discrimination capability.

Next, we discuss four modalities: audio (speech), face texture, lip motion, and driving behavior.

### Audio (speech)

We can classify speaker recognition tasks according to their text dependence. Text-independent systems assume no prior knowledge of input information, whereas text-dependent systems constrain the speaker to speak a set of personal password phrases for identification. Text-dependent systems can be more accurate; text-independent systems can be more flexible and assume less user cooperation.

State-of-the-art systems use hidden Markov models (HMMs) for text-dependent—and Gaussian mixture models (GMMs) for text-independent—speaker recognition.[1] Recently, researchers introduced multi-grained GMM modeling, which uses separate GMMs for phone groups.[2] These statistical models are trained on acoustic features extracted from the short-time spectrum of speech signals. From numerous experiments with different feature sets, researchers have found Mel-frequency cepstral coefficients (MFCCs) to be more robust and effective compared to, for example, various derivatives of linear predictive coefficients.[1] In addition, other researchers are pursuing speaker recognition using a support vector machine based on MFCC and its differences.[3]

Speaker recognition in a moving vehicle, which is a unique environment, requires robust algorithms that can withstand background noise and acoustic echoes. We can achieve noise robustness by preprocessing speech to remove background noise (a technique called speech enhancement) before extracting the features. Researchers have traditionally done this by spectral subtraction,[4] Wiener filtering,[5] or—more recently—by nonlinear and adaptive noise-removal techniques.[6] Recently, the European Telecommunications Standards Institute published an advanced front-end standard (ES 202 050) for speech recognition applications, which yields promising performance gains under noise using two-stage Wiener filter-based noise reduction.[7] For noise robustness, it could be beneficial to use beam-forming in a microphone array to combine information from each microphone to decrease the noise effects.[5]

The speaker recognition community generally prefers to perform feature and score normalization for noise robustness, usually without preprocessing the speech data. Several feature normalization methods modify extracted features to be more noise-robust and invariant. These include cepstral mean and variance normalization, and distribution normalization such as feature warping and short-time Gaussianization.[8] The quest to extract more noise-robust features from noisy speech as compared to MFCCs is ongoing; however, one alternative could be to use articulatory features which capture the shape, place, and manner of the speech production organs.[9]

In speech recognition, systems trained on clean speech can be adapted to different channels and noisy environments by affine feature and model transformations.[1] It's also possible to use explicit noise models to model noisy speech.[10] These model adaptation techniques aren't as popular in speaker recognition, in which researchers more easily induce a similar effect by performing score normalization—for example, using cohort models, Z-norm, and T-norm normalization methods.[8]

Previous work on noise-robust speaker and speech recognition focused mostly on circuit- or packet-switched telephony channel and noise types. In a vehicular application, the audio isn't sent over a telephone channel to a processing center but instead processed within the vehicle. An in-vehicle application's focus, therefore, should be to handle vehicular noise and channel effects within the vehicle.

### Face texture

Face texture is widely accepted as one of the most common biometric characteristics used for person recognition. Many proposed methods have been based on image intensities; see Zhao et al. for a comprehensive review.[11] Most popularly, face recognition approaches are based on either the location and shape of facial attributes (such as the eyes, nose, and lips and their spatial relationships);[12,13] the overall (global) analysis of a face image representing the face as a weighted combination of numerous canonical faces;[14] or on hybrid methods. Phillips et al. conducted a comprehensive evaluation of the Facial Recognition Technology (Feret) Database[15] to evaluate different systems using the same image database. They found that the neural network method based on Elastic Bunch Graph Matching (EBGM),[16] the statistical method based on subspace linear discriminant analysis (LDA), and the probabilistic principal component analysis (PCA) method[17] were the top-performing methods, each showing different levels of performance on different subsets of images.

It's difficult to recognize people from in-vehicle video for several reasons. One difficulty arises because, in vehicles, the subjects—especially the driver—aren't expected to pose for a camera since their first priority is to operate the vehicle safely. Large illumination and pose variations can occur as a result. Additionally, partial occlusions and disguises are common. Changes induced by illumination are often larger than the differences between individuals, causing systems based on image comparison to misclassify input images. Researchers observed these changes using a data set of 25 individuals[18] and theoretically proved for systems based on eigenface projection.[11] In most algorithms evaluated under Feret, changing the illumination resulted in a significant performance loss. For some algorithms, this loss was equivalent to comparing images taken a year and a half apart. Changing facial position can also affect performance: A 15-degree difference in position between the query image and the database image will degrade performance. At a difference of 45 degrees, recognition is ineffective.

A second difficulty in identifying subjects in vehicles is the low spatial resolution and video

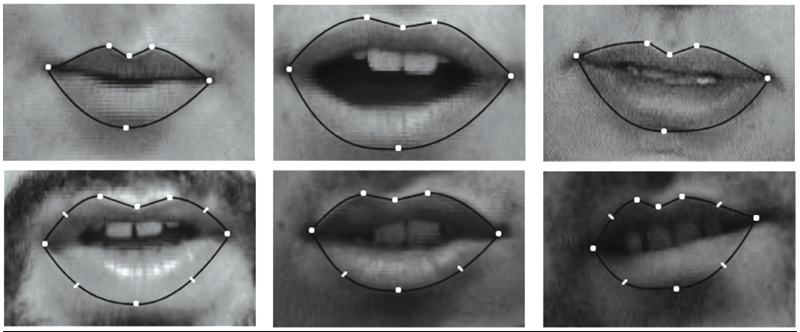


Figure 1. Lip tracking examples taken from the Multimedia, Vision, and Graphics Laboratory Audiovisual Database, acquired in a controlled environment. The results in the second row needed user intervention to hand-label some assisting points on the lip contour because of color ambiguity. Those in the first row were obtained automatically.

quality in cars. Because of the acquisition conditions, the face images are smaller (sometimes much smaller) than the assumed sizes in most existing still-image-based face recognition systems. Small images not only make the recognition task difficult, but they also affect face segmentation accuracy as well as the detection accuracy of fiducial points or landmarks often required (at least for registering face images) in recognition methods. For example, the EBGM-based system,[16] one of the top three performers under Feret, requires a large image, for example, 128 × 128, which severely restricts its application to video-based surveillance, where face images are smaller.

Despite these disadvantages, video-based recognition has three advantages over still images:

- *Video provides abundant data*. We can, therefore, select good frames on which to perform classification.

- *Video allows face tracking*. Accordingly, we can compensate for phenomena such as facial expressions and pose changes, resulting in improved recognition. In-car visual tracking is easier because drivers are expected to remain in a fixed location in the car.

- *Video provides temporal continuity*. This lets us use super-resolution techniques to improve the quality of face images.

Although many face recognition algorithms work well in constrained environments, face recognition is an open and challenging problem for in-vehicle applications.

### Lip motion

Researchers have commonly used lip information in speech recognition.[19] Lip motion correlates highly with the speech signal, and lip-reading reveals speech content. For speech recognition, it's usually sufficient to extract the principal components of lip movement and establish a one-to-one correspondence with speech phonemes and lip movement visemes. For person recognition, however, the use of lip motion might require more sophisticated processing,[20,21] because the principal components of lip movement aren't usually sufficient to discriminate a speaker's biometric properties. Researchers might need to model high-frequency or nonprincipal components of lip motion to model the biometrics—that is, a speaker's specific lip movements rather than the speaker's words. The success of a lip-based person recognition system depends largely on the accuracy and precision of the lip tracking or lip motion estimation procedure. For instance, parametric lip models commonly used for lip contour tracking fail to adequately represent the discriminative biometric details specific to a speaker. Figure 1 shows examples of lip tracking, which we obtained with parametric model fitting using the technique presented by Eveno et al.[22]

The audiovisual speech and speaker recognition literature examines three basic alternatives for initial representation of lip motion features. These alternatives use

- raw pixel intensity values on a rectangular grid about the mouth region,[19,23]

- motion vectors instead of intensity values,[20] and

- lip shape parameters.[21]

The last option would seem to be the most powerful representation, provided that the lip contour can be accurately tracked. However, this task is challenging in adverse circumstances, such as in a moving vehicle, because lip-contour-tracking algorithms are generally sensitive to light conditions and image quality. In such cases, detecting the rectangular mouth region is relatively easier to accomplish. Thus, the first two alternatives—raw pixel intensity values and motion vectors—should be suitable for in-vehicle applications.

Typically, the dimension of the initial lip motion feature vector is reduced by subspace transform techniques such as discrete cosine transform (DCT),[23] and subjected to an analysis via techniques like LDA.[20] A lip-based speaker identification system's success eventually depends on how much of the discriminative information is retained in the final reduced-dimensional feature set.

### Driving behavior

Can drivers be identified from their driving behavior? Together with researchers at CIAIR, we've been studying pressure readings from accelerator and brake pedals as well as vehicle speed variations[5] to see if our driving behavior is unique and, if so, to see if we could use this knowledge for driver recognition. We selected five driving signals, based on automotive industry recommendations and results of preliminary experiments, as candidates in the CIAIR data collection initiative. The signals are accelerator pedal pressure and brake pedal pressure readings in kilogram force per square centimeter (kgf/cm$^2$), vehicle speed in kilometers per hour (km/h), engine speed in revolutions per minute (rpm), and steering angle within the range of –720 to +720 degrees. The CIAIR database sampled the signals from these sensors at 1.0 kHz and obtained the location information from a differential GPS at one reading per second. Table 1 summarizes the data specifications.[5,24,25]

*Table 1. Data collection specifications.*

| Data source | Data characteristics |
|---|---|
| Speech | Sampling: 16 kHz; 16 bits per sample; 12 channels |
| Video | MPEG-1; 29.97 frames per second; 3 channels |
| Driving signals | Acceleration, accelerator pedal pressure, brake pedal pressure, steering wheel angle, engine rpm, and vehicle speed: each at 16 bits/sample and 1.0 kHz |
| Location | Differential GPS: one reading per second |

Initially, we explored methods based on fast Fourier transform, interdriver, and intradriver distributions; we also explored multidimensional, multichannel extensions of the linear predictive theory. We had limited success in identifying a driver.[24,26] Later, to represent each driver's characteristics, we employed Gaussian mixture modeling, a technique regularly and successfully employed in speaker modeling. We used smoothed and subsampled driving signals (acceleration and brake pedal pressures) and their first derivatives as features for statistical modeling. Our initial findings, based only on driving signals, are encouraging for genuine versus impostor modeling for driver authentication.[24]

Driver authentication using driving signals can't be used for determining whether the driver is an authorized driver or not. Driver authentication should be done before the vehicle moves. However, we can use driving behavior signals to verify an authorized driver's driving condition in a safe driving scenario. Assuming that the driver has already been authenticated, the driving behavior signals let us verify whether the driver is alert, sleepy, or drunk. If we determine that the driver isn't driving normally, we could deploy active and passive safety enforcement systems. In addition, driver signals could be useful for forensic purposes to identify drivers in a stolen-and-found vehicle or after a crash when audiovisual sensors aren't available or can't be relied on.

## Open-Set Unimodal Person Identification

The maximum a posteriori probability solution to the $N$-person open-set problem requires computing $P(\lambda_n \mid \mathbf{f})$ for each class $\lambda_n$, $n = 1, \ldots, N, N+1$, given a feature vector $\mathbf{f}$ representing the sample data of an unknown individual. An alternative is to employ the maximum likelihood solution, which maximizes the class-conditional probability, $P(\mathbf{f} \mid \lambda_n)$, for $n = 1, \ldots, N, N+1$. Because it's difficult to accurately model the imposter class, $\lambda_{N+1}$, we can use the following approach, which includes a reject strategy through the definition of the likelihood ratio:

$$\rho(\lambda_n) = \log \frac{P(\mathbf{f} \mid \lambda_n)}{P(\mathbf{f} \mid \lambda_{N+1})} = \log P(\mathbf{f} \mid \lambda_n) - \log P(\mathbf{f} \mid \lambda_{N+1})$$

We can implement the decision strategy in two steps. First, we determine

$$\lambda^* = \arg \max_{\lambda_1, \ldots, \lambda_N} \rho(\lambda_n)$$

and then

$$\text{if } \rho(\lambda^*) \geq \tau \text{ accept; otherwise reject}$$

where $\tau$ is the optimal threshold that's usually determined experimentally to achieve the desired accept and false reject rates.

Computation of class-conditional probabilities needs a prior modeling step, through which we estimate a probability density function of feature vectors for each class $n = 1, \ldots, N, N+1$ from the available training data. A common and effective approach for modeling the impostor class is to use a universal background model, which we estimate by using all available training data regardless of which class they belong to.

### Multimodal person recognition

The person recognition problem is often formalized in a probabilistic framework,[27] reviewed in the "Open-Set Unimodal Person Identification" sidebar. Here, we discuss extensions of this

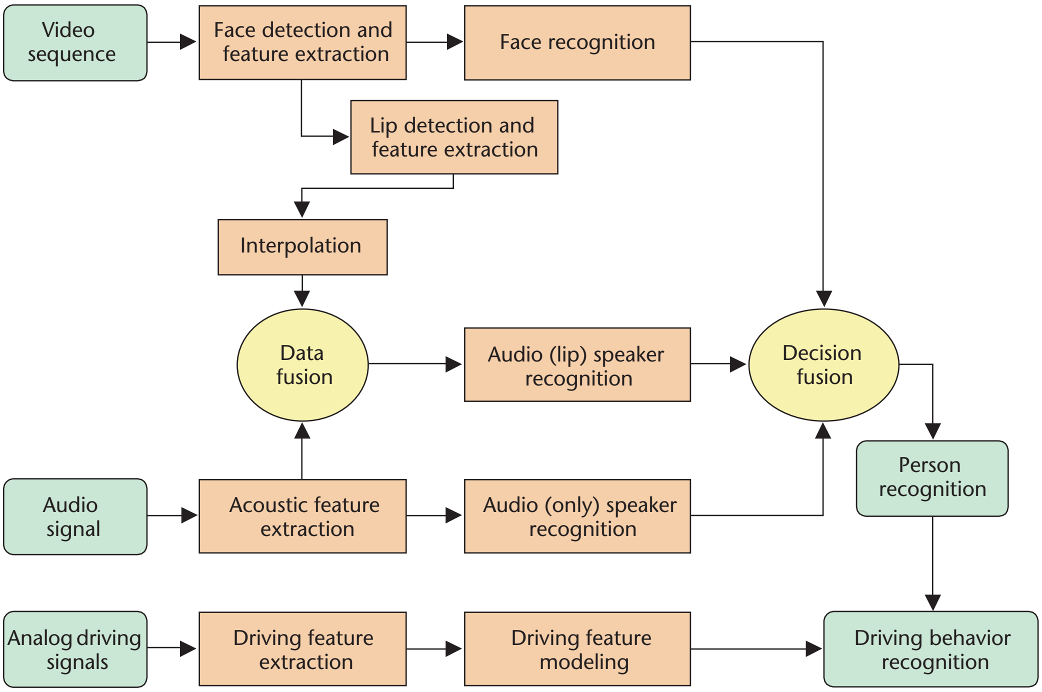

Video sequence
Face detection and feature extraction
Face recognition
Lip detection and feature extraction
Interpolation
Data fusion
Audio (lip) speaker recognition
Decision fusion
Person recognition
Audio signal
Acoustic feature extraction
Audio (only) speaker recognition
Analog driving signals
Driving feature extraction
Driving feature modeling
Driving behavior recognition

*Figure 2. Multimodal decision fusion for in-vehicle person recognition.*

framework for multimodal person recognition.

The multimodal person recognition framework considers features and decisions from audio, video, and analog driving signals. Figure 2 shows the framework we propose for in-vehicle person identification and authentication, consisting of three sensor inputs (video sequence, audio signal, and analog driving signal measurements) and two fusion techniques (data fusion and decision fusion). We implement data fusion by concatenating correlated speech and lip motion features prior to the decision fusion stage.

### Decision fusion versus data fusion

The main objective in multimodal fusion is to compensate for possible misclassification errors resulting from a given modality classifier with other available modalities and to achieve a more reliable overall decision. Different strategies for multimodal fusion are possible: In the so-called early-integration modalities, fusion is implemented at the data or feature level.[19] In late-integration modalities, decisions or scores resulting from each unimodal classification are combined to arrive at a conclusion.[27,28] This latter strategy—combining decisions—is also referred to as decision or opinion fusion. It's especially effective when contributing modalities aren't correlated and resulting partial decisions are statistically independent. Early integration techniques, on the other hand, might be preferable if modalities are tightly correlated, as in the fusion of audio and lip movement. We can also view multimodal decision fusion more broadly as a way of combining classifiers, which is a well-studied problem in pattern recognition.[28] We obtain best results typically through a combination of data and decision fusion in a single framework, as Figure 2 shows.

### Multimodal decision fusion

Here, we consider two decision fusion techniques, reliability weighted summation and the adaptive cascade rule.

Suppose that $P$ different classifiers, one for each of the $P$ modalities $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_P$ are available, and the $p$th classifier produces a set of $N$-class log-likelihood ratios $\rho_p(\lambda_n)$, $n = 1, 2, \ldots, N$. The problem then reduces to computing a single set of joint log-likelihood ratios, $\rho(\lambda_1), \rho(\lambda_2), \ldots \rho(\lambda_n)$, for each class. The most generic way of computing joint ratios (or scores) can be expressed as a weighted summation:

$$\rho(\lambda_n) = \sum_{p=1}^{P} \omega_p \rho_p(\lambda_n), \quad n = 1,2,\ldots,N \tag{1}$$

where $\omega_p$ denotes the weighting coefficient for modality $p$, such that $\sum_p \omega_p = 1$. The decision fusion problem then becomes a coefficient optimization task. Note that when $\omega_p = 1/P$ for all $p$, Equation 1 is equivalent to the product rule, which is the optimal setting in the Bayesian sense, given that the classifier decisions are statistically independent and free of modeling and measurement errors. However, in practice, this isn't generally the case, and the optimal assignment of weights still remains an important problem.

Score normalization is another critical issue in multimodal fusion schemes because the statistics and the numerical range of the likelihood scores can vary from one modality classifier to another. In the literature, researchers describe several different ways to achieve score normalization.[29] In our work, we used the sigmoid and variance normalization[23] to map the likelihood scores coming from each separate modality into a (0, 1) interval before the fusion process.

**Reliability-weighted summation (RWS)**. Most classifier fusion schemes[28,30] vary in the way they interpret the weighting coefficients in Equation 1. Hard-level fusion techniques, such as max rule, min rule, and median rule,[28] use binary values for the weighting coefficients. Soft-level fusion techniques, on the other hand, use reliability values as the weighting coefficients and compute a weighted average of the classifier output scores. We describe this fusion scheme in the "Reliability-Weighted Summation Rule" sidebar.

In a multimodal fusion system, environmental noise, modeling errors, and time-varying characteristics of signals at hand might corrupt some modalities, yielding erroneous likelihood scores. It's critical, therefore, that a multimodal fusion system be able to assign a reliability measure to each modality and incorporate this information into the fusion scheme.

Two main approaches exist for estimating a modality's reliability. In one approach, the data is analyzed from which the feature vector is extracted. Techniques based on this approach estimate how much the observed data deviates from an estimated distribution model.[19] In practice, however, the source of statistical deviation varies and is difficult to model because of such factors as acoustic or visual noise, time-varying characteristics of signals, and lighting and pose variations for visual data. In an alternative to this method, the statistics and rank correlation of the resulting likelihood scores are analyzed.[21] Although this more general approach addresses all possible types of corruption, techniques based on this approach aren't generally designed to measure the reliability of a modality classifier's reject decisions. Among the available reliability measures, we prefer the one we proposed in our implementation of the RWS rule,[23] because it's better suited to the open-set speaker identification problem, assessing both accept and reject decisions of a classifier.

## Reliability-Weighted Summation Rule

The Reliability Weighted Summation (RWS) rule combines likelihood ratio values of the *P* modalities weighted by their reliability values,

$$\rho(\lambda_n) = \sum_{p=1}^{P} R_p \rho_p(\lambda_n), \quad n = 1, 2, \ldots, N$$

where $\lambda_n$ denotes class *n*. The resulting joint likelihood ratio, $\rho(\lambda_n)$, can then be used as described in the main text's accompanying "Open-Set Unimodal Person Identification" sidebar.

We estimate the reliability value $R_p$ based on the difference of likelihood ratios of the best two candidate classes $\lambda^*$ and $\lambda^{**}$, that is, $\Delta_p = \rho_p(\lambda^*) - \rho_p(\lambda^{**})$ for modality *p*. In the presence of a reject class, we would expect a high likelihood ratio $\rho_p(\lambda^*)$ and a high $\Delta_p$ value for the true accept decisions, but a low likelihood ratio $\rho_p(\lambda^*)$ and a low $\Delta_p$ value for true reject decisions. Accordingly, we can estimate a normalized reliability measure $R_p$ by

$$R_p = \frac{\gamma_p}{\sum_i \gamma_i}$$

where

$$\gamma_p = \left(e^{\left(\rho_p(\lambda^*)+\Delta_p\right)} - 1\right) + \left(e^{\left(\kappa-\rho_p(\lambda^*)-\Delta_p\right)} - 1\right)$$

The first and second terms in $\gamma_p$ are associated with the true accept and true reject decisions, respectively. The symbol $\kappa$, $0 < \kappa < 2$, stands for an experimentally determined factor to reach the best compromise between accept and reject scenarios in a given training data set.

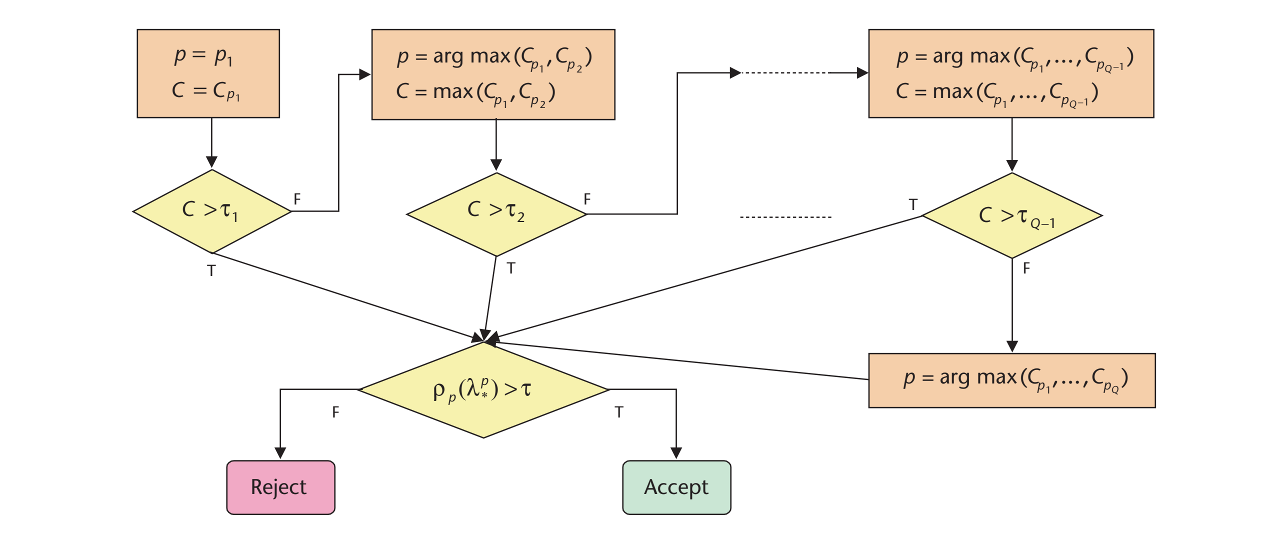
**Adaptive cascade rule**. There's no formal justification that the RWS rule will probabilistically result in a minimum error classifier. Consequently, we've proposed the adaptive cascade rule as an alternative method for multiple-modality fusion.[23]

In this rule, we assume *P* different classifiers, each associated with a single modality. Theoretically, it's possible to create a total of $Q = 2^{P-1}$ classifier combinations, including *P* unimodal classifiers and $2^{P-1} - P$ for combined modalities. Each multimodal combination corresponds to a classifier that produces another set of likelihood scores by some linear combination, for example, the RWS rule, of the corresponding likelihood scores. In a multimodal fusion system, results obtained by linear fusion of two modalities may

## Adaptive Cascade Rule

The adaptive cascade rule employs a cascade of $Q = 2^{P-1}$ classifier combinations with $Q$ reliability measures $R_1$, $R_2$, …, $R_Q$ and $Q$ confidence measures $C_1$, $C_2$, …, $C_Q$. The confidence measure for the decision of classifier $p$ is defined as the absolute difference between the best likelihood score $\rho_p(\lambda^*)$ and the threshold $\tau$ (see the third equation in the main text's sidebar, "Open-Set Unimodal Person Identification"). The order in the classifier cascade $\{p_i\}$ is then arranged so that $R_{p1} \geq R_{p2} \geq \ldots \geq R_{pQ}$. This ordering implicitly defines a priority on each modality or modality combination. Starting with the most reliable classifier $p_1$, the cascade rule successively searches for a decision with a sufficiently high confidence. As soon as a classifier with sufficiently high confidence measure is encountered, the decision cascade concludes with an accept or reject decision. Figure A depicts the adaptive cascade rule's structure.

The adaptive cascade rule uses $Q$ confidence thresholds $\tau$ and $\tau_1$, $\tau_2$, …, $\tau_{Q-1}$, each of which has to be determined experimentally to achieve the desired error rate. We can reduce the algorithmic complexity by considering only a few classifiers (usually three is sufficient) each time, depending on the reliability order that varies from one decision instant to another.



$p = p_1$
$C = C_{p_1}$
$C > \tau_1$
F
T
$p = \arg\max(C_{p_1}, C_{p_2})$
$C = \max(C_{p_1}, C_{p_2})$
$C > \tau_2$
F
T
$p = \arg\max(C_{p_1}, \ldots, C_{p_{Q-1}})$
$C = \max(C_{p_1}, \ldots, C_{p_{Q-1}})$
T
$C > \tau_{Q-1}$
F
$p = \arg\max(C_{p_1}, \ldots, C_{p_Q})$
$\rho_p(\lambda^p_*) > \tau$
F
T
Reject
Accept

*Figure A. The adaptive cascade rule's structure.*

in general outperform those obtained from each modality alone. However, on other occasions, a single highly reliable modality alone might yield a correct decision, whereas its fusion with some other less reliable modality would give incorrect results.

We can implement the adaptive cascade rule through a reliability-ordered cascade of classifier combinations, as the "Adaptive Cascade Rule" sidebar illustrates. Reliability is thus regarded as a means of giving priority to some single or combined modality in the fusion process, rather than using it as a numerical weight. In particular, this rule chooses the decision of the most reliable of the classifier combinations having sufficiently high confidence and disregards the others. We need two measures, therefore—one for reliability and one for confidence. In the context of person recognition, *reliability* measures how much one can rely on a classifier's decision, whereas *confidence* measures how confident a classifier is of its own decision. We can show that, under the adaptive cascade rule, the upper bound for the system error rate becomes the expected occurrence rate of the cases where all classifier combinations fail.[23]

### Case studies

To test our proposed framework, we conducted two case studies. The first addressed multimodal person identification using two different multimodal data sets. We collected the first data set in the Multimedia, Vision, and Graphics Laboratory (MVGL) of Koç University,[23] which we later artificially contaminated with car noise. For our second data set, we used a subset of the CIAIR vehicular corpus.[25] The second case study addressed authentication of the driving behavior of known drivers. In that study, we used driving signals available with the CIAIR database.

### Case study 1: Identification

The quality and spatial resolution of the MVGL database's visual data are substantially higher than those of the CIAIR database. We used audio (speech) and still-face image modalities on both databases. With the MVGL data set, we were able to track lip position on high-quality, high-resolution videos and incorporated it as a separate modality in addition to audio and face.

We evaluated the open-set identification performance via the equal error rate (EER) measure. The EER is calculated as the operating point, where false accept rate (FAR) equals false reject rate (FRR). False accept and false reject rates are defined as

$$\text{FAR} = 100 \times \frac{\text{\# of false accepts}}{N_{\text{a}} + N_r} \text{ and}$$

$$\text{FRR} = 100 \times \frac{\text{\# of false rejects}}{N_a}$$

where $N_a$ and $N_r$ are the total number of trials for the genuine and imposter clients in the testing, respectively.

**MVGL database examples**. The MVGL database includes 50 subjects, in which each subject utters 10 repetitions of his or her name as the password phrase. A set of impostor data is also available for each subject in the population uttering five different names from the population. We considered a multimodal open-set speaker identification system that integrated information coming from audio, face, and lip motion modalities. We used the adaptive cascade and RWS rules to compare the fusion of multiple modalities.

The audio modality, A, is represented using a set of 39 features: 13 MFCCs along with the first- and second-order derivatives. We obtained the multistream audio-lip features, AL, by concatenating the audio MFCCs and the 2D-DCT coefficients of the lip intensity for each frame. To achieve temporal characterization, we used an HMM-based classifier. For face, F, identification, we used the eigenface technique applied to a set of video images for each speaker. Table 2 presents the unimodal and multimodal equal error rate performances that we obtained, with car noise present, with the RWS rule (⊕) and the adaptive cascade rule (∗).

*Table 2. Open-set speaker identification results using the Multimedia, Vision, and Graphics Laboratory database.*

| Equal Error Rate (%) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Modality | Noise Level (dB SNR) | | | | | | |
| (M = fused modality) | Clean | 20 | 10 | 0 | –5 | –10 | –15 |
| A | 2.4 | 2.4 | 2.5 | 3.7 | 5.6 | 12.1 | 29.5 |
| F | | | | 8.4 | | | |
| AL | 13.6 | 13.6 | 13.6 | 13.8 | 14.0 | 14.8 | 15.3 |
| $M_0 = A \oplus F$ | 0.4 | 0.4 | 0.4 | 0.9 | 1.4 | 3.0 | 8.6 |
| $M_1 = F \oplus AL$ | 3.2 | 3.2 | 3.2 | 3.3 | 3.4 | 3.7 | 3.9 |
| $M_2 = A \oplus F \oplus AL$ | 0.6 | 0.6 | 0.6 | 0.7 | 1.0 | 1.2 | 3.5 |
| $M_0 \oplus M_1 \oplus M_2 \oplus A \oplus F$ | 0.4 | 0.4 | 0.4 | 0.5 | 0.8 | 1.3 | 3.0 |
| $A * F * AL$ | 1.4 | 1.4 | 1.4 | 1.7 | 2.0 | 3.1 | 6.6 |
| $M_0 * M_1 * M_2 * A * F$ | 0.2 | 0.2 | 0.3 | 0.3 | 0.7 | 1.1 | 2.7 |

The adaptive cascade rule incorporates the modality combinations, generated by the RWS rule, as additional sources of further decision fusion. Three such combined modalities are considered by fusing audio, face, and multistream audio-lip modalities in different RWS combinations, specifically (A ⊕ F ⊕ AL), (A ⊕ F), and (F ⊕ AL). When these combined modalities are adaptively cascaded with relatively reliable unimodal streams, that is, audio and face, we observe a further performance gain.

**CIAIR database examples**. Nagoya University's CIAIR has been collecting an in-car speech database since 1999 with a data collection vehicle the center has designed. This vehicle supports synchronous recording of multichannel audio data from 12 microphones, multichannel video data from 3 cameras, and vehicle-related data such as vehicle speed, engine rpm, steering wheel angle, acceleration, and brake pedal pressures. Each channel is sampled at 1.0 kHz. During the data collection stage, each subject converses with three types of dialogue systems: a human navigator, a conversational system, and a Wizard of Oz system. Wizard of Oz systems, commonly used in the speech recognition community, simulate computer-based spoken-language systems with the help of a human "wizard" such that the user believes he is using the real system.

For purposes of the case study, we carried out open-set person identification experiments using audio and video from a 20-person subset of the CIAIR database. We used a camera facing the driver and the audio signal from a headset microphone for each person as video and audio sources, respectively. The faces were hand-cropped to 64 × 40 pixel size and nonsilence audio sections were hand-selected. We performed this manual preprocessing to decrease the effects of preprocessing errors in our experiment. When we performed a

*Table 3. Open-set identification results using a 20-person subset of the Center for Acoustic Information Research database. The ⊕ symbol denotes the RWS rule, and • stands for the product rule.*

| Modality | Equal Error Rate (%) |
|---|---|
| A | 2.44 |
| F | 7.89 |
| A • F | 1.26 |
| A ⊕ F | 1.04 |

fully automatic operation, we observed performance degradation from possible misdetection of the face region and the use of background audio instead of audio from the driver.

Additionally, we used 12 static MFCC features (excluding the energy coefficient, $c_0$) at 100 Hz as audio features. For faces, we used the PCA method to reduce the feature dimension to 20 for each image frame.

For each driver, we used 50 image frames and 50 seconds of nonsilence audio extracted from the database. We divided the data set obtained separately for each driver and modality into 20 equal-length parts. Then, we performed a leave-one-out training procedure so that one data set out of 20 was used for testing and the remaining 19 were used for training. This gave us 20 different test data sets for each person (and the training data was different each time), leading to 400 (20 × 20) trials in total. We then modeled the audio and face data, using GMMs with eight and one mixture components, respectively. For each modality, we used GMMs to construct and train a background model.

Next, we conducted an open-set person identification experiment using a leave-one-out testing, in which 19 people made up the genuine population and one, the imposter class. Repeating the leave-one-out test 20 times, we obtained 7,600 (20 × 20 × 19) genuine, and 400 (20 × 20) impostor, tests.

For each modality, we first normalized log-likelihood scores using the background model and a sigmoid function. Then we combined the audio and face modalities by the so-called product and RWS rules. Table 3 presents our findings on the unimodal and multimodal experiments. The product rule, which assumes independence of modalities and combines the equally weighted modality scores, achieved a 1.26 percent EER rate by improving the unimodal identification rates. On the other hand, the fusion of audio and face modalities with the RWS rule resulted in a 1.04 percent EER rate by outperforming the product rule. These results indicate that combining audio and face modalities for open-set person identification considerably improves the overall performance.

In situations where audio is collected in high-noise regimes, such as in a high-speed vehicle or in windy or rainy weather, we expect that the system will rely more on the face data to obtain a robust open-set person identification performance. Under such conditions, the RWS rule adaptively favors the more reliable modality scores to enhance the fusion process, which our experiments with MVGL data demonstrated (see Table 2).

### Case study 2: Authentication

In our second case study, we used driving signals to verify whether a known driver was in normal driving condition, using a 20-person subset of the CIAIR database. If authentication didn't succeed, it indicated a potential problem such as fatigue or drunkenness. In this scenario, the impostor data for each driver should theoretically be gathered in actual fatigued or drunk-driving conditions. However, the CIAIR database doesn't contain such data; collecting such data is obviously difficult. Accordingly, we assumed that the impostor data for each driver was given by the driving signals of the remaining 19 drivers.

In this study, the brake and acceleration pedal pressure readings, originally taken at 1 kHz, were smoothed and downsampled by a factor of 10. We used those signals and their first derivatives as features for modeling the driving behavior. This resulted in a four-dimensional feature vector at 100 Hz. We gathered 600 seconds of driving signals from each person in the database. We divided the data for each person into 20 equal parts, and used an 8-mixture GMM to model each driver's behavior. We then performed a leave-one-out training procedure to obtain 400 genuine and 7,600 impostor testing samples to compute the authentication performance.

We achieved an authentication EER of 4 percent by using driving signals only, where

$$\text{FAR} = 100 \times \frac{\text{\# of false accepts}}{N_r} \text{ and}$$

$$\text{FRR} = 100 \times \frac{\text{\# of false rejects}}{N_a}$$

and $N_a$ and $N_r$ are the total number of trials for the genuine (400) and imposter (7,600) clients in the testing, respectively. This result was encouraging, in that the driving signals apparently verified the driving behavior, and so could be used to detect fatigued or drunk-driving conditions.

### Discussion and conclusion

Although the performance of each individual modality could be increased in adverse conditions, such as using denoising and echo cancellation for in-vehicle speech signal,[5] the multimodal performance we obtained in our case studies surpassed each unimodal system once the best features or decisions from each modality were fused. In particular, in case study 1, we demonstrated that the required levels of accuracy for biometric person recognition in adverse environments could be achieved by fusing multiple modalities on data from two different databases. Results from the CIAIR data set served as a realistic evaluation of the system performance using low-resolution cameras, whereas results on data from the MVGL database demonstrated the benefit from improving the data acquisition setup employed in a vehicle. Moreover, we observed that as the number of available modalities increases, the benefit of using the adaptive cascade rule compared to the RWS rule became clearer, as our experiments with the MVGL data set illustrated. In the CIAIR case, the RWS rule worked reasonably well for fusing the available audio and face modalities.

Furthermore, multimodal person identification enables a fault-tolerant design in case one of the sensors (for example, a camera or acoustic sensor) fails. We can assign reliability measures to each modality, and to the output of each sensor, so that we can consider only the features for the most reliable modalities and sensors in the fusion strategy. This includes scenarios such as "disregard speech modality" when the background noise can't be suppressed effectively, or "disregard lip features" if the driver isn't looking straight into the camera, and so on.

We also demonstrated, in case study 2, that driving behavior signals can verify the current driving condition of an identified driver in a drive-safe scenario, where active or passive safety enforcement systems could be deployed if the driver's behavior doesn't comply with predetermined normal behavior.

Potential future research directions include detecting the best features for each modality, evaluating optimum fusion strategies, and improving behavioral modeling of drivers. **MM**

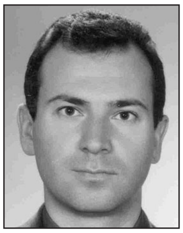
## Acknowledgments

We acknowledge Fumitada Itakura of Meijo University, Japan, and Kazuya Takeda and Nobuo Kawaguchi of Nagoya University, Japan, for providing the CIAIR database used in our case studies.

## References

1. J. Campbell, "Speaker Recognition: A Tutorial," *Proc. IEEE*, vol. 85, no. 9, 1997, IEEE Press, pp. 1437-1462.
2. U.V. Chaudhari, J. Navratil, and S.H. Maes, "Multi-Grained Modeling with Pattern Specific Maximum Likelihood Transformations for Text-Independent Speaker Recognition," *IEEE Trans. Speech and Audio Processing*, vol. 11., no. 1, 2003, pp. 61-69.
3. M. Schmidt and H. Gish, "Speaker Identification via Support Vector Machines," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 96), IEEE Press, 1996, pp. 105-108.
4. S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, 1979, pp. 113-120.
5. H. Abut, J.H.L. Hansen, and K. Takeda, eds., *DSP in Vehicular and Mobile Systems*, Springer, 2005.
6. I. Cohen and B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, 2002, pp. 12-15.
7. *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms*, ETSI ES 202 050 V1.1.1, European Telecomm. Standards Inst., Oct. 2002.
8. C. Barras and J.-L. Gauvain, "Feature and Score Normalization for Speaker Verification of Cellular Data," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP 03), IEEE Press 2003, vol. 2, pp. 49-52.
9. K. Kirchhoff, "Combining Articulatory and Acoustic Information for Speech Recognition in Noisy and Reverberant Environments," *Proc. 5th Int'l Conf. Spoken Language Processing* (ICSLP 98), Int'l Speech Communication Assoc., 1998, pp. 891-894.
10. P.J. Moreno, *Speech Recognition in Noisy Environments*, doctoral dissertation, Electrical and Computer Engineering Dept., Carnegie Mellon Univ., 1996.
11. W. Zhao and R. Chellappa, Eds., Face Processing: Advanced Modeling and Methods, Academic Press, 2005.
12. B.S. Manjunath, R. Chellappa, and C.V.D. Malsburg, "A Feature Based Approach to Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (CVPR 92), IEEE Press, 1992, pp. 373-378.

13. K. Okada et al., "The Bochum/USC Face Recognition System and How it Fared," *FERET Phase III Test,* in *Face Recognition: From Theory to Applications*, H. Wechsler et al., eds., Springer-Verlag, 1998, pp. 186-205.
14. W. Zhao, "Subspace Methods in Object/Face Recognition," *Proc. Int'l Joint Conf. Neural Networks*, vol. 5, IEEE Press, 1999, pp. 3260-3264.
15. P.J. Phillips et al., "The FERET Database and Evaluation Procedure for Face Recognition Algorithms," *Image and Vision Computing*, vol. 16, issue 5, 1998, pp. 295-306.
16. L. Wiskott, J.M. Fellous, and C. von der Malsburg, "Face Recognition by Elastic Bunch Graph Matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, 1997, pp. 775-779.
17. M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience*, vol. 3, no. 1, 1991, pp. 586-591.
18. Y. Adini, Y. Moses, and S. Ullman, "Face Recognition: The Problem of Compensating for Changes in Illumination Direction," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, 1997, pp. 721-732.
19. U.V. Chaudhari et al., "Information Fusion and Decision Cascading for Audiovisual Speaker Recognition Based on Time-Varying Stream Reliability Prediction," *Proc. Int'l Conf. Multimedia & Expo* (ICME 03), vol. 3, IEEE Press, 2003, pp. 9-12.
20. H.E. Cetingul et al., "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading," to appear in *IEEE Trans. Image Processing*.
21. T. Wark and S. Sridharan, "Adaptive Fusion of Speech and Lip Information for Robust Speaker Identification," *Digital Signal Processing*, vol. 11, no. 3, 2001, pp. 169-186.
22. N. Eveno, A. Caplier, and P.Y. Coulon, "A Parametric Model for Realistic Lip Segmentation," *Proc. 7th Int'l Conf. Control, Automation, Robotics and Vision* (ICARCV 02), vol. 3, IEEE Press, 2002, pp. 1426-1431.
23. E. Erzin, Y. Yemez, and A.M. Tekalp, "Multimodal Speaker Identification Using an Adaptive Classifier Cascade Based on Modality Reliability," *IEEE Trans. MultiMedia*, vol. 7, no. 5, 2005, pp. 840-852.
24. K. Igarashi et al., "Biometric Identification Using Driving Behavior*," Proc. IEEE Int'l Conf. Multimedia & Expo* (ICME 04)*,* vol. 1, IEEE Press, 2004, pp. 65-68.
25. N. Kawaguchi et al., "Construction and Analysis of the Multi-layered In-Car Spoken Dialogue Corpus," *DSP in Vehicular and Mobile Systems*, H. Abut, J.H.L. Hansen, and K. Takeda, eds., Springer, 2005,pp. 1-18.
26. H. Abut, J.H.L. Hansen, and K. Takeda, eds., *DSP in Vehicular and Mobile Systems*, Springer, 2005, pp. 257-274.
27. R. Snelick et al., "Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, 2005, pp. 450-455.
28. J. Kittler et al., "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, 1998, pp. 226-239.
29. A.K. Jain, K. Nandakumar, and A. Ross, "Score Normalization in Multimodal Biometric Systems," *Pattern Recognition*, vol. 38, issue 12, Elsevier, Dec. 2005, pp. 2270-2285.
30. A. Baykut and A. Erçil, "Towards Automated Classifier Combination for Pattern Recognition," *Multiple Classifier Systems,* LNCS 2709, T. Wideatt and Fabio Roli, eds., Springer Verlag, 2003, pp. 94-105.



**Engin Erzin** is an assistant professor in the Electrical and Electronics Engineering and the Computer Engineering Departments of Koç University, Istanbul, Turkey. His research interests include speech signal processing, pattern recognition, and adaptive signal processing. Erzin received a PhD, MS, and BS from the Bilkent University, Ankara, Turkey, all in electrical engineering.

**Yücel Yemez** is an assistant professor in the Computer Engineering Department at Koç University. His current research is focused on various fields of computer vision and 3D computer graphics. Yemez received a BS from Middle East Technical University, Ankara, Turkey, and an MS and PhD from Bogaziçi University, Istanbul, Turkey, all in electrical engineering.

**A. Murat Tekalp** is a professor at Koç University. His research interests are in digital image and video processing. Tekalp received an MS and a PhD in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute. He received a Fulbright Senior Scholarship in 1999 and the Tubitak Science Award in 2004. Tekalp, editor in chief of the EURASIP journal *Signal Processing: Image Communication*, authored *Digital Video Processing* (Prentice Hall, 1995). He holds seven US patents and is a Fellow of the IEEE.

**Aytül Erçil** is a professor at the Faculty of Engineering at Sabanci University, Istanbul, Turkey and director of the Computer Vision and Pattern Recognition Lab. Her research interests include invariant object recognition, shape modeling, texture analysis, pattern recognition, biometrics, and multimodal classification. She is a governing board member of the International Association of Pattern Recognition, president of the Turkish Pattern Recognition and Image Processing Society, and a member of the IEEE Computer Society. Erçil received a BS in electrical engineering and a BS in mathematics from Bogaziçi University and an MS and a PhD in applied mathematics from Brown University.

**Hakan Erdogan** is an assistant professor at Sabanci University in Istanbul, Turkey. His research interests include statistical methods for multimedia information extraction and pattern classification with an emphasis on audio processing. Erdogan received a BS in electrical engineering and mathematics from Middle East Technical University, Ankara, and an MS and a PhD from the University of Michigan, Ann Arbor.

**Hüseyin Abut** is a faculty early retirement program (FERP) professor of Electrical and Computer Engineering at San Diego State University. He is the editor of *Vector Quantization* (IEEE Press, 1990) and has recently co-edited *DSP for In-Vehicle and Mobile Systems* (Springer Verlag, 2005). The author/co-author of more than 100 publications in journals, books, and book chapters in the field of speech processing and signal compression, Abut holds two patents on speaker verification systems for consumer applications. Abut received an MS and a PhD from North Carolina State University, Raleigh.

Readers may contact Engin Erzin at Koç University, Sariyer, Istanbul, 34450, Turkey; eerzin@ku.edu.tr.

**For further information on this or any other computing topic, please visit our Digital Library at http://computer.org/publications/dlib.**