

An Overview of Text-Independent Speaker Recognition: from Features to Supervectors

Tomi Kinnunen^{*a}, Haizhou Li^b

^aDepartment of Computer Science and Statistics, Speech and Image Processing Unit
University of Joensuu, P.O.Box 111, 80101 Joensuu, FINLAND
WWW homepage: <http://cs.joensuu.fi/sipu/>

^bDepartment of Human Language Technology, Institute for Infocomm Research (I²R)
1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632
WWW homepage: <http://hlt.i2r.a-star.edu.sg/>

Abstract

This paper gives an overview of automatic speaker recognition technology, with an emphasis on text-independent recognition. Speaker recognition has been studied actively for several decades. We give an overview of both the classical and the state-of-the-art methods. We start with the fundamentals of automatic speaker recognition, concerning feature extraction and speaker modeling. We elaborate advanced computational techniques to address robustness and session variability. The recent progress from vectors towards supervectors opens up a new area of exploration and represents a technology trend. We also provide an overview of this recent development and discuss the evaluation methodology of speaker recognition systems. We conclude the paper with discussion on future directions.

Key words: Speaker recognition, text-independence, feature extraction, statistical models, discriminative models, supervectors, intersession variability compensation

1. Introduction

Speaker recognition refers to recognizing persons from their voice. No two individuals sound identical because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. In addition to these *physical* differences, each speaker has his or her characteristic *manner of speaking*, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, choice of vocabulary and so on. State-of-the-art speaker recognition systems use a number of these features in parallel, attempting to cover these different aspects and employing them in a complementary way to achieve more accurate recognition.

An important application of speaker recognition technology is *forensics*. Much of information is exchanged between two parties in telephone conversations, including between criminals, and in recent years there has been increasing interest to integrate automatic speaker

recognition to supplement auditory and semi-automatic analysis methods [3, 76, 174, 185, 223].

Not only forensic analysts but also ordinary persons will benefit from speaker recognition technology. It has been predicted that telephone-based services with integrated speech recognition, speaker recognition, and language recognition will supplement or even replace human-operated telephone services in the future. An example is automatic password reset over the telephone¹. The advantages of such automatic services are clear - much higher capacity compared to human-operated services with hundreds or thousands of phone calls being processed simultaneously. In fact, the focus of speaker recognition research over the years has been tending towards such telephony-based applications.

In addition to telephony speech data, there is a continually increasing supply of other *spoken documents* such as TV broadcasts, teleconference meetings, and video clips from vacations. Extracting *metadata* like topic of discussion or participant names and genders from

*Corresponding author

Email addresses: tkinnu@cs.joensuu.fi (Tomi Kinnunen),
hli@i2r.a-star.edu.sg (Haizhou Li)

Preprint submitted to *Speech Communication*

¹See e.g. http://www.pcworld.com/article/106142/visa_gets_behind_voice_recognition.html

these documents would enable automated information searching and indexing. *Speaker diarization* [226], also known as “who spoke when”, attempts to extract speaking turns of the different participants from a spoken document, and is an extension of the “classical” speaker recognition techniques applied to recordings with multiple speakers.

In forensics and speaker diarization, the speakers can be considered non-cooperative as they do not specifically wish to be recognized. On the other hand, in telephone-based services and access control, the users are considered cooperative. Speaker recognition systems, on the other hand, can be divided into *text-dependent* and *text-independent* ones. In text-dependent systems [91], suited for cooperative users, the recognition phrases are fixed, or known beforehand. For instance, the user can be prompted to read a randomly selected sequence of numbers as described in [101]. In text-independent systems, there are no constraints on the words which the speakers are allowed to use. Thus, the reference (what are spoken in training) and the test (what are uttered in actual use) utterances may have completely different content, and the recognition system must take this phonetic mismatch into account. Text-independent recognition is the much more challenging of the two tasks.

In general, phonetic variability represents one adverse factor to accuracy in text-independent speaker recognition. Changes in the acoustic environment and technical factors (transducer, channel), as well as “within-speaker” variation of the speaker him/herself (state of health, mood, aging) represent other undesirable factors. In general, any variation between two recordings of the same speaker is known as *session variability* [111, 231]. Session variability is often described as mismatched training and test conditions, and it remains to be the most challenging problem in speaker recognition.

This paper represents an overview of speaker recognition technologies, including a few representative techniques from 1980s until today. In addition, we give emphasis to the recent techniques that have presented a paradigm shift from the traditional vector-based speaker models to so-called supervector models. This paper serves as a quick overview of the research questions and their solutions for someone who would like to start research in speaker recognition. The paper may also be useful for speech scientists to have a glance at the current trends in the field. We assume familiarity with basics of digital signal processing and pattern recognition.

We recognize that a thorough review of the field with more than 40 years of active research is challenging.

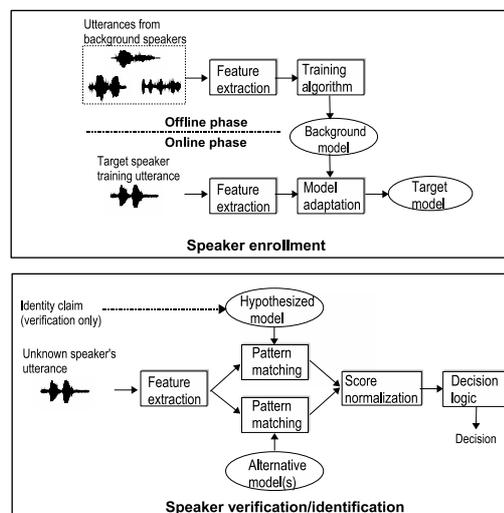


Figure 1: Components of a typical automatic speaker recognition system. In the enrollment mode, a speaker model is created with the aid of previously created background model; in recognition mode, both the hypothesized model and the background model are matched and background score is used in normalizing the raw score.

For the interested reader we therefore point to other useful surveys. Campbell’s tutorial [33] includes in-depth discussions of feature selection and stochastic modeling. A more recent overview, with useful discussions of normalization methods and speaker recognition applications, can be found in [22]. Recent collection of book chapters on various aspects of speaker classification can also be found in [167, 168]. For an overview of text-dependent recognition, refer to [91].

Section 2 provides fundamentals of speaker recognition. Sections 3 and 4 then elaborate feature extraction and speaker modeling principles. Section 5 describes robust methods to cope with real-life noisy and session mismatched conditions, with the focus on feature and score normalization. Section 6 is then devoted to the current supervector classifiers and their session compensation. In Section 7 we discuss the evaluation of speaker recognition performance and give pointers to software packages as well. Finally, possible future horizons of the field are outlined in Section 8, followed by conclusions in Section 9.

2. Fundamentals

Figure 1 shows the components of an automatic speaker recognition system. The upper is the enrollment process, while the lower panel illustrates the recognition process. The *feature extraction* module first transforms

the raw signal into feature vectors in which speaker-specific properties are emphasized and statistical redundancies suppressed. In the enrollment mode, a *speaker model* is trained using the feature vectors of the target speaker. In the recognition mode, the feature vectors extracted from the unknown person’s utterance are compared against the model(s) in the system database to give a similarity score. The decision module uses this similarity score to make the final decision.

Virtually all state-of-the-art speaker recognition systems use a set of *background speakers* or *cohort speakers* in one form or another to enhance the robustness and computational efficiency of the recognizer. In the enrollment phase, background speakers are used as the negative examples in the training of a discriminative model [36], or in training a *universal background model* from which the target speaker models are adapted [197]. In the recognition phase, background speakers are used in the normalization of the speaker match score [71, 101, 139, 193, 197, 206].

2.1. Selection of Features

Speech signal includes many features of which not all are important for speaker discrimination. An ideal feature would [201, 234]

- have large between-speaker variability and small within-speaker variability
- be robust against noise and distortion
- occur frequently and naturally in speech
- be easy to measure from speech signal
- be difficult to impersonate/mimic
- not be affected by the speaker’s health or long-term variations in voice.

The number of features should be also relatively low. Traditional statistical models such as the Gaussian mixture model [197, 198] cannot handle high-dimensional data. The number of required training samples for reliable density estimation grows exponentially with the number of features. This problem is known as *the curse of dimensionality* [104]. The computational savings are also obvious with low-dimensional features.

There are different ways to categorize the features (Fig. 2). From the viewpoint of their physical interpretation, we can divide them into (1) *short-term spectral features*, (2) *voice source features*, (3) *spectro-temporal features*, (4) *prosodic features* and (5) *high-level features*. Short-term spectral features, as the name suggests, are computed from short frames of about 20-30

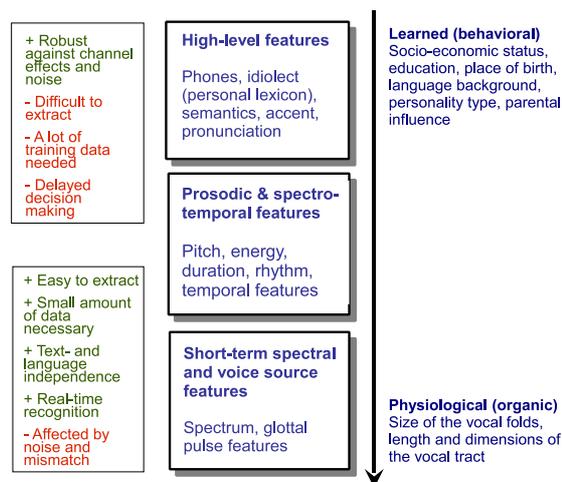


Figure 2: A summary of features from viewpoint of their physical interpretation. The choice of features has to be based on their discrimination, robustness, and practicality. Short-term spectral features are the simplest, yet most discriminative; prosodics and high-level features have received much attention at high computational cost.

milliseconds in duration. They are usually descriptors of the short-term *spectral envelope* which is an acoustic correlate of *timbre*, i.e. the “color” of sound, as well as the resonance properties of the supralaryngeal vocal tract. The voice source features, in turn, characterize the voice source (glottal flow). Prosodic and spectro-temporal features span over tens or hundreds of milliseconds, including intonation and rhythm, for instance. Finally, high-level features attempt to capture conversation-level characteristics of speakers, such as characteristic use of words (“uh-huh”, “you know”, “oh yeah”, etc.) [57].

Which features one should use? It depends on the intended application, computing resources, amount of speech data available (for both development purposes and in run-time) and whether the speakers are cooperative or not. For someone who would like to start research in speaker recognition, we recommend to begin with the short-term spectral features since they are easy to compute and yield good performance [195]. Prosodic and high-level features are believed to be more robust, but less discriminative and easier to impersonate; for instance, it is relatively well known that professional impersonators tend to modify the overall pitch contour towards the imitated speaker [10, 126]. High-level features also require considerably more complex front-end, such as automatic speech recognizer. To conclude, there does not yet exist globally “best” feature but the choice is a trade-off between speaker discrimination, robust-

ness, and practicality.

2.2. Speaker Modeling

By using feature vectors extracted from a given speaker’s training utterance(s), a speaker model is trained and stored into the system database. In text-dependent mode, the model is utterance-specific and it includes the temporal dependencies between the feature vectors. Text-dependent speaker verification and speech recognition do share similarities in their pattern matching processes, and these can also be combined [18, 93].

In text-independent mode we often model the feature distribution, i.e. the shape of the “feature cloud” rather than the temporal dependencies. Note that, in text-dependent recognition, we can temporally align the test and training utterances because they contain (are assumed to contain) the same phoneme sequences. However, in text-independent recognition, since there are little or absolutely no correspondence between the frames in the test and reference utterances, alignment at the frame level is not possible. Therefore, *segmentation* of the signal into phones or broad phonetic classes can be used as a pre-processing step, or alternatively, the speaker models can be structured phonetically. Such approaches have been proposed in [61, 81, 79, 92, 180, 107]. It is also possible to use data-driven units instead of the strictly linguistic phonemes as segmentation units [80].

Classical speaker models can be divided into *template models* and *stochastic models* [33], also known as *nonparametric* and *parametric* models, respectively. In template models, training and test feature vectors are directly compared with each other with the assumption that either one is an imperfect replica of the other. The amount of distortion between them represents their degree of similarity. *Vector quantization* (VQ) [213] and *dynamic time warping* (DTW) [70] are representative examples of template models for text-independent and text-dependent recognition, respectively.

In stochastic models, each speaker is modeled as a probabilistic source with an unknown but fixed probability density function. The training phase is to estimate the parameters of the probability density function from a training sample. Matching is usually done by evaluating the likelihood of the test utterance with respect to the model. The *Gaussian mixture model* (GMM) [198, 197] and the *hidden Markov model* (HMM) [19, 171] are the most popular models for text-independent and text-dependent recognition, respectively.

According to the training paradigm, models can also be classified into *generative* and *discriminative* models. The generative models such as GMM and VQ estimate

the feature distribution *within* each speaker. The discriminative models such as *artificial neural networks* (ANNs) [62, 94, 239] and *support vector machines* (SVMs) [36], in contrast, model the *boundary* between speakers. For more discussions, refer to [190].

In summary, a speaker is characterized by a speaker model such as VQ, GMM or SVM. At run-time, a unknown voice is first represented by a collection of feature vectors or a supervector - a concatenation of multiple vectors, then evaluated against the target speaker models.

3. Feature Extraction

3.1. Short-Term Spectral Features

The speech signal continuously changes due to articulatory movements, and therefore, the signal must be broken down in short *frames* of about 20-30 milliseconds in duration. Within this interval, the signal is assumed to remain stationary and a spectral feature vector is extracted from each frame.

Usually the frame is *pre-emphasized* and multiplied by a smooth *window function* prior to further steps. Pre-emphasis boosts the higher frequencies whose intensity would be otherwise very low due to downward sloping spectrum caused by glottal voice source [82, p. 168]. The window function (usually Hamming), on the other hand, is needed because of the finite-length effects of the discrete Fourier transform (DFT); for details, refer to [83, 56, 177]. In practice, choice of the window function is not critical. Although the frame length is usually fixed, *pitch-synchronous analysis* has also been studied [172, 247, 75]. The experiments in [172, 247] indicate that recognition accuracy reduces with this technique, whereas [75] obtained some improvement in noisy conditions. Pitch-dependent speaker models have also been studied [9, 60].

The well-known *fast Fourier transform* (FFT), a fast implementation of DFT, decomposes a signal into its frequency components [177]. Alternatives to FFT-based signal decomposition such as non-harmonic bases, aperiodic functions and data-driven bases derived from independent component analysis (ICA) have been studied in literature [77, 103, 105]. The DFT, however, remains to be used in practice due to its simplicity and efficiency. Usually only the magnitude spectrum is retained, based on the belief that phase has little perceptual importance. However, [179] provides opposing evidence while [96] described a technique which utilizes phase information.

The global shape of the DFT magnitude spectrum (Fig. 3), known as *spectral envelope*, contains information about the resonance properties of the vocal tract and

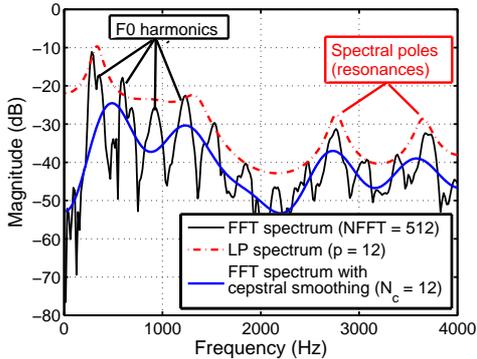


Figure 3: Extraction of spectral envelope using cepstral analysis and linear prediction (LP). Spectrum of NFFT = 512 points can be effectively reduced to only $N_c = 12$ cepstral coefficients or $p = 12$ LP coefficients. Both the cepstral and LP features are useful and complementary to each other when used in speaker recognition.

has been found out to be the most informative part of the spectrum in speaker recognition. A simple model of spectral envelope uses a set of band-pass filters to do energy integration over neighboring frequency bands. Motivated by psycho-acoustic studies, the lower frequency range is usually represented with higher resolution by allocating more filters with narrow bandwidths [82].

Although the subband energy values have been used directly as features [20, 21, 49, 205], usually the dimensionality is further reduced using other transformations. The so-called *mel-frequency cepstral coefficients* (MFCCs) [50] are popular features in speech and audio processing. MFCCs were introduced in early 1980s for speech recognition and then adopted in speaker recognition. Even though various alternative features, such as spectral subband centroids (SSCs) [125, 221] have been studied, the MFCCs seem to be difficult to beat in practice.

MFCCs are computed with the aid of a psychoacoustically motivated filterbank, followed by logarithmic compression and discrete cosine transform (DCT). Denoting the outputs of an M -channel filterbank as $Y(m), m = 1, \dots, M$, the MFCCs are obtained as follows:

$$c_n = \sum_{m=1}^M \left[\log Y(m) \right] \cos \left[\frac{\pi n}{M} \left(m - \frac{1}{2} \right) \right]. \quad (1)$$

Here n is the index of the cepstral coefficient. The final MFCC vector is obtained by retaining about 12-15 lowest DCT coefficients. More details of MFCCs can be found in [56, 102]. Alternative features that emphasize speaker-specific information have been studied in [43, 165, 113, 178]. For study of speaker-discriminative

information in spectrum, refer to [144]. Finally, some new trends in feature extraction can be found in [6].

Linear prediction (LP) [152, 155] is an alternative spectrum estimation method to DFT that has good intuitive interpretation both in time domain (adjacent samples are correlated) and frequency domain (all-pole spectrum corresponding to the resonance structure). In time domain, LP predictor equation is defined as,

$$\tilde{s}[n] = \sum_{k=1}^p a_k s[n-k]. \quad (2)$$

Here $s[n]$ is the observed signal, a_k are the *predictor coefficients* and $\tilde{s}[n]$ is the predicted signal. The prediction error signal, or *residual*, is defined as $e[n] = s[n] - \tilde{s}[n]$, and illustrated in the middle panel of Fig. 4. The coefficients a_k are usually determined by minimizing the residual energy using the so-called *Levinson-Durbin* algorithm [82, 102, 189]. The spectral model is defined as,

$$H(z) = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}}, \quad (3)$$

and it consists of spectral peaks or *poles* only (dash-dotted line in Fig. 3).

The predictor coefficients $\{a_k\}$ themselves are rarely used as features but they are transformed into robust and less correlated features such as *linear predictive cepstral coefficients* (LPCCs) [102], *line spectral frequencies* (LSFs) [102], and *perceptual linear prediction* (PLP) coefficients [97]. Other, somewhat less successful features, include *partial correlation coefficients* (PARCORs), *log area ratios* (LARs) and *formant frequencies and bandwidths* [189].

Given all the alternative spectral features, which one should be used for speaker recognition and how should the parameters (e.g. the number of coefficients) be selected? Some comparisons can be found in [12, 114, 118, 198], and it has been observed that in general channel compensation methods are much more important than the choice of the base feature set [198]. Different spectral features, however, are complementary and can be combined to enhance accuracy [28, 36, 118]. In summary, for practical use we recommend any of the following features: MFCC, LPCC, LSF, PLP.

3.2. Voice Source Features

Voice source features characterize the glottal excitation signal of voiced sounds such as glottal pulse shape and fundamental frequency, and it is reasonable to assume that they carry speaker-specific information. Fundamental frequency, the rate of vocal fold vibration, is

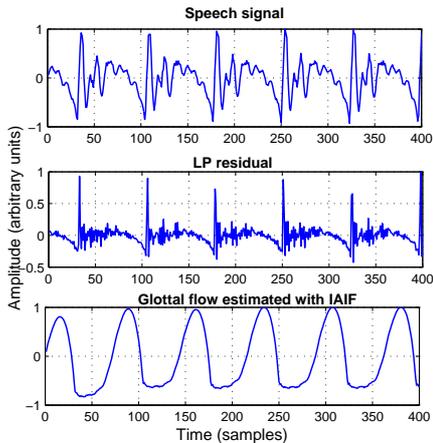


Figure 4: Glottal feature extraction [116]. Speech frame (top), linear prediction (LP) residual (middle), and glottal flow estimated via inverse filtering (bottom). ©2009 IEEE. Reprinted by permission.

popular and will be discussed in Section 3.4. Other parameters are related to the shape of the glottal pulse, such as the degree of vocal fold opening and the duration of the closing phase. These contribute to voice quality which can be described for example, as modal, breathy, creaky or pressed [59].

The glottal features are not directly measurable due to the vocal tract filtering effect. By assuming that the glottal source and the vocal tract are independent of each other, vocal tract parameters can be first estimated using, for instance, the linear prediction model, followed by *inverse filtering* of the original waveform to obtain an estimate of the source signal [116, 170, 186, 188, 220, 242]. An alternative method uses *closed-phase covariance analysis* during the portions when the vocal folds are closed [78, 186, 208]. This leads to improved estimate of the vocal tract but accurate detection of closed phase is required which is difficult in noisy conditions. As an example, Fig. 4 shows a speech signal together with its LP residual and glottal flow estimated with a simple inverse filtering method [4].

Features of the inverse filtered signal can be extracted, for instance, by using an auto-associative neural network [188]. Other approaches have used parametric glottal flow model parameters [186], wavelet analysis [242], residual phase [170], cepstral coefficients [78, 47, 116] and higher-order statistics [47] to mention a few.

Based on the literature, voice source features are not as discriminative as vocal tract features but fusing these two complementary features can improve accu-

racy [170, 242]. Experiments of [42, 188] also suggest that the amount of training and testing data for the voice source features can be significantly less compared to the amount of data needed for the vocal tract features (10 seconds vs 40 seconds in [188]). A possible explanation for this is that vocal tract features depend on the phonetic content and thus require sufficient phonetic coverage for both the training and test utterances. Voice source features, in turn, depend much less on phonetic factors.

3.3. Spectro-Temporal Features

It is reasonable to assume that the spectro-temporal signal details such as formant transitions and energy modulations contain useful speaker-specific information. A common way to incorporate some temporal information to features is through 1st and 2nd order time derivative estimates, known as *delta* (Δ) and *double-delta* (Δ^2) coefficients, respectively [70, 102, 214]. They are computed as the time differences between the adjacent vectors feature coefficients and usually appended with the base coefficients on the frame level (e.g. 13 MFCCs with Δ and Δ^2 coefficients, implying 39 features per frame). An alternative, potentially more robust, method fits a regression line [189] or an orthogonal polynomial [70] to the temporal trajectories, although in practice simple differentiation seems to yield equal or better performance [114]. *Time-frequency principal components* [148] and *data-driven temporal filters* [153] have also been studied.

In [115, 123], we proposed to use *modulation frequency* [13, 98] as a feature for speaker recognition as illustrated in Fig. 5. Modulation frequency represents the frequency content of the subband amplitude envelopes and it potentially contains information about speaking rate and other stylistic attributes. Modulation frequencies relevant for speech intelligibility are approximately in the range 1-20 Hz [13, 98]. In [115], the best recognition result was obtained by using a temporal window of 300 milliseconds and by including modulation frequencies in the range 0-20 Hz. The dimensionality of the modulation frequency vector depends on the number of FFT points of the spectrogram and the number of frames spanning the FFT computation in the temporal direction. For the best parameter combination, the dimension of the feature vector was 3200 [115].

In [122] and [123] we studied reduced-dimensional spectro-temporal features. The *temporal discrete cosine transform* (TDCT) method, proposed in [122] and illustrated in Fig. 6, applies DCT on the temporal trajectories of the cepstral vectors rather than on the spectrogram magnitudes. Using DCT rather than DFT mag-

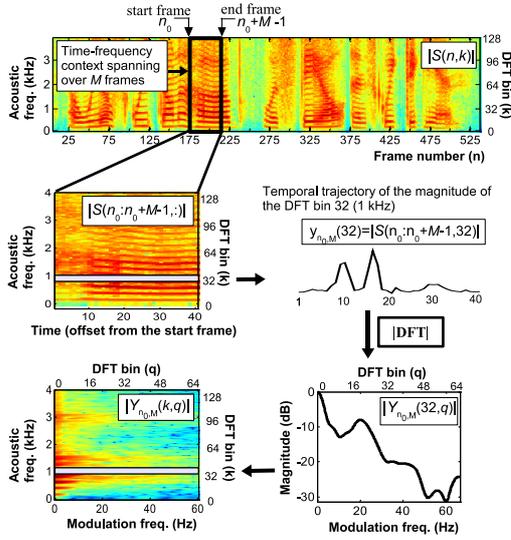


Figure 5: Extracting modulation spectrogram features [123]. A time-frequency context, including M short-term spectra over the interval $[n_0 \dots n_0 + M - 1]$, is first extracted. The DFT magnitude spectra of all feature trajectories are then computed and stacked as a feature vector with high dimensionality (here $129 \times 65 = 8385$ elements).

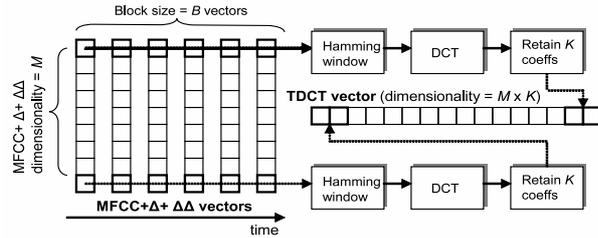


Figure 6: Temporal discrete cosine transform (TDCT) [122]. Short-term MFCC features with their delta features are taken as input; their low-frequency modulation characteristics are then represented by computing discrete cosine transform (DCT) over a context of B frames. The lowest DCT coefficients are retained as features.

nitude here has an advantage that it retains the relative phases of the feature coefficient trajectories, and hence, it can preserve both phonetic and speaker-specific information. This, however, requires more research. In [123], DCT was used in a different role: reducing the dimensionality of the modulation magnitude spectra. The best results in [115, 123] were obtained by using a time context of 300-330 milliseconds, which is significantly longer compared with the typical time contexts of the delta features.

Even though we obtained some improvement over the cepstral systems by fusing the match scores of the cepstral and temporal features [115, 122], the gain was rather modest and more research is required before these

features can be recommended for practical applications. One problem could be that we have applied speaker modeling techniques that are designed for short-term features. Due to larger temporal context, the number of training vectors is usually less compared with short-term features. Furthermore, as the short-term and longer-term features have different frame rates, they cannot be easily combined at the frame level. Perhaps a completely different modeling and fusion technique is required for these features.

An alternative to amplitude-based methods considers *frequency modulations* (FM) instead [222]. In FM-based methods, the input signal is first divided into sub-band signals using a bank of bandpass filters. The dominant frequency components (such as the frequency centroids) in the subbands then capture formant-like features. As an example, the procedure described in [222] uses 2nd order all-pole analysis to detect the dominant frequency. The FM features are then obtained by subtracting the center frequency of the subband from the pole frequency, yielding a measure of deviation from the “default” frequency of the bandpass signal. This feature was applied to speaker recognition in [223], showing promise when fused with conventional MFCCs.

3.4. Prosodic Features

Prosody refers to non-segmental aspects of speech, including for instance syllable stress, intonation patterns, speaking rate and rhythm. One important aspect of prosody is that, unlike the traditional short-term spectral features, it spans over long segments like syllables, words, and utterances and reflects differences in speaking style, language background, sentence type, and emotions to mention a few. A challenge in text-independent speaker recognition is modeling the different levels of prosodic information (instantaneous, long-term) to capture speaker differences; at the same time, the features should be free of effects that the speaker can voluntarily control.

The most important prosodic parameter is the fundamental frequency (or F0). Combining F0-related features with spectral features has been shown to be effective, especially in noisy conditions. Other prosodic features for speaker recognition have included duration (e.g. pause statistics, phone duration), speaking rate, and energy distribution/modulations among others [2, 16, 195, 204]. Interested reader may refer to [204] for further details. In that study, it was found out, among a number of other observations, that F0-related features yielded the best accuracy, followed by energy and duration features in this order. Since F0 is the predominant prosodic feature, we will now discuss it in more detail.

Reliable F0 determination itself is a challenging task. For instance, in telephone quality speech, F0 is often outside of the narrowband telephone network passband (0.3–3.4 kHz) and the algorithms can only rely on the information in the upper harmonics for F0 detection. For a detailed discussion of classical F0 estimation approaches, refer to [100]. More recent comparison of F0 trackers can be found in [48]. For practical use, we recommend the YIN method [51] and the autocorrelation method as implemented in Praat software [26].

For speaker recognition, F0 conveys both physiological and learned characteristics. For instance, the mean value of F0 can be considered as an acoustic correlate of the larynx size [201], whereas the temporal variations of pitch are related to the manner of speaking. In text-dependent recognition, temporal alignment of pitch contours have been used [11]. In text-independent studies, long-term F0 statistics - especially the mean value - have been extensively studied [39, 117, 158, 176, 209, 210]. The mean value combined with other statistics such as variance and kurtosis can be used as speaker model [16, 39, 117], even though histograms [117], latent semantic analysis [46] and support vector machines [204] perform better. It has also been found through a number of experiments that $\log(F0)$ is a better feature than F0 itself [117, 210].

F0 is a one-dimensional feature, therefore mathematically, not expected to be very discriminative. Multi-dimensional pitch- and voicing-related features can be extracted from the auto-correlation function without actual F0 extraction as done in [131, 146, 233] for example. Another way to improve accuracy is modeling both the local and long-term temporal variations of F0.

Capturing local F0 dynamics can be achieved by appending the delta features with the instantaneous F0 value. For longer-term modeling, F0 contour can be segmented and presented by a set of parameters associated with each segment [1, 2, 160, 204, 209]. The segments may be syllables obtained using automatic speech recognition (ASR) system [204]. An alternative, ASR-free approach, is to divide the utterance into syllable-like units using, for instance, vowel onsets [161] or F0/energy inflection points [1, 55] as the segment boundaries.

For parameterization of the segments, prosodic feature statistics and their local temporal slopes (tilt) within each segment are often used. In [2, 209], each voiced segment was parameterized by a piece-wise linear model whose parameters formed the features. In [204], the authors used N -gram counts of discretized feature values as features to an SVM classifier with promising results. In [55], prosodic features were ex-

tracted using polynomial basis functions.

3.5. High-Level Features

Speakers differ not only in their voice timbre and accent/pronunciation, but also in their lexicon - the kind of words the speakers tend to use in their conversations. The work on such “high-level” conversational features was initiated in [57] where a speaker’s characteristic vocabulary, the so-called *idiolect*, was used to characterize speakers. The idea in “high-level” modeling is to convert each utterance into a sequence of *tokens* where the co-occurrence patterns of tokens characterize speaker differences. The information being modeled is hence in categorical (discrete) rather than in numeric (continuous) form.

The tokens considered have included words [57], phones [8, 35], prosodic gestures (rising/falling pitch/energy) [2, 46, 204], and even articulatory tokens (manner and place of articulation) [137]. The top-1 scoring Gaussian mixture component indices have also been used as tokens [147, 225, 235].

Sometimes several parallel tokenizers are utilized [35, 106, 147]. This is partly motivated by the success of parallel phone recognizers in state-of-the-art spoken language recognition [248, 145]. This direction is driven by the hope that different tokenizers (e.g. phone recognizers trained on different languages or with different phone models) would capture complementary aspects of the utterance. As an example, in [147] a set of parallel GMM tokenizers [225, 235] were used. Each tokenizer was trained from a different group of speakers obtained by clustering.

The baseline classifier for token features is based on N -gram modeling. Let us denote the token sequence of the utterance by $\{\alpha_1, \alpha_2, \dots, \alpha_T\}$, where $\alpha_t \in V$ and V is a finite vocabulary. An N -gram model is constructed by estimating the joint probability of N consecutive tokens. For instance, $N = 2$ gives the *bigram* model where the probabilities of token pairs (α_t, α_{t+1}) are estimated. A *trigram* model consists of triplets $(\alpha_t, \alpha_{t+1}, \alpha_{t+2})$, and so forth. As an example, the bigrams of the token sequence `hello_world` are (h,e), (e,l), (l,l), (l,o), (o,_), (_,w), (w,o), (o,r), (r,l) and (l,d).

The probability of each N -gram is estimated in the same way as N -gram in statistical language models in automatic speech recognition [173]. It is the *maximum likelihood* (ML) or *maximum a posteriori* (MAP) estimate of the N -gram in the training corpus [137]. The N -gram statistics have been used in vector space [35, 147] and with entropy measures [7, 137] to assess similarity between speakers.

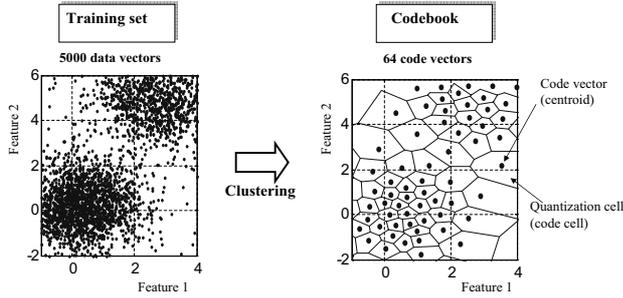


Figure 7: Codebook construction for vector quantization using the K -means algorithm. The original training set consisting of 5000 vectors is reduced to a set of $K = 64$ code vectors (centroids).

4. Speaker Modeling: Classical Approaches

This section describes some of the popular models in text-independent speaker recognition. The models presented here have co-evolved with the short-term spectral features such as MFCCs in the literature.

4.1. Vector Quantization

Vector quantization (VQ) model [32, 88, 90, 109, 120, 213, 214], also known as *centroid model*, is one of the simplest text-independent speaker models. It was introduced to speaker recognition in the 1980s [32, 213] and its roots are originally in data compression [73]. Even though VQ is often used for computational speed-up techniques [142, 120, 199] and lightweight practical implementations [202], it also provides competitive accuracy when combined with background model adaptation [88, 124]. We will return to adaptation methods in Subsection 4.2.

In the following, we denote the test utterance feature vectors by $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ and the reference vectors by $\mathcal{R} = \{\mathbf{r}_1, \dots, \mathbf{r}_K\}$. The *average quantization distortion* is defined as,

$$D_Q(\mathcal{X}, \mathcal{R}) = \frac{1}{T} \sum_{t=1}^T \min_{1 \leq k \leq K} d(\mathbf{x}_t, \mathbf{r}_k), \quad (4)$$

where $d(\cdot, \cdot)$ is a distance measure such as the Euclidean distance $\|\mathbf{x}_t - \mathbf{r}_k\|$. A smaller value of (4) indicates higher likelihood for \mathcal{X} and \mathcal{R} originating from the same speaker. Note that (4) is not symmetric [109]: $D_Q(\mathcal{X}, \mathcal{R}) \neq D_Q(\mathcal{R}, \mathcal{X})$.

In theory, it is possible to use all the training vectors directly as the reference template \mathcal{R} . For computational reasons, however, the number of vectors is usually reduced by a clustering method such as K -means [140]. This gives a reduced set of vectors known as *codebook* (Fig. 7). The choice of the clustering method is not as important as optimizing the codebook size [121].

4.2. Gaussian Mixture Model

Gaussian mixture model (GMM) [197, 198] is a stochastic model which has become the *de facto* reference method in speaker recognition. The GMM can be considered as an extension of the VQ model, in which the clusters are overlapping. That is, a feature vector is not assigned to the nearest cluster as in (4), but it has a nonzero probability of originating from each cluster.

A GMM is composed of a finite mixture of multivariate Gaussian components. A GMM, denoted by λ , is characterized by its probability density function:

$$p(\mathbf{x}|\lambda) = \sum_{k=1}^K P_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad (5)$$

In (5), K is the number of Gaussian components, P_k is the prior probability (mixing weight) of the k^{th} Gaussian component, and

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\} \quad (6)$$

is the d -variate Gaussian density function with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. The prior probabilities $P_k \geq 0$ are constrained as $\sum_{k=1}^K P_k = 1$.

For numerical and computational reasons, the covariance matrices of the GMM are usually diagonal (i.e. variance vectors), which restricts the principal axes of the Gaussian ellipses in the direction of the coordinate axes. Estimating the parameters of a full-covariance GMM requires, in general, much more training data and is computationally expensive. As an example for estimating the parameters of a full-covariance GMM, refer to [241].

Monogaussian model uses a single Gaussian component with a full covariance matrix as the speaker model [21, 20, 23, 33, 246]. Sometimes only the covariance matrix is used because the cepstral mean vector is affected by convolutive noise (e.g. due to the microphone/handset). The monogaussian and covariance-only models have a small number of parameters and are therefore computationally efficient, although their accuracy is clearly behind GMM.

Training a GMM consists of estimating the parameters $\lambda = \{P_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ from a training sample $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$. The basic approach is *maximum likelihood* (ML) estimation. The average log-likelihood of \mathcal{X} with respect to model λ is defined as,

$$\text{LL}_{\text{avg}}(\mathcal{X}, \lambda) = \frac{1}{T} \sum_{t=1}^T \log \sum_{k=1}^K P_k \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad (7)$$

The higher the value, the higher the indication that the unknown vectors originate from the model λ . The popular *expectation-maximization* (EM) algorithm [24] can be used for maximizing the likelihood with respect to a given data. Note that *K*-means [140] can be used as an initialization method for EM algorithm; a small number or even no EM iterations are needed according to [124, 128, 181]. This is by no means a general rule, but the iteration count should be optimized for a given task.

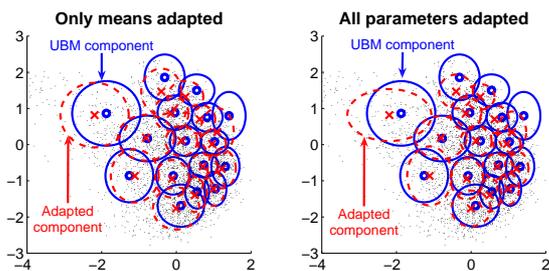


Figure 8: Examples of GMM adaptation using *maximum a posteriori* (MAP) principle. The Gaussian components of a universal background model (solid ellipses) are adapted to the target speaker’s training data (dots) to create speaker model (dashed ellipses).

In speech applications, *adaptation* of the acoustic models to new operating conditions is important because of data variability due to different speakers, environments, speaking styles and so on. In GMM-based speaker recognition, a speaker-independent *world model* or *universal background model* (UBM) is first trained with the EM algorithm from tens or hundreds of hours of speech data gathered from a large number of speakers [197]. The background model represents speaker-independent distribution of the feature vectors. When enrolling a new speaker to the system, the parameters of the background model are adapted to the feature distribution of the new speaker. The adapted model is then used as the model of that speaker. In this way, the model parameters are not estimated from scratch, with prior knowledge (“speech data in general”) being utilized instead. Practice has shown that it is advantageous to train two separate background models, one for female and the other one for male speakers. The new speaker model is then adapted from the background model of the same gender as the new speaker. Let us now look how the adaptation is carried out.

As indicated in Fig. 8, it is possible to adapt all the parameters, or only some of them from the background model. Adapting the means only has been found to work well in practice [197] (this also motivates for a simplified adapted VQ model [88, 124]). Given the enrollment sample, $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$, and the UBM,

$\lambda_{\text{UBM}} = \{P_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, the adapted mean vectors ($\boldsymbol{\mu}'_k$) in the *maximum a posteriori* (MAP) method [197] are obtained as weighted sums of the speaker’s training data and the UBM mean:

$$\boldsymbol{\mu}'_k = \alpha_k \tilde{\mathbf{x}}_k + (1 - \alpha_k) \boldsymbol{\mu}_k, \quad (8)$$

where

$$\alpha_k = \frac{n_k}{n_k + r} \quad (9)$$

$$\tilde{\mathbf{x}}_k = \frac{1}{n_k} \sum_{t=1}^T P(k|\mathbf{x}_t) \mathbf{x}_t \quad (10)$$

$$n_k = \sum_{t=1}^T P(k|\mathbf{x}_t) \quad (11)$$

$$P(k|\mathbf{x}_t) = \frac{P_k \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{m=1}^K P_m \mathcal{N}(\mathbf{x}_t | \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}. \quad (12)$$

The MAP adaptation is to derive a speaker-specific GMM from the UBM. The relevance parameter r , and thus α_k , controls the effect of the training samples on the resulting model with respect to the UBM.

In the recognition mode, the MAP-adapted model and the UBM are coupled, and the recognizer is commonly referred to as *Gaussian mixture model - universal background model*, or simply “GMM-UBM”. The match score depends on both the target model (λ_{target}) and the background model (λ_{UBM}) via the average log likelihood ratio:

$$\begin{aligned} & \text{LLR}_{\text{avg}}(\mathcal{X}, \lambda_{\text{target}}, \lambda_{\text{UBM}}) \\ &= \frac{1}{T} \sum_{t=1}^T \{ \log p(\mathbf{x}_t | \lambda_{\text{target}}) - \log p(\mathbf{x}_t | \lambda_{\text{UBM}}) \}, \end{aligned} \quad (13)$$

which essentially measures the *difference* of the target and background models in generating the observations $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$. The use of a common background model for all speakers makes the match score ranges of different speakers comparable. It is common to apply test segment dependent normalization [14] on top of UBM normalization to account for test-dependent score offsets.

There are alternative adaptation methods to MAP, and selection of the method depends on the amount of available training data [150, 157]. For very short enrollment utterances (a few seconds), some other methods have shown to be more effective. *Maximum likelihood linear regression* (MLLR) [135], originally developed for speech recognition, has been successfully applied to speaker recognition [108, 150, 157, 216]. Both the MAP and MLLR adaptations form a basis for the recent supervector classifiers that we will cover in Section 6.

Gaussian mixture model is computationally intensive due to the frame-by-frame matching. In the GMM-UBM framework [197], the score (13) can be evaluated fast by finding for each test utterance vector the top- C (where usually $C \approx 5$) scoring Gaussians from the UBM [197, 203, 227]. Other speed-up techniques include reducing the numbers of vectors, Gaussian component evaluations, or speaker models [15, 120, 143, 163, 183, 199, 203, 236, 238].

Unlike the hidden Markov models (HMM) in speech recognition, GMM does not explicitly utilize any phonetic information - the training set for GMM simply contains all the spectral features of different phonetic classes pooled together. Because the features of the test utterance and the Gaussian components are not phonetically aligned, the match score may be biased due to different phonemes in training and test utterances.

This phonetic mismatch problem has been attacked with phonetically-motivated tree structures [44, 92] and by using a separate GMM for each phonetic class [40, 61, 81, 180] or for parts of syllables [25]. As an example, *phonetic GMM* (PGMM) described in [40] used neural network classifier for 11 language independent broad phone classes. In the training phase, a separate GMM was trained for each phonetic class and in runtime the GMM corresponding to the frame label was used in scoring. Promising results were obtained when combining PGMM with feature-level intersession combination and with conventional (non-phonetic) GMM. Phonetic modeling in GMMs is clearly worth further studying.

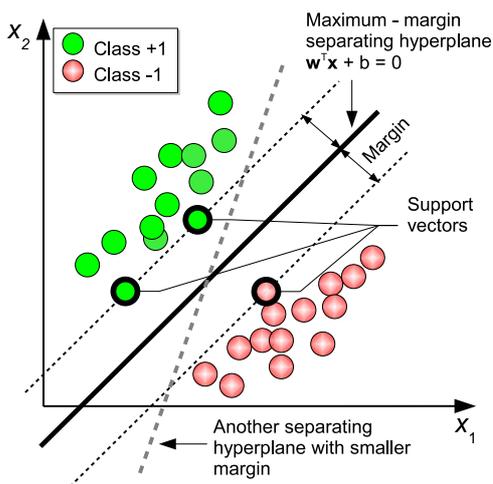


Figure 9: Principle of support vector machine (SVM). A maximum-margin hyperplane that separates the positive (+1) and negative (-1) training examples is found by an optimization process. SVMs have excellent generalization performance.

4.3. Support Vector Machine

Support vector machine (SVM) is a powerful discriminative classifier that has been recently adopted in speaker recognition. It has been applied both with spectral [36, 38], prosodic [204, 67], and high-level features [35]. Currently SVM is one of the most robust classifiers in speaker verification, and it has also been successfully combined with GMM to increase accuracy [36, 38]. One reason for the popularity of SVM is its good generalization performance to classify unseen data.

The SVM, as illustrated in Fig. 9, is a *binary* classifier which models the decision boundary between two classes as a *separating hyperplane*. In speaker verification, one class consists of the target speaker training vectors (labeled as +1), and the other class consists of the training vectors from an “impostor” (background) population (labeled as -1). Using the labeled training vectors, SVM optimizer finds a separating hyperplane that maximizes the *margin* of separation between these two classes.

Formally, the discriminant function of SVM is given by [36],

$$f(\mathbf{x}) = \sum_{i=1}^N \alpha_i t_i K(\mathbf{x}, \mathbf{x}_i) + d. \quad (14)$$

Here $t_i \in \{+1, -1\}$ are the ideal output values, $\sum_{i=1}^N \alpha_i t_i = 0$ and $\alpha_i > 0$. The *support vectors* \mathbf{x}_i , their corresponding weights α_i and the bias term d , are determined from a training set using an optimization process. The *kernel function* $K(\cdot, \cdot)$ is designed so that it can be expressed as $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$, where $\phi(\mathbf{x})$ is a mapping from the input space to kernel feature space of high dimensionality. The kernel function allows computing inner products of two vectors in the kernel feature space. In a high-dimensional space, the two classes are easier to separate with a hyperplane. Intuitively, linear hyperplane in the high-dimensional kernel feature space corresponds to a nonlinear decision boundary in the original input space (e.g. the MFCC space). For more information about SVM and kernels, refer to [24, 169].

4.4. Other Models

Artificial neural networks (ANNs) have been used in various pattern classification problems, including speaker recognition [62, 94, 130, 239]. A potential advantage of ANNs is that feature extraction and speaker modeling can be combined into a single network, enabling joint optimization of the (speaker-dependent) feature extractor and the speaker model [94]. They are also handy in fusing different subsystems [195, 224].

Speaker-specific mapping has been proposed in [153, 164]. The idea is to extract two parallel feature streams with the same frame rate: a feature set representing purely phonetic information (speech content), and a feature set representing a mixture of phonetic and speaker-specific information. The speaker modeling is thus essentially to find a mapping from the “phonetic” spectrum to the “speaker-specific” spectrum by using subspace method [153] or neural network [164].

Representing a speaker relative to other speakers is proposed in [154, 218]. Each speaker model is presented as a combination of some reference models known as the *anchor* models. The combination weights - coordinates in the anchor model space - compose the speaker model. The similarity score between the unknown speech sample and a target model is determined as the distance between their coordinate vectors.

4.5. Fusion

Like in other pattern classification tasks, combining information from multiple sources of evidence - a technique called *fusion* - has been widely applied in speaker recognition [5, 80, 45, 49, 63, 69, 118, 149, 166, 190, 200, 207]. Typically, a number of different feature sets are first extracted from the speech signal; then an individual classifier is used for each feature set; following that the sub-scores or decisions are combined. This implies that each speaker has multiple speaker models stored in the database.

It is also possible to obtain fusion through modelling the *same* features using different classifier architectures, feature normalizations, or training sets [28, 63, 124, 166]. A general belief is that successful fusion system should combine as independent features as possible - low-level spectral features, prosodic features and high-level features. But improvement can also be obtained by fusion of different low-level spectral features (e.g. MFCCs and LPCCs) and different classifiers for them [28, 36, 118]. Fusing dependent (correlated) classifiers can enhance the robustness of the score due to variance reduction [187].

Simplest form of fusion is combining the classifier output scores by weighted sum. That is, given the sub-scores s_k , where k indices the classifier, the fused match score is $s = \sum_{n=1}^{N_c} w_n s_n$. Here N_c is the number of classifiers and w_n is the relative contribution of the n^{th} classifier. The fusion weights w_n can be optimized using a development set, or they can be set as equal ($w_n = 1/N_c$) which does not require weight optimization - but is likely to fail if the accuracies of the individual classifiers are diverse. In cases where the classifier outputs can be interpreted as posterior probability estimates, product

can be used instead of sum. However, the sum rule is the preferred option since the product rule amplifies estimation errors [127]. A theoretically elegant technique for optimizing the fusion weights based on *logistic regression* has been proposed in [28, 29]. An implementation of the method is available in the Fusion and Calibration (FoCal) toolkit². This method, being simple and robust at the same time, is usually the first choice in our own research.

By considering outputs from the different classifiers as another random variable, *score vector*, a backend classifier can be built on top of the individual classifiers. For instance, a support vector machine or a neural network can be trained to separate the genuine and impostor score vectors (e.g. [86, 195, 224, 68]). Upon verifying a person, each of the individual classifiers gives an output score and these scores are in turn arranged into a vector. The vector is then presented to the SVM and the SVM output score is compared against the verification threshold.

Majority of fusion approaches in speaker recognition are based on trial-and-error and optimization on given datasets. The success of a particular combination depends on the performance of the individual systems, as well as their complementariness. Whether the combiner yields improvement on an unseen dataset depends on how the optimization set matches the new dataset (in terms of signal quality, gender distribution, lengths of the training and test material, etc.).

Recently, some improvements to fusion methodology have been achieved by integrating *auxiliary side information*, also known as *quality measures*, into the fusion process [66, 72, 129, 211]. Unlike the traditional methods where the fusion system is trained on development data and kept fixed during run-time, the idea in side-information fusion is to adapt the fusion on each test case. Signal-to-noise ratio (SNR) [129] and nonnative-ness score of the test segment [66] have been used as the auxiliary side information, for instance. Another recent enhancement is to model the correlations between the scores of individual subsystems, since intuitively uncorrelated systems fuse better than correlated ones [68]. Both the auxiliary information and correlation modeling were demonstrated to improve accuracy and are certainly worth further studying.

5. Robust Speaker Recognition

As a carrier wave of phonetic information, affective attributes, speaker characteristics and transmission

²<http://www.dsp.sun.ac.za/~nbrummer/focal/>

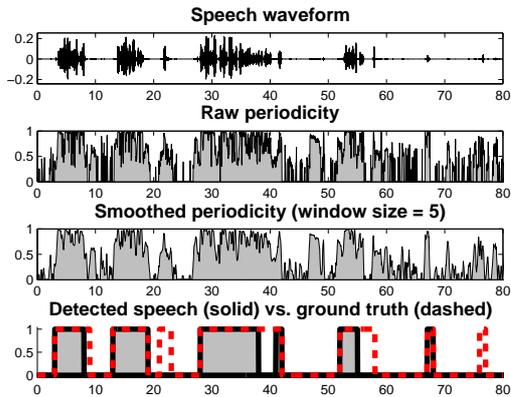


Figure 10: Voice activity detector (VAD) based on periodicity [89]. It is known that voiced speech sounds (vowels, nasals) are more discriminative than fricative and stop sounds. By using periodicity rather than energy may lead to better performance in noisy environments.

path information, the acoustic speech signal is subject to much variations, most of which are undesirable. It is well-known that any mismatch between the training and testing conditions dramatically decreases the accuracy of speaker recognition. The main focus of speaker recognition research has been in tackling this mismatch. Normalization and adaptation methods have been applied to all the parts of speaker recognition systems.

5.1. Voice Activity Detection

Voice activity detector (VAD), as illustrated in Fig. 10, aims at locating the speech segments from a given audio signal [17]. The problem is analogous to face detection from images: we wish to locate the objects of interest before any further processing. VAD is an important sub-component for any real-world recognition system. Even though a seemingly simple binary classification task, it is, in fact, rather challenging to implement a VAD that works consistently across different environments. Moreover, short-duration utterances (few seconds) require special care [64].

A simple solution that works satisfactorily on typical telephone-quality speech data, uses signal energy to detect speech. As an example, we provide a Matlab code fragment in the following:

```
E = 20*log10(std(Frames')+eps); % Energies
max1 = max(E); % Maximum
I = (E>max1-30) & (E>-55); % Indicator
```

Here `Frames` is a matrix that contains the short-term frames of the whole utterance as its row vectors (it is also assumed that the signal values are normalized to the range $[-1, 1]$). This VAD first computes the energies

of all frames, selects the maximum, and then sets the detection threshold as 30 dB below the maximum. Another threshold (-55 dB) is needed for canceling frames with too low an absolute energy. The entire utterance (file) is required before the VAD detection is carried out. A real-time VAD, such as the *long-term spectral divergence* (LTSD) method [191] is required in most real-world systems. Periodicity-based VAD (Fig. 10), an alternative to energy-based methods, was studied in [89].

5.2. Feature Normalization

In principle, it is possible to use generic noise suppression techniques to enhance the quality of the original time-domain signal prior to feature extraction. However, signal enhancement as an additional step in the entire recognition process will increase the computational load. It is more desirable to design a feature extractor which is itself robust [155], or to normalize the features before feeding them onto the modeling or matching algorithms.

The simplest method of feature normalization is to subtract the mean value of each feature over the entire utterance. With the MFCC and LPCC features, this is known as *cepstral mean subtraction* (CMS) or *cepstral mean normalization* (CMN) [12, 70]. In the log-spectral and cepstral domains, convolutive channel noise becomes additive. By subtracting the mean vector, the two feature sets obtained from different channels both become zero-mean and the effect of the channel is correspondingly reduced. Similarly, the variances of the features can be equalized by dividing each feature by its standard deviation. When VAD is used, the normalization statistics are usually computed from the detected speech frames only.

The utterance-level mean and variance normalization assume that channel effect is constant over the entire utterance. To relax this assumption, mean and variance estimates can be updated over a sliding window [228]. The window should be long enough to allow good estimates for the mean and variance, yet short enough to capture time-varying properties of the channel. A typical window size is 3-5 seconds [182, 237].

Feature warping [182] and *short-term Gaussianization* [237] aim at modifying the short-term feature distribution to follow a reference distribution. This is achieved by “warping” the cumulative distribution function of the features so that it matches the reference distribution function, for example a Gaussian. In [182], each feature stream was warped independently. In [237] the independence assumption was relaxed by applying a global linear transformation prior to warping, whose

purpose was to achieve short-term decorrelation or independence of the features. Although Gaussianization was observed to improve accuracy over feature warping [237], it is considerably more complex to implement.

Relative SpecTrAl (RASTA) filtering [99, 153] applies a bandpass filter in the log-spectral or cepstral domain. The filter is applied along the temporal trajectory of each feature, and it suppresses modulation frequencies which are outside of typical speech signals. For instance, a slowly varying convolutive channel noise can be seen as a low-frequency part of the modulation spectrum. Note that the RASTA filter is signal-independent, whereas CMS and variance normalization are adaptive in the sense that they use statistics of the given signal. For useful discussions on data-driven temporal filters versus RASTA, refer to [153].

Mean and variance normalization, Gaussianization, feature warping and RASTA filtering are unsupervised methods which do not explicitly use any channel information. *Feature mapping* (FM) [194] is a supervised normalization method which transforms the features obtained from different channel conditions into a channel-independent feature space so that channel variability is reduced. This is achieved with a set of channel-dependent GMMs adapted from a channel-independent root model. In the training or operational phase, the most likely channel (highest GMM likelihood) is detected, and the relationship between the root model and the channel-dependent model is used for mapping the vectors into channel-independent space. A generalization of the method which does not require detection of the top-1 Gaussian component was proposed in [245].

Often different feature normalizations are used in combination. A typical robust front-end [196] consists of extracting MFCCs, followed by RASTA filtering, delta feature computation, voice activity detection, feature mapping and global mean/variance normalization in that order. Different orders of the normalization steps are possible; in [31] cepstral vectors were first processed through global mean removal, feature warping, and RASTA filtering, followed by adding first-, second-, and third-order delta features. Finally, voice activity detector and dimensionality reduction using heteroscedastic linear discriminant analysis (HLDA) were applied.

Graph-theoretic compensation method was proposed in [87]. This method considered the training and test utterances as graphs where the graph nodes correspond to “feature points” in the feature space. The matching was then carried out by finding the corresponding feature point pairs from the two graphs based on graph isomorphism, and used for global transformation of the feature space, followed by conventional match-

ing. The graph structure was motivated by invariance against the affine feature distortion model for cepstral features (e.g. [151, 155]). The method requires further development to validate the assumptions of the feature distortion model and to improve computational efficiency.

5.3. Speaker Model Compensation

Model-domain compensation involves modifying the speaker model parameters instead of the feature vectors. One example is *speaker model synthesis* (SMS) [219], which adapts the target GMM parameters into a new channel condition, if this condition has not been present in the enrollment phase. This is achieved with the help of transformations between a channel-independent background model and channel-dependent adapted models. Roughly, speaker model synthesis is a model-domain equivalent of feature mapping (FM) [194]. Feature mapping can be considered more flexible since the mapped features can be used with any classifier and not only with the GMM.

Both SMS and FM require a labeled training set with training examples from a variety of different channel conditions. In [162], an unsupervised clustering of the channel types was proposed so that labeling would not be needed. The results indicate that feature mapping based on unsupervised channel labels achieves equal or better accuracy compared with supervised labeling. It should be noted, however, that state-of-the-art speaker modeling with supervectors use *continuous* intersession variability models and therefore extend the SMS and FM methods to handle with unknown conditions. The continuous model compensation methods have almost completely surpassed the SMS and FM methods, and will be the focus of Section 6.

5.4. Score Normalization

In *score normalization*, the “raw” match score is normalized relative to a set of other speaker models known as *cohort*. The main purpose of score normalization is to transform scores from different speakers into a similar range so that a common (speaker-independent) verification threshold can be used. Score normalization can correct some speaker-dependent score offsets not compensated by the feature and model domain methods.

A score normalization of the form

$$s' = \frac{s - \mu_I}{\sigma_I} \quad (15)$$

is commonly used. In (15), s' is the normalized score, s is the original score, and μ_I and σ_I are the estimated mean and standard deviation of the impostor

score distribution, respectively. In *zero normalization* (“Z-norm”), the impostor statistics μ_I and σ_I are target speaker dependent and they are computed off-line in the speaker enrollment phase. This is done by matching a batch of non-target utterances against the target model, and obtaining the mean and standard deviation of those scores. In *test normalization* (“T-norm”) [14], the parameters are test utterance dependent and they are computed “on the fly” in the verification phase. This is done by matching the unknown speaker’s feature vectors against a set of impostor models and obtaining the statistics.

Usually the cohort models are common for all speakers, however, speaker-dependent cohort selection for T-norm has been studied in [192, 217]. Z-norm and T-norm can also be combined. According to [229], Z-norm followed by T-norm does produce good results.

Score normalization can be improved by using side information such as channel type. Handset-dependent background models were used in [95]. The handset type (carbon button or electret) through which the training utterance is channeled was automatically detected, and the corresponding background model was used for score normalization in the verification phase. In [197], handset-dependent mean and variance of the likelihood ratio were obtained for each target speaker. In the matching phase, the most likely handset was detected and the corresponding statistics were used to normalize the likelihood ratio. In essence, this approach is a handset-dependent version of Z-norm, which the authors call “H-norm”. In a similar way, handset-dependent T-norm (“HT-norm”) has been proposed [58]. Note that the handset-dependent normalization approaches [58, 95, 197] require an automatic handset labeler which inevitable makes classification errors.

Although Z-norm and T-norm can be effective in reducing speaker verification error rates, they may seriously fail if the cohort utterances are badly selected, that is, if their acoustic and channel conditions differ too much from the typical enrollement and test utterances of the system. According to [31], score normalization may not be needed at all if the other components, most notable eigenchannel compensation of speaker models, are well-optimized. However, Z- and T-norms and their combinations seem to be an essential necessity for the more complete joint factor analysis model [112]. In summary, it remains partly a mystery when score normalization is useful, and would deserve more research.

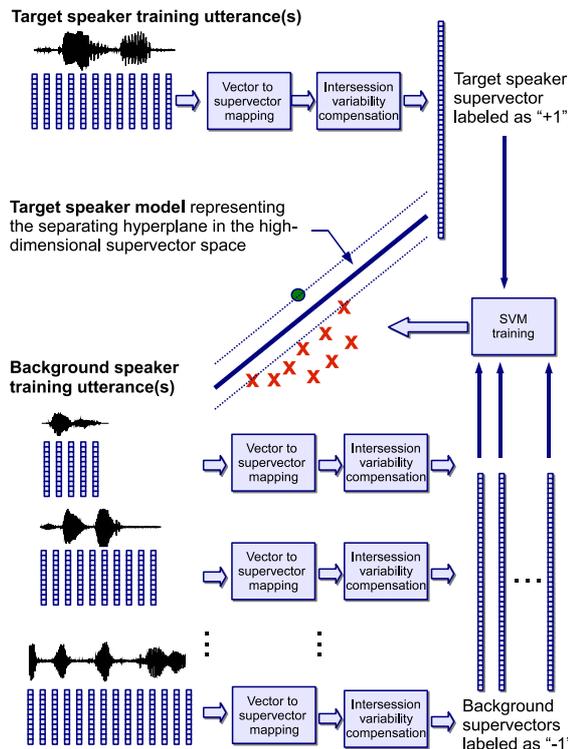


Figure 11: The concept of modern sequence kernel SVM. Variable-length utterances are mapped into fixed-dimensional *supervectors*, followed by intersession variability compensation and SVM training.

6. Supervector Methods: a Recent Research Trend

6.1. What is a Supervector?

One of the issues in speaker recognition is how to represent utterances that, in general, have a varying number of feature vectors. In early studies [158] speaker models were generated by time-averaging features so that each utterance could be represented as a single vector. The average vectors would then be compared using a distance measure [119], which is computationally very efficient but gives poor recognition accuracy. Since the 1980s, the predominant trend has been creating a model of the training utterance followed by “data-to-model” type of matching at run-time (e.g. likelihood of an utterance with respect to a GMM). This is computationally more demanding but gives good recognition accuracy.

Interestingly, the speaker recognition community has recently re-discovered a robust way to present utterances using a single vector, a so-called *supervector*. On one hand, these supervectors can be used as inputs to support vector machine (SVM) as illustrated in Fig. 11. This leads to *sequence kernel SVMs*, where the

utterances with variable number of feature vectors are mapped to a fixed-length vector using the sequence kernel; for review and useful insights, refer to [141, 232]. On the other hand, conventional adapted Gaussian mixture speaker model [197] can also be seen as a supervector. Combinations of generative models and SVM have also lead to good results [38].

Often “supervector” refers to combining many smaller-dimensional vectors into a higher-dimensional vector; for instance, by stacking the d -dimensional mean vectors of a K -component adapted GMM into a Kd -dimensional Gaussian supervector [38]. In this paper, we understand supervector in a broader sense as any high- and fixed-dimensional representation of an utterance. It is important that the supervectors of different utterances arise from a “common coordinate system” such as being adapted from a universal background model, or being generated using a fixed polynomial basis [36]. In this way the supervector elements are meaningfully aligned and comparable when doing similarity computations in the supervector space. With SVMs, normalizing the dynamic ranges of the supervector elements is also crucial since SVMs are not scale invariant [232].

An important recent advance in speaker recognition has been the development of *explicit inter-session variability compensation* techniques [31, 112, 231]. Since each utterance is now presented as a single point in the supervector space, it becomes possible to directly quantify and remove the unwanted variability from the supervectors. Any variation in different utterances of the same speaker, as characterized by their supervectors – be it due to different handsets, environments, or phonetic content – is harmful.

Does this mean that we will need several training utterances recorded through different microphones or environments when enrolling a speaker? Not necessarily. Rather, the intersession variability model is trained on an independent development data and then removed from the supervectors of a new speaker. The intersession variability model itself is continuous, which is in contrast with speaker model synthesis (SMS) [219] and feature mapping (FM) [194] discussed in Section 5. Both SMS and FM assume a discrete collection of recording conditions (such as mobile/landline channels or carbon button/electrec handsets). However, the explicit inter-session variability normalization techniques enable modeling channel conditions that “fall in between” some conditions that are not seen in training data.

Various authors have independently developed different session compensation methods for both GMM- and SVM-based speaker models. *Factor analysis* (FA) tech-

niques [110] are designed for the GMM-based recognizer and take explicit use of stochastic properties of the GMM, whereas the methods developed for SVM supervectors are often based on numerical linear algebra [212]. To sum up, two core design issues with the modern supervector based recognizers are 1) how to create the supervector of an utterance, 2) how to estimate and apply the session variability compensation in the supervector space. In addition, the question of how to compute the match score with the session-compensated models needs to be solved [74].

6.2. GLDS Kernel SVM

One of the simplest SVM supervectors is *generalized linear discriminant sequence* (GLDS) kernel [36]. The GLDS method creates the supervector by explicit mapping into kernel feature space using a *polynomial expansion* [34], denoted here as $\mathbf{b}(\mathbf{x})$. As an example, 2nd order polynomial expansion for a 2-dimensional vector $\mathbf{x} = (x_1, x_2)^T$ is given by $\mathbf{b}(\mathbf{x}) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^T$. During enrollment, all the background speaker and target speaker utterances $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ are represented as average expanded feature vectors:

$$\mathbf{b}_{\text{avg}} = \frac{1}{T} \sum_{t=1}^T \mathbf{b}(\mathbf{x}_t). \quad (16)$$

The averaged vectors are further variance-normalized using the background utterances, and assigned with the appropriate label for SVM training (+1=target speaker vectors; -1=background speaker vectors). The SVM optimization (using standard linear kernel) yields a set of support vectors \mathbf{b}_i , their corresponding weights α_i and a bias d . These are collapsed into a single model vector as,

$$\mathbf{w} = \sum_{i=1}^L \alpha_i t_i \mathbf{b}_i + \mathbf{d}, \quad (17)$$

where $\mathbf{d} = (d, 0, 0, \dots, 0)^T$ and $t_i \in \{+1, -1\}$ are the ideal outputs (class labels of the support vectors), and L is the number of support vectors. In this way, the speaker model can be presented as a single supervector. The collapsed model vector \mathbf{w} is also normalized using background utterances, and it serves as the model of the target speaker.

The match score in the GLDS method is computed as an inner product $s = \mathbf{w}_{\text{target}}^T \mathbf{b}_{\text{test}}$, where $\mathbf{w}_{\text{target}}$ denotes the normalized model vector of the target speaker and \mathbf{b}_{test} denotes the normalized average expanded feature vector of the test utterance. Since all the speaker models and the test utterance are represented as single vectors,

the verification phase is computationally efficient. The main drawback of the GLDS method is that it is difficult to control the dimensionality of the supervectors; in practice, the polynomial expansion includes either 2nd or 3rd order monomials before the dimensionality gets infeasible.

6.3. Gaussian Supervector SVM

Since the universal background model (UBM) is included as a part in most speaker recognition systems, it provides a natural way to create supervectors [38, 52, 132]. This leads to *hybrid* classifier where the generative GMM-UBM model is used for creating “feature vectors” for the discriminative SVM.

In [38] the authors derive the *Gaussian supervector* (GSV) kernel by bounding the Kullback-Leibler (KL) divergence measure between GMMs. Suppose that we have the UBM, $\lambda_{\text{UBM}} = \{P_k, \mu_k, \Sigma_k\}_{k=1}^K$, and two utterances a and b which are described by their MAP-adapted GMMs (Subsection 4.2). That is, $\lambda_a = \{P_k, \mu_k^a, \Sigma_k\}_{k=1}^K$ and $\lambda_b = \{P_k, \mu_k^b, \Sigma_k\}_{k=1}^K$ (note that the models differ only in their means). The KL divergence kernel is then defined as,

$$K(\lambda_a, \lambda_b) = \sum_{k=1}^K (\sqrt{P_k \Sigma_k^{-1/2}} \mu_k^a)^T (\sqrt{P_k \Sigma_k^{-1/2}} \mu_k^b). \quad (18)$$

From the the implementation point of view, this just means that all the Gaussian means μ_k need to be normalized with $\sqrt{P_k \Sigma_k^{-1/2}}$ before feeding them into SVM training. Again, this is a form of variance normalization. Hence, even though only the mean vectors of the GMM are included in the supervector, the variance and weight information of the GMM is implicitly present in the role of normalizing the Gaussian supervector. It is also possible to normalize all the adapted GMM supervectors to have a constant distance from the UBM [53]. As in the GLDS kernel, the speaker model obtained via SVM optimization can be compacted as a single model supervector.

A recent extension to Gaussian supervectors is based on bounding the Bhattacharyya distance [240]. This leads to a *GMM-UBM mean interval* (GUMI) kernel to be used in conjunction with SVM. The GUMI kernel exploits the speaker’s information conveyed by the mean of GMM as well as those by the covariance matrices in an effective manner. Another alternative kernel known as *probabilistic sequence kernel* (PSK) [132, 133] uses output values of the Gaussian functions rather than the Gaussian means to create a supervector. Since the individual Gaussians can be assumed to present phonetic

classes [198], the PSK kernel can be interpreted as presenting high-level information related to phone occurrence probabilities.

6.4. MLLR Supervector SVM

In [108, 216], the authors use *Maximum likelihood linear regression* (MLLR) transformation parameters as inputs to SVM. MLLR transforms the mean vectors of a speaker-independent model as $\mu'_k = \mathbf{A}\mu_k + \mathbf{b}$, where μ'_k is the adapted mean vector, μ_k is the world model mean vector and the parameters \mathbf{A} and \mathbf{b} define the linear transform. The parameters \mathbf{A} and \mathbf{b} are estimated by maximizing the likelihood of the training data with a modified EM algorithm [135]. Originally MLLR was developed for speaker adaptation in speech recognition [135] and it has also been used in speaker recognition as an alternative to maximum a posteriori (MAP) adaptation of the universal background model (UBM) [150].

The key differences between MLLR and Gaussian supervectors are in the underlying speech model - phonetic hidden Markov models versus GMMs, and the adaptation method employed - MLLR versus *maximum a posteriori* (MAP) adaptation. MLLR is motivated to benefit from more detailed speech model and the efficient use of data through transforms that are shared across Gaussians [216]. Independent studies [41, 136] have shown that detailed speech model improve the speaker characterization ability of supervectors.

A similar work to MLLR supervectors is to use *feature transformation* (FT) parameters as inputs to SVM [243], where a flexible FT function clusters transformation matrices and bias vectors with different regression classes. The FT framework is based on GMM-UBM rather than hidden Markov model, therefore, does not require a phonetic acoustic system. The FT parameters are estimated with the MAP criteria that overcome possible numerical problems with insufficient training. A recent extension of this framework [244] includes the joint MAP adaptation of FT and GMM parameters.

6.5. High-Level Supervector SVM

The GLDS-, GMM- and MLLR-supervectors are suitable for modeling short-term spectral features. For the prosodic and high-level features (Subsections 3.4 and 3.5), namely, features created using a tokenizer front-end, it is customary to create a supervector by concatenating the uni-, bi- and tri-gram ($N = 1, 2, 3$) frequencies into a vector or *bag-of-N-grams* [35, 204]. The authors of [35] developed *term frequency log likelihood ratio* (TFLLR) kernel that normalizes the original N -gram frequency by $1/\sqrt{f_i}$, where f_i is the overall frequency of that N -gram. Thus the value of rare N -grams

is increased and the value of frequent N -grams is decreased, thereby equalizing their contribution in kernel computations.

The high-level features created by a phone tokenizer, or by quantization of prosodic feature values by binning [204], are inherently noisy: tokenizer error (e.g. phone recognizer error) or small variation in the original feature value may cause the feature to fall into a wrong category (bin). To tackle this problem, the authors of [67] proposed to use *soft* binning with the aid of Gaussian mixture model and use the weights of the Gaussians as the features for SVM supervector.

6.6. Normalizing SVM Supervectors

Two forms of SVM supervector normalizations are necessary: normalizing the dynamic range of features and intersession variability compensation. The first one, normalizing the dynamic range, is related to the inherent property of the SVM model. SVM is not invariant to linear transformations in feature space and some form of variance normalization is required so that certain supervector dimensions do not dominate the inner product computations. Often variance normalization is included in the definition of the kernel function and specific to a given kernel as seen in the previous subsections. Kernel-independent *rank normalization* has also been successfully applied [215]. Rank normalization replaces each feature by its relative position (rank) in the background data. For useful insights on normalization, refer to [215, 232]. Let us now turn our focus to the other necessary normalization, the intersession variability compensation.

Nuisance attribute projection (NAP) is a successful method for compensating SVM supervectors [37, 212]. It is not specific to some kernel, but can be applied to *any* kind of SVM supervectors. The NAP transformation removes the directions of undesired sessions variability from the supervectors before SVM training. The NAP transformation of a given supervector \mathbf{s} is [28],

$$\mathbf{s}' = \mathbf{s} - \mathbf{U}(\mathbf{U}^T \mathbf{s}), \quad (19)$$

where \mathbf{U} is the *eigenchannel* matrix. The eigenchannel matrix is trained using a development dataset with a large number of speakers, each having several training utterances (sessions). The training set is prepared by subtracting the mean of the supervectors within each speaker and pooling all the supervectors from different speakers together; this removes most of the speaker variability but leaves session variability. By performing eigen-analysis on this training set, one captures the

principal directions of channel variability. The underlying assumption is that the session variability lies in a speaker-independent low-dimensional subspace; after training the projection matrix, the method can be applied for unseen data with different speakers. The equation (19) then just means subtracting the supervector that has been projected on the channel space. For practical details of NAP, refer to [28, 65].

removed by NAP may contain speaker-specific information [230]. Moreover, session compensation and SVM optimization processes are treated independently from each other. Motivated with these facts, discriminative variant of NAP has been studied in [30, 230]. In [230], *scatter difference analysis* (SDA), a similar method to linear discriminant analysis (LDA), was used for optimizing the NAP projection matrix, and in [30], the session variability model was directly integrated within the optimization criterion of the SVM; this leaves the decision about usefulness of the supervector dimensions for the SVM optimizer. This approach improved recognition accuracy over the NAP baseline in [30], albeit introducing a new control parameter that controls the contribution of the nuisance subspace constraint. Nevertheless, discriminative session compensation is certainly an interesting new direction for future studies.

Within-class covariance normalization (WCCN), another SVM supervector compensation method similar to NAP, was proposed in [85]. The authors considered generalized linear kernels of the form $K(\mathbf{s}_1, \mathbf{s}_2) = \mathbf{s}_1 \mathbf{R} \mathbf{s}_2$, where \mathbf{s}_1 and \mathbf{s}_2 are supervectors and \mathbf{R} is a positive semidefinite matrix. With certain assumptions, a bound of a binary classification error metric can be minimized by choosing $\mathbf{R} = \mathbf{W}^{-1}$, where \mathbf{W} is the expected within-class (within-speaker) covariance matrix. The WCCN was then combined with principal component analysis (PCA) in [84] to attack the problem of estimating and inverting \mathbf{W} to large data sets. The key difference between NAP and WCCN is the way how they weight the dimensions in the supervector space [216]. The NAP method completely removes some of the dimensions by projecting the supervectors to a lower-dimensional space, whereas WCCN weights rather than completely removes the dimensions.

6.7. Factor Analysis Techniques

In the previous subsection we focused on compensating SVM supervectors. We will now discuss a different technique based on generative modeling, that is, Gaussian mixture model (GMM) with factor analysis (FA) technique. Recall that the MAP adaptation technique for GMMs [197], as described in Section 4.2, adapts

the mean vectors of the universal background model (UBM) while the weights and covariances are shared between all speakers. Thus a speaker model is uniquely represented as the concatenation of the mean vectors, which can be interpreted as a supervector.

For a given speaker, the supervectors estimated from different training utterances may not be the same especially when these training samples come from different handsets. Channel compensation is therefore necessary to make sure that test data obtained from different channel (than that of the training data) can be properly scored against the speaker models. For channel compensation to be possible, the channel variability has to be modelled explicitly. The technique of *joint factor analysis* (JFA) [110] was proposed for this purpose.

The JFA model considers the variability of a Gaussian supervector as a linear combination of the speaker and channel components. Given a training sample, the speaker-dependent and channel-dependent supervector \mathbf{M} is decomposed into two statistically independent components, as follows

$$\mathbf{M} = \mathbf{s} + \mathbf{c}, \quad (20)$$

where \mathbf{s} and \mathbf{c} are referred to as the *speaker* and *channel* supervectors, respectively. Let d be the dimension of the acoustic feature vectors and K be the number of mixtures in the UBM. The supervectors \mathbf{M} , \mathbf{s} and \mathbf{c} live in a Kd -dimensional parameter space. The channel variability is explicitly modeled by the channel model of the form,

$$\mathbf{c} = \mathbf{U}\mathbf{x}, \quad (21)$$

where \mathbf{U} is a rectangular matrix and \mathbf{x} are the *channel factors* estimated from a given speech sample. The columns of the matrix \mathbf{U} are the *eigenchannels* estimated for a given dataset. During enrollment, the channel factors \mathbf{x} are to be estimated jointly with the speaker factors \mathbf{y} of the speaker model of the following form:

$$\mathbf{s} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z}. \quad (22)$$

In the above equation, \mathbf{m} is the UBM supervector, \mathbf{V} is a rectangular matrix with each of its columns referred to as the eigenvoices, \mathbf{D} is $Kd \times Kd$ diagonal matrix and \mathbf{z} is a $Kd \times 1$ column vector. In the special case $\mathbf{y} = \mathbf{0}$, $\mathbf{s} = \mathbf{m} + \mathbf{D}\mathbf{z}$ describes exactly the same adaptation process as the MAP adaptation technique (Section 4.2). Therefore, the speaker model in the JFA technique can be seen as an extension to the MAP technique with the eigenvoice model $\mathbf{V}\mathbf{y}$ included, which has been shown to be useful for short training samples.

The matrices \mathbf{U} , \mathbf{V} and \mathbf{D} are called the *hyperparameters* of the JFA model. These matrices are estimated beforehand on large datasets. One possible way is to first estimate \mathbf{V} followed by \mathbf{U} and \mathbf{D} [110, 112]. For a given training sample, the latent factors \mathbf{x} and \mathbf{y} are jointly estimated and followed by estimation of \mathbf{z} . Finally, the channel supervector \mathbf{c} is discarded and the speaker supervector \mathbf{s} is used as the speaker model. By doing so, channel compensation is accomplished via the explicit modeling of the channel component during training. For detailed account of estimation procedure the reader should refer to [110, 112]. For comparing various scoring methods, refer to [74].

The JFA model dominated the latest NIST 2008 speaker recognition evaluation (SRE) [175] and it was pursued further in the Johns Hopkins University (JHU) summer 2008 workshop [30]. Independent evaluations by different research groups have clearly indicated the potential of JFA. The method has a few practical deficiencies, however. One is sensitivity to training and test lengths (and their mismatch), especially for short utterances (10–20 seconds). The authors of [30] hypothesized that this was caused by *within-session* variability (due to phonemic variability) rather than *inter-session* variability captured by the baseline JFA. The authors then extended the JFA model by explicitly adding a model of the within-session variability. Other choices to tackle the JFA dependency on utterance length were studied as well - namely, utilizing variable length development utterances to create *stacked* channel matrix. The extended JFA and the stacking approach both showed improvement over the baseline JFA when the training and test utterance lengths were not matched, hence improving the generalization of JFA for unknown utterance lengths. The within-session variability modeling, however, has a price: a phone recognizer was used for generating data for within-session modeling. It may be worthwhile to study simplified approach – segmenting the data into fixed-length chunks – as proposed in [30].

Given the demonstrated excellent performance of the JFA compensation and Gaussian supervector SVMs [38], it seems appropriate to ask how they compare with each other, and whether they could be combined? These questions were recently addressed in [53, 54]. In [53] the authors compared JFA and SVM both with linear and nonlinear kernels, compensated with nuisance attribute projection (NAP). They concluded that JFA without speaker factors gives similar accuracy to SVM with Gaussian supervectors; however, JFA outperformed SVM when speaker factors were added. In [54] the same authors used the speaker factors of the JFA model as inputs to SVM. Within-class covariance

normalization (WCCN) [216] was used instead of NAP. The results indicated that using the speaker factors in SVM is effective but the accuracy was not improved over the JFA-compensated GMM. The combined JFA-SVM method, however, results in faster scoring.

6.8. Summary: Which Supervector Method to Use?

Given the multiple choices to create a supervector and to model intersession variability, which one to choose for practical use? It is somewhat difficult to compare the methods in literature due to differences in data set selections, parameter settings and other implementation details. However, there are some common practice that we can follow. To facilitate discussion, we present here the results of the latest NIST 2008 speaker recognition evaluation submission by the I4U consortium [138]. All the classifiers of I4U used short-term spectral features and the focus was in the supervectors classifiers. Three well-known methods - Gaussian mixture model-universal background model (GMM-UBM) [197], generalized linear discriminant sequence (GLDS) kernel SVM [36] and Gaussian supervector (GSV) kernel SVM (GSV-SVM) [38] were studied. In addition, three novel SVM kernels were proposed: feature transformation kernel (FT-SVM) [244], probabilistic sequence kernel (PSK-SVM) [132, 133] and Bhattacharyya kernel (BK-SVM) [240].

Table 1 reports the performance of individual systems, together with the weighted summation fusion of the classifiers. The accuracy is measured in *equal error rate* (EER), a verification error measure that gives the accuracy at decision threshold for which the probabilities of false rejection (miss) and false acceptance (false alarm) are equal (see Section 7).

From the results in Table 1 it is clear that intersession compensation significantly improves the accuracy of the GMM-UBM system. It can also be seen that the best individual classifier is the GMM-UBM system with JFA compensation, and that JFA outperforms the eigenchannel method (which is a special case of JFA). Finally, fusing all the session-compensated classifiers improves accuracy as expected.

Even though JFA outperforms the SVM-based methods, for practitioners we recommend to start with the two simplest approaches at this moment: GLDS-SVM and GSV-SVM. The former does not require much optimization whereas the latter comes almost as a by-product when a GMM-UBM system is used. Furthermore, they do not require as many datasets as JFA does, are simple to implement and fast in computation. They should be augmented with nuisance attribute projection (NAP) [28] and test normalization (T-norm) [14].

Table 1: Performance of individual classifiers and their fusion of I4U system on I4U’s telephone quality development dataset [138]. UNC = Uncompensated, EIG = Eigenchannel, JFA = Joint factor analysis, GLDS = Generalized linear discriminant sequence, GSV = Gaussian supervector, FT = Feature transformation, PSK = Probabilistic sequence kernel, BK = Bhattacharyya kernel. All the SVM-based systems use nuisance attribute projection (NAP) compensation.

	Tuning set EER (%)	Eval. set EER (%)
Gaussian mixture model		
1. GMM-UBM (UNC)	8.45	8.10
2. GMM-UBM (EIG) [112]	5.47	5.22
3. GMM-UBM (JFA) [112]	3.19	3.11
Support vector machine with different kernels		
4. GLSD-SVM [36]	4.30	4.44
5. GSV-SVM [38]	4.47	4.43
6. FT-SVM [243]	4.20	3.66
7. PSK-SVM [132]	5.29	4.77
8. BK-SVM [240]	4.46	5.16
Fusing systems 2 to 8	2.49	2.05

7. Performance Evaluation and Software Packages

7.1. Performance Evaluation

Assessing the performance of new algorithms on a common dataset is essential to enable meaningful performance comparison. In early studies, corpora consisted of a few or at the most a few dozen speakers, and data was often self-collected. Recently, there has been significant effort directed towards standardizing the evaluation methodology in speaker verification.

The National Institute of Standards and Technology (NIST)³ provides a common evaluation framework for text-independent speaker recognition methods [156]. NIST evaluations include test trials under both matched conditions such as telephone only, and unmatched conditions such as language effects (matched languages vs unmatched languages), cross channel and two-speaker detection. NIST has conducted speaker recognition benchmarking on an annual basis since 1997, and registration is open to all parties interested in participating in this benchmarking activity. During the evaluation, NIST releases a set of speech files as the development data to the participants. At this initial phase, the participants do not have access to the “ground truth”, that is, the speaker labels. Each participating group then runs their algorithms “blindly” on the given data and submits the recognition scores and verification decisions. NIST then evaluates the performances of the submissions and

³<http://nist.gov/>

the results are discussed in a follow-up workshop. The use of “blind” evaluation data makes it possible to conduct an unbiased comparison of the various algorithms. These activities would be difficult without a common evaluation dataset or a standard evaluation protocol.

Visual inspections of the *detection error trade-off* (DET) curves [159] and *equal error rate* (EER) are commonly used evaluation tools in the speaker verification literature. The problem with EER is that it corresponds to an arbitrary detection threshold, which is not a likely choice in a real application where it is critical to maintain the balance between user convenience and security. NIST uses a *detection cost function* (DCF) as the primary evaluation metric to assess speaker verification performance:

$$\text{DCF}(\Theta) = 0.1 \times P_{\text{miss}}(\Theta) + 0.99 \times P_{\text{fa}}(\Theta). \quad (23)$$

Here $P_{\text{miss}}(\Theta)$ and $P_{\text{fa}}(\Theta)$ are the probabilities of *miss* (i.e. rejection of a genuine speaker) and *false alarm* (i.e. acceptance of an impostor), respectively. Both of them are functions of a global (speaker-independent) verification threshold Θ .

Minimum DCF (MinDCF), defined as the DCF value at the threshold for which (23) is smallest, is the optimum cost. When the decision threshold is optimized on a development set and applied to the evaluation corpus, this produces *actual DCF*. Therefore, the difference between the minimum DCF and the actual DCF indicates how well the system is *calibrated* for a certain application and how robust is the threshold setting method. For an in-depth and thorough theoretical discussion as well as the alternative formulations of application-independent evaluation metrics, refer to [29].

While the NIST speaker recognition benchmarking considers mostly conversational text-independent speaker verification in English, there have been a few alternative evaluations, for instance the NFI-TNO evaluation⁴ which considered authentic forensic samples (mostly in Dutch), including wiretap recordings. Another evaluation, specifically for Chinese, was organized in conjunction with the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP’06)⁵. This evaluation included open-set speaker identification and text-dependent verification tasks in addition to text-independent verification.

Some of the factors affecting speaker recognition accuracy in the NIST and NFI-TNO evaluations have been

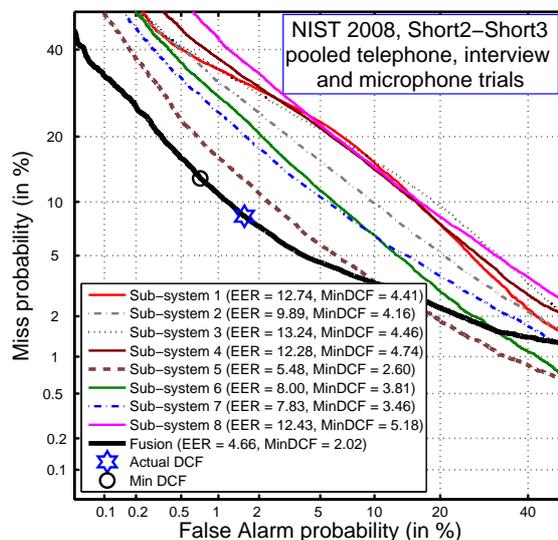


Figure 12: Example of detection error trade-off (DET) plot presenting various subsystems and a combined system using score-level fusion.

analyzed in [134]. It is widely known that cross-channel training and testing display a much lower accuracy compared to that with same channel. Including different handsets in the training material also improves recognition accuracy. Another factor significant to performance is the duration of training and test utterances. The greater the amount of speech data used for training and/or testing, the better the accuracy. Training utterance duration seems to be more significant than test segment duration.

7.2. Software Packages for Speaker Recognition

As can be seen throughout this article, the state-of-the-art speaker recognition methods are getting more and more advanced and they often combine several complementary techniques. Implementing a full system from scratch may not be meaningful. In this sub-section we point out a few useful software packages that can be used for creating a state-of-the-art speaker recognition system.

Probably the most comprehensive and up-to-date software package is ALIZE toolkit⁶, an open-source software developed at Université d’Avignon, France. For more details, the interested reader is referred to [65].

⁴<http://speech.tn.tno.nl/aso/>

⁵<http://www.iscslp2006.org/>

⁶Now under “Mistral” platform for biometrics authentication. Available at: <http://mistral.univ-avignon.fr/en/>

For research purposes, it is possible to build up a complete speaker recognition system using various different software packages. The Matlab software by MathWorks Inc. is excellent especially for developing new feature extraction methods. Octave⁷ is an open-source alternative to Matlab is, and there are a plenty of free toolboxes for both of them such as Statistical Pattern Recognition Toolbox⁸ and NetLab⁹. Aside from Matlab/Octave, the Hidden Markov Model Toolkit (HTK)¹⁰ is also popular in statistical modeling, whereas Torch¹¹ software represents state-of-the-art SVM implementation.

For score fusion of multiple sub-systems, we recommend the FoCal toolkit¹². For evaluation purposes, such as plotting DET curves, we recommend the DETware toolbox (for Matlab) by NIST¹³. A similar tool but with more features is SRETools¹⁴.

8. Future Horizons of Speaker Recognition

During the past ten years, speaker recognition community has made significant advances in the technology. In summary, we have selected a few of the most influential techniques that have been proven to work in practice in independent studies, or shown significant promise in the past few NIST technology evaluation benchmarks:

- Universal background modeling (UBM) [197]
- Score normalization, calibration, fusion [14, 31]
- Sequence kernel SVMs [36, 38]
- Use of prosodics and high-level features with SVM [35, 204, 216]
- Phonetic normalization using ASR [41, 216]
- Explicit session variability modeling and compensation [28, 41, 84, 112].

Even though effective, these methods are highly data-driven and massive amounts of data are needed for training the background models, cohort models for score

normalization, and modeling session and speaker variabilities. The data sets need to be labeled and organized in a controlled manner requiring significant human efforts. It is not trivial to decide how to split the system development data for UBM training, session modeling, and score normalization. If the development data conditions do not match to those of the expected operation environment, the accuracy will drop significantly, sometimes to unusable level. It is clear that laborious design of data set splits cannot be expected, for instance, from forensic investigators who just want to use speaker recognition software in “turnkey” fashion.

For transferring the technology into practice, therefore, in future it will be important to focus on making the methods less sensitive to selection of the data sets. The methods also require computational simplifications before they can be used in real-world applications such as in smart cards or mobile phones, for instance. Finally, the current techniques require several *minutes* of training and test data to give satisfactory performance, that presents a challenge for applications where real-time decision is desired. For instance, the core evaluation condition in recent NIST benchmarkings uses about 2.5 minutes of speech data. New methods for short training and test utterances (less than 10 seconds) will be needed. The methods for long data do not readily generalize to short-duration tasks as indicated in [27, 30, 64].

The NIST speaker recognition evaluations [156, 134] have systematized speaker recognition methodology development and constant positive progress has been observed in the past years. However, the NIST evaluations have mostly focused on combating *technical* error sources, most notably that of training/test channel mismatch (for instance, using different microphones in training and test material). There are also many other factors that have impacts on the speaker recognition performance. We should also address *human-related* error sources, such as the effects of emotions, vocal organ illness, aging, and level of attention. Furthermore, one of the most popular questions asked by laymen is “what if someone or some machine imitates me or just plays previously recorded signal back?”. Before considering speaker recognition in large-scale commercial applications, the research community must answer such questions. These questions have been considered in some studies, mostly in the context of phonetic sciences, but always for a limited number of speakers and using non-public corpora. As voice transformation technique advances, low cost voice impersonation becomes possible [27, 184]. This opens up a new horizon to study attack and defense in voice biometrics.

⁷<http://www.gnu.org/software/octave/>

⁸<http://cmp.felk.cvut.cz/cmp/software/stprtool/>

⁹<http://www.ncrg.aston.ac.uk/netlab/index.php>

¹⁰<http://htk.eng.cam.ac.uk/>

¹¹<http://www.torch.ch/>

¹²<http://niko.brummer.googlepages.com/focal>

¹³http://www.itl.nist.gov/iad/mig/tools/DETware_v2.1.targz.htm

¹⁴<http://sretools.googlepages.com/>

Much of the recent progress in speaker recognition is attributed to the success in classifier design and session compensation, which largely rely on traditional short-term spectral features. These features were introduced nearly 30 years ago for speech recognition [50]. Despite there is a strong belief that temporal, prosodic and high level features are salient speaker cues, we have not benefited much from them. So far, they are playing a secondary role complementary to short-term spectral features. This warrants further investigation, especially as to how temporal and prosodic features can capture high-level phenomena (robust) without using computationally intensive speech recognizer (practical). It remains a great challenge in the near future to understand what features to exactly look for in speech signal.

9. Summary

We have presented an overview of the classical and new methods of automatic text-independent speaker recognition. The recognition accuracy of current speaker recognition systems under controlled conditions is high. However, in practical situations many negative factors are encountered including mismatched handsets for training and testing, limited training data, unbalanced text, background noise and non-cooperative users. The techniques of robust feature extraction, feature normalization, model-domain compensation and score normalization methods are necessary. The technology advancement as represented by NIST evaluations in the recent years has addressed several technical challenges such as text/language dependency, channel effects, speech durations, and cross-talk speech. However, many research problems remain to be addressed, such as human-related error sources, real-time implementation, and forensic interpretation of speaker recognition scores.

Acknowledgements

The authors would like to thank Ms. Sharifah Mahani Aljunied for spell-checking an earlier version of the manuscript, and Dr. Kong-Aik Lee for providing insights into channel compensation of supervectors.

References

- [1] ADAMI, A. Modeling prosodic differences for speaker recognition. *Speech Communication* 49, 4 (April 2007), 277–291.
- [2] ADAMI, A., MIHAESCU, R., REYNOLDS, D., AND GODFREY, J. Modeling prosodic dynamics for speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)* (Hong Kong, China, April 2003), pp. 788–791.
- [3] ALEXANDER, A., BOTTI, F., DESSIMOZ, D., AND DRYGAJLO, A. The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic Science International* 146S (December 2004), 95–99.
- [4] ALKU, P., TIITINEN, H., AND NÄÄTÄNEN, R. A method for generating natural-sounding speech stimuli for cognitive brain research. *Clinical Neurophysiology* 110, 8 (1999), 1329–1333.
- [5] ALTINCAY, H., AND DEMIREKLER, M. Speaker identification by combining multiple classifiers using dempster-shafer theory of evidence. *Speech Communication* 41, 4 (November 2003), 531–547.
- [6] AMBIKAIKAJAH, E. Emerging features for speaker recognition. In *Proc. 6th International IEEE Conference on Information, Communications & Signal Processing* (Singapore, December 2007), pp. 1–7.
- [7] ANDREWS, W., KOHLER, M., AND CAMPBELL, J. Phonetic speaker recognition. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (Aalborg, Denmark, September 2001), pp. 2517–2520.
- [8] ANDREWS, W., KOHLER, M., CAMPBELL, J., GODFREY, J., AND HERNANDEZ-CORDERO, J. Gender-dependent phonetic refraction for speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002)* (Orlando, Florida, USA, May 2002), vol. 1, pp. 149–152.
- [9] ARCIENEGA, M., AND DRYGAJLO, A. Pitch-dependent GMMs for text-independent speaker recognition systems. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (Aalborg, Denmark, September 2001), pp. 2821–2824.
- [10] ASHOUR, G., AND GATH, I. Characterization of speech during imitation. In *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)* (Budapest, Hungary, September 1999), pp. 1187–1190.
- [11] ATAL, B. Automatic speaker recognition based on pitch contours. *Journal of the Acoustic Society of America* 52, 6 (1972), 1687–1697.
- [12] ATAL, B. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustic Society of America* 55, 6 (1974), 1304–1312.
- [13] ATLAS, L., AND SHAMMA, S. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing* 7 (2003), 668–675.
- [14] AUCKENTHALER, R., CAREY, M., AND LLOYD-THOMAS, H. Score normalization for text-independent speaker verification systems. *Digital Signal Processing* 10, 1-3 (January 2000), 42–54.
- [15] AUCKENTHALER, R., AND MASON, J. Gaussian selection applied to text-independent speaker verification. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)* (Crete, Greece, June 2001), pp. 83–88.
- [16] BARTKOVA, K., D.L.GAC, CHARLET, D., AND JOUVET, D. Prosodic parameter for speaker identification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)* (Denver, Colorado, USA, September 2002), pp. 1197–1200.
- [17] BENYASSINE, A., SCHLOMOT, E., AND SU, H. ITU-T recommendation g729 annex b: A silence compression scheme for use with g729 optimized for v.70 digital simultaneous voice and data applications. *IEEE Communications Magazine* 35 (1997), 64–73.
- [18] BENZEGHIBA, M., AND BOURLAND, H. On the combination of speech and speaker recognition. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)* (Geneva, Switzerland, September 2003), pp. 1361–1364.

- [19] BENZEGHIBA, M., AND BOURLAND, H. User-customized password speaker verification using multiple reference and background models. *Speech Communication* 48, 9 (September 2006), 1200–1213.
- [20] BESACIER, L., BONASTRE, J., AND FREDOUILLE, C. Localization and selection of speaker-specific information with statistical modeling. *Speech Communication* 31 (June 2000), 89–106.
- [21] BESACIER, L., AND BONASTRE, J.-F. Subband architecture for automatic speaker recognition. *Signal Processing* 80 (July 2000), 1245–1259.
- [22] BIMBOT, F., BONASTRE, J.-F., FREDOUILLE, C., GRAVIER, G., MAGRIN-CHAGNOLLEAU, I., MEIGNIER, S., MERLIN, T., ORTEGA-GARCIA, J., PETROVSKA-DELACRETAZ, D., AND REYNOLDS, D. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing* 2004, 4 (2004), 430–451.
- [23] BIMBOT, F., MAGRIN-CHAGNOLLEAU, I., AND MATHAN, L. Second-order statistical measures for text-independent speaker identification. *Speech Communication* 17 (August 1995), 177–192.
- [24] BISHOP, C. *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, New York, 2006.
- [25] BOCKLET, T., AND SHRIBERG, E. Speaker recognition using syllable-based constraints for cepstral frame selection. In *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)* (Taipei, Taiwan, April 2009), pp. 4525–4528.
- [26] BOERSMA, P., AND WEENINK, D. Praat: doing phonetics by computer [computer program]. WWW page, June 2009. <http://www.praat.org/>.
- [27] BONASTRE, J.-F., MATROUF, D., AND FREDOUILLE, C. Artificial impostor voice transformation effects on false acceptance rates. In *Proc. Interspeech 2007 (ICSLP)* (Antwerp, Belgium, August 2007), pp. 2053–2056.
- [28] BRÜMMER, N., BURGET, L., ČERNOCKÝ, J., GLEMBEK, O., GRÉZL, F., KARAFIÁT, M., LEEUWEN, D., MATĚJKA, P., SCHWARTZ, P., AND STRASHEIM, A. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Trans. Audio, Speech and Language Processing* 15, 7 (September 2007), 2072–2084.
- [29] BRÜMMER, N., AND PREEZ, J. Application-independent evaluation of speaker detection. *Computer Speech and Language* 20 (April–July 2006), 230–275.
- [30] BURGET, L., BRÜMMER, N., REYNOLDS, D., KENNY, P., PELECANOS, J., VOGT, R., CASTALDO, F., DEHAK, N., DEHAK, R., GLEMBEK, O., KARAM, Z., NOECKER, J., NA, E., COSTIN, C., HUBEIKA, V., KAJAREKAR, S., SCHEFFER, N., AND ČERNOCKÝ, J. Robust speaker recognition over varying channels - report from JHU workshop 2008. Technical report, March 2009. http://www.cslsp.jhu.edu/workshops/ws08/documents/jhu_report_main.pdf (URL valid June 2009).
- [31] BURGET, L., MATĚJKA, P., SCHWARZ, P., GLEMBEK, O., AND ČERNOCKÝ, J. Analysis of feature extraction and channel compensation in a GMM speaker recognition system. *IEEE Trans. Audio, Speech and Language Processing* 15, 7 (September 2007), 1979–1986.
- [32] BURTON, D. Text-dependent speaker verification using vector quantization source coding. *IEEE Trans. Acoustics, Speech, and Signal Processing* 35, 2 (February 1987), 133–143.
- [33] CAMPBELL, J. Speaker recognition: a tutorial. *Proceedings of the IEEE* 85, 9 (September 1997), 1437–1462.
- [34] CAMPBELL, W., ASSALEH, K., AND BROUN, C. Speaker recognition with polynomial classifiers. *IEEE Trans. on Speech and Audio Processing* 10, 4 (May 2002), 205–212.
- [35] CAMPBELL, W., CAMPBELL, J., REYNOLDS, D., JONES, D., AND LEEK, T. Phonetic speaker recognition with support vector machines. In *Advances in Neural Information Processing Systems* 16, S. Thrun, L. Saul, and B. Schölkopf, Eds. MIT Press, Cambridge, MA, 2004.
- [36] CAMPBELL, W., CAMPBELL, J., REYNOLDS, D., SINGER, E., AND TORRES-CARRASQUILLO, P. Support vector machines for speaker and language recognition. *Computer Speech and Language* 20, 2-3 (April 2006), 210–229.
- [37] CAMPBELL, W., STURIM, D., AND REYNOLDS, D. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)* (Philadelphia, USA, March 2005), pp. 637–640.
- [38] CAMPBELL, W., STURIM, D., AND REYNOLDS, D. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters* 13, 5 (May 2006), 308–311.
- [39] CAREY, M., PARRIS, E., LLOYD-THOMAS, H., AND BENNETT, S. Robust prosodic features for speaker identification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1996)* (Philadelphia, Pennsylvania, USA, 1996), pp. 1800–1803.
- [40] CASTALDO, F., COLIBRO, D., DALMASSO, E., LAFACE, P., AND VAIR, C. Compensation of nuisance factors for speaker and language recognition. *IEEE Trans. Audio, Speech and Language Processing* 15, 7 (September 2007), 1969–1978.
- [41] CASTALDO, F., COLIBRO, D., DALMASSO, E., LAFACE, P., AND VAIR, C. Compensation of nuisance factors for speaker and language recognition. *IEEE Trans. Audio, Speech and Language Processing* 15, 7 (September 2007), 1969–1978.
- [42] CHAN, W., ZHENG, N., AND LEE, T. Discrimination power of vocal source and vocal tract related features for speaker segmentation. *IEEE Trans. Audio, Speech and Language Processing* 15, 6 (August 2007), 1884–1892.
- [43] CHARBUILLET, C., GAS, B., CHETOUANI, M., AND ZARADER, J. Filter bank design for speaker diarization based on genetic algorithms. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006)* (Toulouse, France, May 2006), vol. 1, pp. 673–676.
- [44] CHAUDHARI, U., NAVRATIL, J., AND MAES, S. Multigrained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition. *IEEE Trans. on Speech and Audio Processing* 11, 1 (January 2003), 61–69.
- [45] CHEN, K., WANG, L., AND CHI, H. Methods of combining multiple classifiers with different features and their applications to text-independent speaker recognition. *International Journal of Pattern Recognition and Artificial Intelligence* 11, 3 (1997), 417–445.
- [46] CHEN, Z.-H., LIAO, Y.-F., AND JUANG, Y.-T. Eigen-prosody analysis for robust speaker recognition under mismatch handset environment. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2004)* (Jeju, South Korea, October 2004), pp. 1421–1424.
- [47] CHETOUANI, M., FAUNDEZ-ZANUY, M., GAS, B., AND ZARADER, J. Investigation on LP-residual presentations for speaker identification. *Pattern Recognition* 42, 3 (March 2009), 487–494.
- [48] CHEVEIGNÉ, A., AND KAWAHARA, H. Comparative evaluation of f_0 estimation algorithms. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (Aalborg, Denmark, September 2001), pp. 2451–2454.
- [49] DAMPER, R., AND HIGGINS, J. Improving speaker identification in noise by subband processing and decision fusion. *Pattern Recognition Letters* 24 (September 2003), 2167–2173.
- [50] DAVIS, S., AND MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, and Signal Processing* 28, 4 (August 1980), 357–366.
- [51] DECHEVEIGNE, A., AND KAWAHARA, H. YIN, a fundamental frequency estimator for speech and music. *The Journal of Acous-*

- tical Society of America 111*, 4 (April 2002), 1917–1930.
- [52] DEHAK, N., AND CHOLLET, G. Support vector GMMs for speaker verification. In *Proc. IEEE Odyssey: the Speaker and Language Recognition Workshop (Odyssey 2006)* (San Juan, Puerto Rico, June 2006).
- [53] DEHAK, N., DEHAK, R., KENNY, P., AND DUMOUCHEL, P. Comparison between factor analysis and GMM support vector machines for speaker verification. In *The Speaker and Language Recognition Workshop (Odyssey 2008)* (Stellenbosch, South Africa, January 2008). Paper 009.
- [54] DEHAK, N., KENNY, P., DEHAK, R., GLEMBEK, O., DUMOUCHEL, P., BURGET, L., HUBEIKA, V., AND CASTALDO, F. Support vector machines and joint factor analysis for speaker verification. In *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)* (Taipei, Taiwan, April 2009), pp. 4237–4240.
- [55] DEHAK, N., KENNY, P., AND DUMOUCHEL, P. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Trans. Audio, Speech and Language Processing* 15, 7 (September 2007), 2095–2103.
- [56] DELLER, J., HANSEN, J., AND PROAKIS, J. *Discrete-Time Processing of Speech Signals*, second ed. IEEE Press, New York, 2000.
- [57] DODDINGTON, G. Speaker recognition based on idiolectal differences between speakers. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (Aalborg, Denmark, September 2001), pp. 2521–2524.
- [58] DUNN, R., QUATIERI, T., REYNOLDS, D., AND CAMPBELL, J. Speaker recognition from coded speech and the effects of score normalization. In *Proc. 35th Asilomar Conference on Signals, Systems and Computers* (Pacific Grove, California, USA, November 2001), vol. 2, pp. 1562–1567.
- [59] ESPY-WILSON, C., MANOCHA, S., AND VISHNUHOTLA, S. A new set of features for text-independent speaker identification. In *Proc. Interspeech 2006 (ICSLP)* (Pittsburgh, Pennsylvania, USA, September 2006), pp. 1475–1478.
- [60] EZZAIDI, H., ROUAT, J., AND O’SHAUGHNESSY, D. Towards combining pitch and MFCC for speaker identification systems. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (Aalborg, Denmark, September 2001), pp. 2825–2828.
- [61] FALTLHAUSER, R., AND RUSKE, G. Improving speaker recognition performance using phonetically structured gaussian mixture models. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)* (Aalborg, Denmark, September 2001), pp. 751–754.
- [62] FARRELL, K., MAMMONE, R., AND ASSALEH, K. Speaker recognition using neural networks and conventional classifiers. *IEEE Trans. on Speech and Audio Processing* 2, 1 (January 1994), 194–205.
- [63] FARRELL, K., RAMACHANDRAN, R., AND MAMMONE, R. An analysis of data fusion methods for speaker verification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1998)* (Seattle, Washington, USA, 1998), vol. 2, pp. 1129–1132.
- [64] FAUVE, B., EVANS, N., AND MASON, J. Improving the performance of text-independent short duration SVM- and GMM-based speaker verification. In *The Speaker and Language Recognition Workshop (Odyssey 2008)* (Stellenbosch, South Africa, January 2008). Paper 018.
- [65] FAUVE, B., MATROUF, D., SCHEFFER, N., BONASTRE, J.-F., AND MASON, J. State-of-the-art performance in text-independent speaker verification through open-source software. *IEEE Trans. Audio, Speech and Language Processing* 15, 7 (September 2007), 1960–1968.
- [66] FERRER, L., GRACIARENA, M., ZYMNIS, A., AND SHRIBERG, E. System combination using auxiliary information for speaker verification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008)* (Las Vegas, Nevada, March–April 2008), pp. 4853–4856.
- [67] FERRER, L., SHRIBERG, E., KAJAREKAR, S., AND SÖNMEZ, K. Parameterization of prosodic feature distributions for SVM modeling in speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007)* (Honolulu, Hawaii, USA, April 2007), vol. 4, pp. 233–236.
- [68] FERRER, L., SÖNMEZ, K., AND SHRIBERG, E. An anticorrelation kernel for improved system combination in speaker verification. In *The Speaker and Language Recognition Workshop (Odyssey 2008)* (Stellenbosch, South Africa, January 2008). Paper 022.
- [69] FREDOUILLE, C., BONASTRE, J.-F., AND MERLIN, T. AMIRAL: A block-segmental multirecognizer architecture for automatic speaker recognition. *Digital Signal Processing* 10, 1-3 (January 2000), 172–197.
- [70] FURUI, S. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing* 29, 2 (April 1981), 254–272.
- [71] FURUI, S. Recent advances in speaker recognition. *Pattern Recognition Letters* 18, 9 (September 1997), 859–872.
- [72] GARCIA-ROMERO, D., FIERREZ-AGUILAR, J., GONZALEZ-RODRIGUEZ, J., AND ORTEGA-GARCIA, J. On the use of quality measures for text-independent speaker recognition. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004)* (Toledo, Spain, May 2004), vol. 4, pp. 105–110.
- [73] GERSHO, A., AND GRAY, R. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1991.
- [74] GLEMBEK, O., BURGET, L., DEHAK, N., BRUMMER, N., AND KENNY, P. Comparison of scoring methods used in speaker recognition with joint factor analysis. In *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)* (Taipei, Taiwan, April 2009), pp. 4057–4060.
- [75] GONG, W.-G., YANG, L.-P., AND CHEN, D. Pitch synchronous based feature extraction for noise-robust speaker verification. In *Proc. Image and Signal Processing (CISP 2008)* (May 2008), vol. 5, pp. 295–298.
- [76] GONZALEZ-RODRIGUEZ, J., GARCIA-GOMAR, D. G.-R. M., RAMOS-CASTRO, D., AND ORTEGA-GARCIA, J. Robust likelihood ratio estimation in Bayesian forensic speaker recognition. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)* (Geneva, Switzerland, September 2003), pp. 693–696.
- [77] GOPALAN, K., ANDERSON, T., AND CUPPLES, E. A comparison of speaker identification results using features based on cepstrum and Fourier-Bessel expansion. *IEEE Trans. on Speech and Audio Processing* 7, 3 (May 1999), 289–294.
- [78] GUDNASON, J., AND BROOKES, M. Voice source cepstrum coefficients for speaker identification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008)* (Las Vegas, Nevada, March–April 2008), pp. 4821–4824.
- [79] GUPTA, S., AND SAVIC, M. Text-independent speaker verification based on broad phonetic segmentation of speech. *Digital Signal Processing* 2, 2 (April 1992), 69–79.
- [80] HANNANI, A., PETROVSKA-DELACRÉTAZ, D., AND CHOLLET, G. Linear and non-linear fusion of ALISP-based and GMM systems for text-independent speaker verification. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004)* (Toledo, Spain, May 2004), pp. 111–116.
- [81] HANSEN, E., SLYH, R., AND ANDERSON, T. Speaker recognition using phoneme-specific GMMs. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004)* (Toledo,

- Spain, May 2004), pp. 179–184.
- [82] HARRINGTON, J., AND CASSIDY, S. *Techniques in Speech Acoustics*. Kluwer Academic Publishers, Dordrecht, 1999.
- [83] HARRIS, F. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE* 66, 1 (January 1978), 51–84.
- [84] HATCH, A., KAJAREKAR, S., AND STOLCKE, A. Within-class covariance normalization for SVM-based speaker recognition. In *Proc. Interspeech 2006 (ICSLP)* (Pittsburgh, Pennsylvania, USA, September 2006), pp. 1471–1474.
- [85] HATCH, A., AND STOLCKE, A. Generalized linear kernels for one-versus-all classification: Application to speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006)* (Toulouse, France, May 2006), pp. 585–588.
- [86] HATCH, A., STOLCKE, A., AND PESKIN, B. Combining feature sets with support vector machines: Application to speaker recognition. In *The 2005 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (November 2005), pp. 75–79.
- [87] HAUTAMÄKI, V., KINNUNEN, T., AND FRÄNTI, P. Text-independent speaker recognition using graph matching. *Pattern Recognition Letters* 29, 9 (2008), 1427–1432.
- [88] HAUTAMÄKI, V., KINNUNEN, T., KÄRKKÄINEN, I., TUONONEN, M., SAASTAMOINEN, J., AND FRÄNTI, P. Maximum *a Posteriori* estimation of the centroid model for speaker verification. *IEEE Signal Processing Letters* 15 (2008), 162–165.
- [89] HAUTAMÄKI, V., TUONONEN, M., NIEMI-LAITINEN, T., AND FRÄNTI, P. Improving speaker verification by periodicity based voice activity detection. In *Proc. 12th International Conference on Speech and Computer (SPECOM 2007)* (Moscow, Russia, October 2007), pp. 645–650.
- [90] HE, J., LIU, L., AND PALM, G. A discriminative training algorithm for VQ-based speaker identification. *IEEE Trans. on Speech and Audio Processing* 7, 3 (May 1999), 353–356.
- [91] HÉBERT, M. Text-dependent speaker recognition. In *Springer handbook of speech processing* (Heidelberg, 2008), J. Benesty, M. Sondhi, and Y. Huang, Eds., Springer Verlag, pp. 743–762.
- [92] HÉBERT, M., AND HECK, L. Phonetic class-based speaker verification. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)* (Geneva, Switzerland, September 2003), pp. 1665–1668.
- [93] HECK, L., AND GENOUD, D. Combining speaker and speech recognition systems. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)* (Denver, Colorado, USA, September 2002), pp. 1369–1372.
- [94] HECK, L., KONIG, Y., SÖNMEZ, M., AND WEINTRAUB, M. Robustness to telephone handset distortion in speaker recognition by discriminative feature design. *Speech Communication* 31 (June 2000), 181–192.
- [95] HECK, L., AND WEINTRAUB, M. Handset-dependent background models for robust text-independent speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1997)* (Munich, Germany, April 1997), pp. 1071–1074.
- [96] HEDGE, R., MURTHY, H., AND RAO, G. Application of the modified group delay function to speaker identification and discrimination. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2004)* (Montreal, Canada, May 2004), vol. 1, pp. 517–520.
- [97] HERMANSKY, H. Perceptual linear prediction (PLP) analysis for speech. *Journal of the Acoustic Society of America* 87 (1990), 1738–1752.
- [98] HERMANSKY, H. Should recognizers have ears? *Speech Communication* 25, 1-3 (August 1998), 3–27.
- [99] HERMANSKY, H., AND MORGAN, N. RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing* 2, 4 (October 1994), 578–589.
- [100] HESS, W. *Pitch determination of speech signals: algorithms and devices*. Springer Verlag, Berlin, 1983.
- [101] HIGGINS, A., BAHLER, L., AND PORTER, J. Speaker verification using randomized phrase prompting. *Digital Signal Processing* 1 (April 1991), 89–106.
- [102] HUANG, X., ACERO, A., AND HON, H.-W. *Spoken Language Processing: a Guide to Theory, Algorithm, and System Development*. Prentice-Hall, New Jersey, 2001.
- [103] IMPERL, B., KACIC, Z., AND HORVAT, B. A study of harmonic features for the speaker recognition. *Speech Communication* 22, 4 (September 1997), 385–402.
- [104] JAIN, A., DUIN, R., AND MAO, J. Statistical pattern recognition: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22, 1 (January 2000), 4–37.
- [105] JANG, G.-J., LEE, T.-W., AND OH, Y.-H. Learning statistically efficient features for speaker recognition. *Neurocomputing* 49 (December 2002), 329–348.
- [106] JIN, Q., SCHULTZ, T., AND WAIBEL, A. Speaker identification using multilingual phone strings. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002)* (Orlando, Florida, USA, May 2002), vol. 1, pp. 145–148.
- [107] KAJAREKAR, S., AND HERMANSKY, H. Speaker verification based on broad phonetic categories. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)* (Crete, Greece, June 2001), pp. 201–206.
- [108] KARAM, Z., AND CAMPBELL, W. A new kernel for SVM MLLR based speaker recognition. In *Proc. Interspeech 2007 (ICSLP)* (Antwerp, Belgium, August 2007), pp. 290–293.
- [109] KARPOV, E., KINNUNEN, T., AND FRÄNTI, P. Symmetric distortion measure for speaker recognition. In *Proc. 9th Int. Conf. Speech and Computer (SPECOM 2004)* (St. Petersburg, Russia, September 2004), pp. 366–370.
- [110] KENNY, P. Joint factor analysis of speaker and session variability: theory and algorithms. technical report CRIM-06/08-14, 2006.
- [111] KENNY, P., BOULIANNE, G., OUELLET, P., AND DUMOUCHEL, P. Speaker and session variability in GMM-based speaker verification. *IEEE Trans. Audio, Speech and Language Processing* 15, 4 (May 2007), 1448–1460.
- [112] KENNY, P., OUELLET, P., DEHAK, N., GUPTA, V., AND DUMOUCHEL, P. A study of inter-speaker variability in speaker verification. *IEEE Trans. Audio, Speech and Language Processing* 16, 5 (July 2008), 980–988.
- [113] KINNUNEN, T. Designing a speaker-discriminative adaptive filter bank for speaker recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)* (Denver, Colorado, USA, September 2002), pp. 2325–2328.
- [114] KINNUNEN, T. *Spectral Features for Automatic Text-Independent Speaker Recognition*. Licentiate’s thesis, University of Joensuu, Department of Computer Science, Joensuu, Finland, 2004.
- [115] KINNUNEN, T. Joint acoustic-modulation frequency for speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006)* (Toulouse, France, 2006), vol. I, pp. 665–668.
- [116] KINNUNEN, T., AND ALKU, P. On separating glottal source and vocal tract information in telephony speaker verification. In *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)* (Taipei, Taiwan, April 2009), pp. 4545–4548.
- [117] KINNUNEN, T., AND GONZÁLEZ-HAUTAMÄKI, R. Long-term f_0 modeling for text-independent speaker recognition. In *Proc. 10th International Conference Speech and Computer*

- (SPECOM'2005) (Patras, Greece, October 2005), pp. 567–570.
- [118] KINNUNEN, T., HAUTAMÄKI, V., AND FRÄNTI, P. Fusion of spectral feature sets for accurate speaker identification. In *Proc. 9th Int. Conf. Speech and Computer (SPECOM 2004)* (St. Petersburg, Russia, September 2004), pp. 361–365.
- [119] KINNUNEN, T., HAUTAMÄKI, V., AND FRÄNTI, P. On the use of long-term average spectrum in automatic speaker recognition. In *5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP'06)* (Singapore, December 2006), pp. 559–567.
- [120] KINNUNEN, T., KARPOV, E., AND FRÄNTI, P. Real-time speaker identification and verification. *IEEE Trans. Audio, Speech and Language Processing* 14, 1 (January 2006), 277–288.
- [121] KINNUNEN, T., KILPILÄINEN, T., AND FRÄNTI, P. Comparison of clustering algorithms in speaker identification. In *Proc. IASTED Int. Conf. Signal Processing and Communications (SPC 2000)* (Marbella, Spain, September 2000), pp. 222–227.
- [122] KINNUNEN, T., KOH, C., WANG, L., LI, H., AND CHNG, E. Temporal discrete cosine transform: Towards longer term temporal features for speaker verification. In *Proc. 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP 2006)* (Singapore, December 2006), pp. 547–558.
- [123] KINNUNEN, T., LEE, K.-A., AND LI, H. Dimension reduction of the modulation spectrogram for speaker verification. In *The Speaker and Language Recognition Workshop (Odyssey 2008)* (Stellenbosch, South Africa, January 2008).
- [124] KINNUNEN, T., SAASTAMOINEN, J., HAUTAMÄKI, V., VINNI, M., AND FRÄNTI, P. Comparative evaluation of maximum *a Posteriori* vector quantization and Gaussian mixture models in speaker verification. *Pattern Recognition Letters* 30, 4 (March 2009), 341–347.
- [125] KINNUNEN, T., ZHANG, B., ZHU, J., AND WANG, Y. Speaker verification with adaptive spectral subband centroids. In *Proc. International Conference on Biometrics (ICB 2007)* (Seoul, Korea, August 2007), pp. 58–66.
- [126] KITAMURA, T. Acoustic analysis of imitated voice produced by a professional impersonator. In *Proc. Interspeech 2008* (September 2008), pp. 813–816.
- [127] KITTLER, J., HATEF, M., DUIN, R., AND MATAS, J. On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 20, 3 (March 1998), 226–239.
- [128] KOLANO, G., AND REGEL-BRIETZMANN, P. Combination of vector quantization and Gaussian mixture models for speaker verification. In *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)* (Budapest, Hungary, September 1999), pp. 1203–1206.
- [129] KRYSZCZUK, K., RICHARDI, J., PRODANOV, P., AND DRYGAJLO, A. Reliability-based decision fusion in multimodal biometric verification systems. *EURASIP Journal of Advances in Signal Processing*, 1 (2007), Article ID 86572.
- [130] LAPIDOT, I., GUTERMAN, H., AND COHEN, A. Unsupervised speaker recognition based on competition between self-organizing maps. *IEEE Transactions on Neural Networks* 13 (July 2002), 877–887.
- [131] LASKOWSKI, K., AND JIN, Q. Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum. In *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)* (Taipei, Taiwan, April 2009), pp. 4541–4544.
- [132] LEE, K., YOU, C., LI, H., KINNUNEN, T., AND ZHU, D. Characterizing speech utterances for speaker verification with sequence kernel SVM. In *Proc. 9th Interspeech (Interspeech 2008)* (Brisbane, Australia, September 2008), pp. 1397–1400.
- [133] LEE, K.-A., YOU, C., LI, H., AND KINNUNEN, T. A GMM-based probabilistic sequence kernel for speaker verification. In *Proc. Interspeech 2007 (ICSLP)* (Antwerp, Belgium, August 2007), pp. 294–297.
- [134] LEEUWEN, D., MARTIN, A., PRZYBOCKI, M., AND BOUTEN, J. NIST and NFI-TNO evaluations of automatic speaker recognition. *Computer Speech and Language* 20 (April-July 2006), 128–158.
- [135] LEGGETTER, C., AND WOODLAND, P. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language* 9 (1995), 171–185.
- [136] LEI, H., AND MIRGHAFORI, N. Word-conditioned HMM super-vectors for speaker recognition. In *Proc. Interspeech 2007 (ICSLP)* (Antwerp, Belgium, August 2007), pp. 746–749.
- [137] LEUNG, K., MAK, M., SU, M., AND KUNG, S. Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification. *Speech Communication* 48, 1 (January 2006), 71–84.
- [138] LI, H., MA, B., LEE, K.-A., SUN, H., ZHU, D., SIM, K., YOU, C., TONG, R., KÄRKKÄINEN, I., HUANG, C.-L., PERVOUCHINE, V., GUO, W., LI, Y., DAI, L., NOSRATIGHODS, M., THARMARAJAH, T., EPPS, J., AMBIKAIKAJAH, E., CHNG, E.-S., SCHULTZ, T., AND JIN, Q. The I4U system in NIST 2008 speaker recognition evaluation. In *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)* (Taipei, Taiwan, April 2009), pp. 4201–4204.
- [139] LI, K.-P., AND PORTER, J. Normalizations and selection of speech segments for speaker recognition scoring. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1988)* (New York, USA, April 1988), pp. 595–598.
- [140] LINDE, Y., BUZO, A., AND GRAY, R. An algorithm for vector quantizer design. *IEEE Transactions on Communications* 28, 1 (January 1980), 84–95.
- [141] LONGWORTH, C., AND GALES, M. Combining derivative and parametric kernels for speaker verification. *IEEE Trans. Audio, Speech and Language Processing* 6, 1 (January 2007), 1–10.
- [142] LOURADOUR, J., AND DAUDI, K. SVM speaker verification using a new sequence kernel. In *Proc. 13th European Conf. on Signal Processing (EUSIPCO 2005)* (Antalya, Turkey, September 2005).
- [143] LOURADOUR, J., DAUDI, K., AND ANDRÉ-OBRECHT, R. Discriminative power of transient frames in speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)* (Philadelphia, USA, 2005), vol. 1, pp. 613–616.
- [144] LU, X., AND DANG, J. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech Communication* 50, 4 (April 2007), 312–322.
- [145] MA, B., LI, H., AND TONG, R. Spoken language recognition with ensemble classifiers. *IEEE Trans. Audio, Speech and Language Processing* 15, 7 (September 2007), 2053–2062.
- [146] MA, B., ZHU, D., AND TONG, R. Chinese dialect identification using tone features based on pitch flux. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006)* (Toulouse, France, May 2006), vol. 1, pp. 1029–1032.
- [147] MA, B., ZHU, D., TONG, R., AND LI, H. Speaker cluster based GMM tokenization for speaker recognition. In *Proc. Interspeech 2006 (ICSLP)* (Pittsburgh, Pennsylvania, USA, September 2006), pp. 505–508.
- [148] MAGRIN-CHAGNOLLEAU, I., DUROU, G., AND BIMBOT, F. Application of time-frequency principal component analysis to text-independent speaker identification. *IEEE Trans. on Speech and Audio Processing* 10, 6 (September 2002), 371–378.
- [149] MAK, M.-W., CHEUNG, M., AND KUNG, S. Robust speaker verification from GSM-transcoded speech based on decision fusion and feature transformation. In *Proc. Int. Conf. on Acoustics,*

- Speech, and Signal Processing (ICASSP 2003)* (Hong Kong, China, April 2003), vol. 2, pp. 745–748.
- [150] MAK, M.-W., HSIAO, R., AND MAK, B. A comparison of various adaptation methods for speaker verification with limited enrollment data. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2006)* (Toulouse, France, May 2006), vol. 1, pp. 929–932.
- [151] MAK, M.-W., AND TSANG, C.-L. Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification. *EURASIP Journal on Applied Signal Processing 4* (January 2004), 452–465.
- [152] MAKHOUL, J. Linear prediction: a tutorial review. *Proceedings of the IEEE 64*, 4 (April 1975), 561–580.
- [153] MALAYATH, N., HERMAN, H., KAJAREKAR, S., AND YEGNANARAYANA, B. Data-driven temporal filters and alternatives to GMM in speaker verification. *Digital Signal Processing 10*, 1-3 (January 2000), 55–74.
- [154] MAMI, Y., AND CHARLET, D. Speaker recognition by location in the space of reference speakers. *Speech Communication 48*, 2 (February 2006), 127–411.
- [155] MAMMONE, R., ZHANG, X., AND RAMACHANDRAN, R. Robust speaker recognition: a feature based approach. *IEEE Signal Processing Magazine 13*, 5 (September 1996), 58–71.
- [156] M.A.PRZYBOCKI, MARTIN, A., AND LE, A. NIST speaker recognition evaluations utilizing the mixer corpora – 2004, 2005, 2006. *IEEE Trans. Audio, Speech and Language Processing 15*, 7 (September 2007), 1951–1959.
- [157] MARIÉTHOZ, J., AND BENGIO, S. A comparative study of adaptation methods for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)* (Denver, Colorado, USA, September 2002), pp. 581–584.
- [158] MARKEL, J., OSHIKA, B., AND A.H. GRAY, J. Long-term feature averaging for speaker recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing 25*, 4 (August 1977), 330–337.
- [159] MARTIN, A., DODDINGTON, G., KAMM, T., ORDOWSKI, M., AND PRZYBOCKI, M. The DET curve in assessment of detection task performance. In *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)* (Rhodos, Greece, September 1997), pp. 1895–1898.
- [160] MARY, L., AND YEGNANARAYANA, B. Prosodic features for speaker verification. In *Proc. Interspeech 2006 (ICSLP)* (Pittsburgh, Pennsylvania, USA, September 2006), pp. 917–920.
- [161] MARY, L., AND YEGNANARAYANA, B. Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication 50*, 10 (2008), 782–796.
- [162] MASON, M., VOGT, R., BAKER, B., AND SRIDHARAN, S. Data-driven clustering for blind feature mapping in speaker verification. In *Proc. Interspeech 2005* (Lisboa, Portugal, September 2005), pp. 3109–3112.
- [163] McLAUGHLIN, J., REYNOLDS, D., AND GLEASON, T. A study of computation speed-ups of the GMM-UBM speaker recognition system. In *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999)* (Budapest, Hungary, September 1999), pp. 1215–1218.
- [164] MISRA, H., IKBAL, S., AND YEGNANARAYANA, B. Speaker-specific mapping for text-independent speaker recognition. *Speech Communication 39*, 3-4 (February 2003), 301–310.
- [165] MIYAJIMA, C., WATANABE, H., TOKUDA, K., KITAMURA, T., AND KATAGIRI, S. A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction. *Speech Communication 35* (October 2001), 203–218.
- [166] MOONASAR, V., AND VENAYAGAMOORTHY, G. A committee of neural networks for automatic speaker recognition (ASR) systems. In *Proc. Int. Joint Conference on Neural Networks (IJCNN 2001)* (Washington, DC, USA, July 2001), pp. 2936–2940.
- [167] MÜLLER, C., Ed. *Speaker Classification I: Fundamentals, Features, and Methods* (2007), vol. 4343 of *Lecture Notes in Computer Science*, Springer.
- [168] MÜLLER, C., Ed. *Speaker Classification II, Selected Projects* (2007), vol. 4441 of *Lecture Notes in Computer Science*, Springer.
- [169] MÜLLER, K.-R., MIKA, S., RÄTSCH, G., TSUDA, K., AND SCHÖLKOPF, B. An introduction to kernel-based learning algorithms. *IEEE Trans. on Neural Networks 12*, 2 (May 2001), 181–201.
- [170] MURTY, K., AND YEGNANARAYANA, B. Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Processing Letters 13*, 1 (January 2006), 52–55.
- [171] NAIK, J., NETSCH, L., AND DODDINGTON, G. Speaker verification over long distance telephone lines. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1989)* (Glasgow, May 1989), pp. 524–527.
- [172] NAKASONE, H., MIMIKOPOULOS, M., BECK, S., AND MATHUR, S. Pitch synchronized speech processing (PSSP) for speaker recognition. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004)* (Toledo, Spain, May 2004), pp. 251–256.
- [173] NEY, H., MARTIN, S., AND WESSEL, F. Statistical language modeling using leaving-one-out. In *Corpus-based Methods in Language and Speech Processing* (1997), S. Young and G. Bloothoof, Eds., Kluwer Academic Publishers, pp. 174–207.
- [174] NIEMI-LAITINEN, T., SAASTAMOINEN, J., KINNUNEN, T., AND FRÄNTI, P. Applying MFCC-based automatic speaker recognition to GSM and forensic data. In *Proc. Second Baltic Conference on Human Language Technologies (HLT'2005)* (Tallinn, Estonia, April 2005), pp. 317–322.
- [175] NIST 2008 SRE results page, September 2008. http://www.nist.gov/speech/tests/sre/2008/official_results/index.html.
- [176] NOLAN, F. *The Phonetic Bases of Speaker Recognition*. Cambridge University Press, Cambridge, 1983.
- [177] OPPENHEIM, A., SCHAFER, R., AND BUCK, J. *Discrete-Time Signal Processing*, second ed. Prentice Hall, 1999.
- [178] ORMAN, D., AND ARSLAN, L. Frequency analysis of speaker identification. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)* (Crete, Greece, June 2001), pp. 219–222.
- [179] PALIWAL, K., AND ALSTERIS, L. Usefulness of phase spectrum in human speech perception. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)* (Geneva, Switzerland, September 2003), pp. 2117–2120.
- [180] PARK, A., AND HAZEN, T. ASR dependent techniques for speaker identification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2002)* (Denver, Colorado, USA, September 2002), pp. 1337–1340.
- [181] PELECANOS, J., MYERS, S., SRIDHARAN, S., AND CHANDRAN, V. Vector quantization based Gaussian modeling for speaker verification. In *Proc. Int. Conf. on Pattern Recognition (ICPR 2000)* (Barcelona, Spain, September 2000), pp. 3298–3301.
- [182] PELECANOS, J., AND SRIDHARAN, S. Feature warping for robust speaker verification. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)* (Crete, Greece, June 2001), pp. 213–218.
- [183] PELLOM, B., AND HANSEN, J. An efficient scoring algorithm for gaussian mixture model based speaker identification. *IEEE Signal Processing Letters 5*, 11 (1998), 281–284.
- [184] PELLOM, B., AND HANSEN, J. An experimental study of speaker verification sensitivity to computer voice-altered imposters. pp. 837–840.

- [185] PFISTER, B., AND BEUTLER, R. Estimating the weight of evidence in forensic speaker verification. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)* (Geneva, Switzerland, September 2003), pp. 701–704.
- [186] PLUMPE, M., QUATIERI, T., AND REYNOLDS, D. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing* 7, 5 (September 1999), 569–586.
- [187] POH, N., AND BENGIO, S. Why do multi-stream, multi-band and multi-modal approaches work on biometric user authentication tasks? In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2004)* (Montreal, Canada, May 2004), vol. 5, pp. 893–896.
- [188] PRASANNA, S., GUPTA, C., AND YEGNANARAYANA, B. Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication* 48 (2006), 1243–1261.
- [189] RABINER, L., AND JUANG, B.-H. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, New Jersey, 1993.
- [190] RAMACHANDRAN, R., FARRRELL, K., RAMACHANDRAN, R., AND MAMMONE, R. Speaker recognition - general classifier approaches and data fusion methods. *Pattern Recognition* 35 (December 2002), 2801–2821.
- [191] RAMIREZ, J., SEGURA, J., BENÍTEZ, C., DE LA TORRE, A., AND RUBIO, A. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication* 42, 3–4 (April 2004), 271–287.
- [192] RAMOS-CASTRO, D., FIERREZ-AGUILAR, J., GONZALEZ-RODRIGUEZ, J., AND ORTEGA-GARCIA, J. Speaker verification using speaker- and test-dependent fast score normalization. *Pattern Recognition Letters* 28, 1 (January 2007), 90–98.
- [193] REYNOLDS, D. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* 17 (August 1995), 91–108.
- [194] REYNOLDS, D. Channel robust speaker verification via feature mapping. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)* (Hong Kong, China, April 2003), vol. 2, pp. 53–56.
- [195] REYNOLDS, D., ANDREWS, W., CAMPBELL, J., NAVRATIL, J., PESHKIN, B., ADAMI, A., JIN, Q., KLUSACEK, D., ABRAMSON, J., MIHAESCU, R., GODFREY, J., JONES, D., AND XIANG, B. The SuperSID project: exploiting high-level information for high-accuracy speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003)* (Hong Kong, China, April 2003), pp. 784–787.
- [196] REYNOLDS, D., CAMPBELL, W., GLEASON, T., QUILLEN, C., STURIM, D., TORRES-CARRASQUILLO, P., AND ADAMI, A. The 2004 MIT Lincoln laboratory speaker recognition system. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)* (Philadelphia, USA, 2005), vol. 1, pp. 177–180.
- [197] REYNOLDS, D., QUATIERI, T., AND DUNN, R. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing* 10, 1 (January 2000), 19–41.
- [198] REYNOLDS, D., AND ROSE, R. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing* 3 (January 1995), 72–83.
- [199] ROCH, M. Gaussian-selection-based non-optimal search for speaker identification. *Speech Communication* 48 (2006), 85–95.
- [200] RODRÍGUEZ-LIÑARES, L., GARCÍA-MATEO, C., AND ALBA-CASTRO, J. On combining classifiers for speaker authentication. *Pattern Recognition* 36, 2 (February 2003), 347–359.
- [201] ROSE, P. *Forensic Speaker Identification*. Taylor & Francis, London, 2002.
- [202] SAASTAMOINEN, J., KARPOV, E., HAUTAMÁKI, V., AND FRÄNTI, P. Accuracy of MFCC based speaker recognition in series 60 device. *EURASIP Journal on Applied Signal Processing* 17 (2005), 2816–2827.
- [203] SAEIDI, R., MOHAMMADI, H., GANCHEV, T., AND R.D.RODMAN. Particle swarm optimization for sorted adapted gaussian mixture models. *IEEE Trans. Audio, Speech and Language Processing* 17, 2 (February 2009), 344–353.
- [204] SHRIBERG, E., FERRER, L., KAJAREKAR, S., VENKATARAMAN, A., AND STOLCKE, A. Modeling prosodic feature sequences for speaker recognition. *Speech Communication* 46, 3-4 (July 2005), 455–472.
- [205] SIVAKUMARAN, P., ARIYAEENIA, A., AND LOOMES, M. Sub-band based text-dependent speaker verification. *Speech Communication* 41 (October 2003), 485–509.
- [206] SIVAKUMARAN, P., FORTUNA, J., AND ARIYAEENIA, A. Score normalization applied to open-set, text-independent speaker identification. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)* (Geneva, Switzerland, September 2003), pp. 2669–2672.
- [207] SŁOMKA, S., SRIDHARAN, S., AND CHANDRAN, V. A comparison of fusion techniques in mel-cepstral based speaker identification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)* (Sydney, Australia, November 1998), pp. 225–228.
- [208] SLYH, R., HANSEN, E., AND ANDERSON, T. Glottal modeling and closed-phase analysis for speaker recognition. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2004)* (Toledo, Spain, May 2004), pp. 315–322.
- [209] SÓNMEZ, K., SHRIBERG, E., HECK, L., AND WEINTRAUB, M. Modeling dynamic prosodic variation for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998)* (Sydney, Australia, November 1998), pp. 3189–3192.
- [210] SÓNMEZ, M., HECK, L., WEINTRAUB, M., AND SHRIBERG, E. A log-normal tied mixture model of pitch for prosody-based speaker recognition. In *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)* (Rhodos, Greece, September 1997), pp. 1391–1394.
- [211] SOLEWICZ, Y., AND KOPPEL, M. Using post-classifiers to enhance fusion of low- and high-level speaker recognition. *IEEE Trans. Audio, Speech and Language Processing* 15, 7 (September 2007), 2063–2071.
- [212] SOLOMONOFF, A., CAMPBELL, W., AND BOARDMAN, I. Advances in channel compensation for SVM speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)* (Philadelphia, USA, March 2005), pp. 629–632.
- [213] SOONG, F., A.E., A. R., JUANG, B.-H., AND RABINER, L. A vector quantization approach to speaker recognition. *AT & T Technical Journal* 66 (1987), 14–26.
- [214] SOONG, F., AND ROSENBERG, A. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing* 36, 6 (June 1988), 871–879.
- [215] STOLCKE, A., KAJAREKAR, S., AND FERRER, L. Nonparametric feature normalization for SVM-based speaker verification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008)* (Las Vegas, Nevada, April 2008), pp. 1577–1580.
- [216] STOLCKE, A., KAJAREKAR, S., FERRER, L., AND SHRIBERG, E. Speaker recognition with session variability normalization based on MLLR adaptation transforms. *IEEE Trans. Audio, Speech and Language Processing* 15, 7 (September 2007), 1987–1998.
- [217] STURIM, D., AND REYNOLDS, D. Speaker adaptive cohort selection for Tnorm in text-independent speaker verification. In

- Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2005)* (Philadelphia, USA, March 2005), vol. 1, pp. 741–744.
- [218] STURIM, D., REYNOLDS, D., SINGER, E., AND CAMPBELL, J. Speaker indexing in large audio databases using anchor models. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2001)* (Salt Lake City, Utah, USA, May 2001), vol. 1, pp. 429–432.
- [219] TEUNEN, R., SHAHSHAHANI, B., AND HECK, L. A model-based transformational approach to robust speaker recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP 2000)* (Beijing, China, October 2000), vol. 2, pp. 495–498.
- [220] THÉVENAZ, P., AND HÜGLI, H. Usefulness of the LPC-residue in text-independent speaker verification. *Speech Communication* 17, 1-2 (August 1995), 145–157.
- [221] THIAN, N., SANDERSON, C., AND BENGIO, S. Spectral subband centroids as complementary features for speaker authentication. In *Proc. First Int. Conf. Biometric Authentication (ICBA 2004)* (Hong Kong, China, July 2004), pp. 631–639.
- [222] THIRUVARAN, T., AMBIKAIKIRAJAH, E., AND EPPS, J. Extraction of FM components from speech signals using all-pole model. *Electronics Letters* 44, 6 (March 2008).
- [223] THIRUVARAN, T., AMBIKAIKIRAJAH, E., AND EPPS, J. FM features for automatic forensic speaker recognition. In *Proc. Interspeech 2008* (Brisbane, Australia, September 2008), pp. 1497–1500.
- [224] TONG, R., MA, B., LEE, K., YOU, C., ZHU, D., KINNUNEN, T., SUN, H., DONG, M., CHNG, E., AND LI, H. Fusion of acoustic and tokenization features for speaker recognition. In *5th Intl. Sym. on Chinese Spoken Language Processing (ISCSLP 2006)* (Singapore, December 2006), pp. 494–505.
- [225] TORRES-CARRASQUILLO, P., REYNOLDS, D., AND JR., J. D. Language identification using Gaussian mixture model tokenization. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002)* (Orlando, Florida, USA, May 2002), vol. 1, pp. 757–760.
- [226] TRANTER, S., AND D.A.REYNOLDS. An overview of automatic speaker diarization systems. *IEEE Trans. Audio, Speech and Language Processing* 14, 5 (September 2006), 1557–1565.
- [227] TYDLITAT, B., NAVRÁTIL, J., PELEGANOS, J., AND RAMASWAMY, G. Text-independent speaker verification in embedded environments. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007)* (Honolulu, Hawaii, April 2007), vol. 4, pp. 293–296.
- [228] VIIKKI, O., AND LAURILA, K. Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication* 25 (August 1998), 133–147.
- [229] VOGT, R., BAKER, B., AND SRIDHARAN, S. Modelling session variability in text-independent speaker verification. In *Proc. Interspeech 2005* (Lisboa, Portugal, September 2005), pp. 3117–3120.
- [230] VOGT, R., KAJAREKAR, S., AND SRIDHARAN, S. Discriminant NAP for SVM speaker recognition. In *The Speaker and Language Recognition Workshop (Odyssey 2008)* (Stellenbosch, South Africa, January 2008). Paper 010.
- [231] VOGT, R., AND SRIDHARAN, S. Explicit modeling of session variability for speaker verification. *Computer Speech and Language* 22, 1 (January 2008), 17–38.
- [232] WAN, V., AND RENALS, S. Speaker verification using sequence discriminant support vector machines. *IEEE Trans. on Speech and Audio Processing* 13, 2 (March 2005), 203–210.
- [233] WILDERMOTH, B., AND PALIWAL, K. Use of voicing and pitch information for speaker recognition. In *Proc. 8th Australian Intern. Conf. Speech Science and Technology* (Canberra, December 2000), pp. 324–328.
- [234] WOLF, J. Efficient acoustic parameters for speaker recognition. *Journal of the Acoustic Society of America* 51, 6 (Part 2) (1972), 2044–2056.
- [235] XIANG, B. Text-independent speaker verification with dynamic trajectory model. *IEEE Signal Processing Letters* 10 (May 2003), 141–143.
- [236] XIANG, B., AND BERGER, T. Efficient text-independent speaker verification with structural gaussian mixture models and neural network. *IEEE Trans. on Speech and Audio Processing* 11 (September 2003), 447–456.
- [237] XIANG, B., CHAUDHARI, U., NAVRÁTIL, J., RAMASWAMY, G., AND GOPINATH, R. Short-time Gaussianization for robust speaker verification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002)* (Orlando, Florida, USA, May 2002), vol. 1, pp. 681–684.
- [238] XIONG, Z., ZHENG, T., SONG, Z., SOONG, F., AND WU, W. A tree-based kernel selection approach to efficient Gaussian mixture model-universal background model based speaker identification. *Speech Communication* 48 (2006), 1273–1282.
- [239] YEGNANARAYANA, B., AND KISHORE, S. AANN: an alternative to GMM for pattern recognition. *Neural Networks* 15 (April 2002), 459–469.
- [240] YOU, C., LEE, K., AND LI, H. An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. *IEEE Signal Processing Letters* 16, 1 (January 2009), 49–52.
- [241] YUO, K.-H., AND WANG, H.-C. Joint estimation of feature transformation parameters and Gaussian mixture model for speaker identification. *Speech Communication* 28, 3 (July 1999), 227–241.
- [242] ZHENG, N., LEE, T., AND CHING, P. Integration of complementary acoustic features for speaker recognition. *IEEE Signal Processing Letters* 14, 3 (March 2007), 181–184.
- [243] ZHU, D., MA, B., AND LI, H. Using MAP estimation of feature transformation for speaker recognition. In *Proc. Interspeech 2008* (Brisbane, Australia, September 2008).
- [244] ZHU, D., MA, B., AND LI, H. Joint MAP adaptation of feature transformation and gaussian mixture model for speaker recognition. In *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)* (Taipei, Taiwan, April 2009), pp. 4045–4048.
- [245] ZHU, D., MA, B., LI, H., AND HUO, Q. A generalized feature transformation approach for channel robust speaker verification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2007)* (Honolulu, Hawaii, April 2007), vol. 4, pp. 61–64.
- [246] ZILCA, R. Text-independent speaker verification using utterance level scoring and covariance modeling. *IEEE Trans. on Speech and Audio Processing* 10, 6 (September 2002), 363–370.
- [247] ZILCA, R., KINGSBURY, B., NAVRÁTIL, J., AND RAMASWAMY, G. Pseudo pitch synchronous analysis of speech with applications to speaker recognition. *IEEE Trans. Audio, Speech and Language Processing* 14, 2 (March 2006), 467–478.
- [248] ZISSMAN, M. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Trans. on Speech and Audio Processing* 4, 1 (January 1996), 31–44.