Impact of variabilities on speech recognition

Mohamed Benzeghiba(1), Renato De Mori(2), Olivier Deroo(3), Stephane Dupont(4), Denis Jouvet(5), Luciano Fissore(6), Pietro Laface(7), Alfred Mertins(8) Christophe Ris(4), Richard Rose(9), Vivek Tyagi(1), Christian Wellekens(1)

1-Institut Eurecom, Sophia Antipolis, France
2-LIA, Avignon, France
3-A Capela, Mons, Belgium
4-Multitel, Mons, Belgium
5-France Telecom, Lannion, France
6-Loquendo, Torino, Italy
7-Politenico, Torino,Italy
8-University Carl von Ossietzky, Oldenburg, Germany
9-University Mc Gill, Montreal, Canada
christian.wellekens@eurecom.fr

Abstract

Major progress is being recorded regularly on both the technology and exploitation of Automatic Speech Recognition (ASR) and spoken language systems. However, there are still technological barriers to flexible solutions and user satisfaction under some circumstances. This is related to several factors, such as the sensitivity to the environment (background noise or channel variability), or the weak representation of grammatical and semantic knowledge.

Current research is also emphasizing deficiencies in dealing with variation naturally present in speech. For instance, the lack of robustness to foreign accents precludes the use by specific populations. There are actually many factors affecting the speech realization: regional, sociolinguistic, or related to the environment or the speaker itself. These create a wide range of variations that may not be modeled correctly (speaker, gender, speech rate, vocal effort, regional accents, speaking style, non stationarity...), especially when resources for system training are scarce.

This paper outlines some current advances related to variabilities in ASR.

1. Introduction

The weaknesses of ASR systems are pointed out even by non-experts: Is it possible to recognize speech in noisy environment? What happens if the speaker has a sore throat or is too stressed? These two questions put into evidence the main sources of variability in ASR. Extrinsic variabilities are due to the environment: signal to noise ratio may be high but also variable within short time, telecommunication channels (wired or wireless) show variable properties and just changing microphones may cause strong error rate. Speech signal not only conveys semantic information (the message) but also a lot of information about the speaker himself: gender, age, social and regional origin, health and emotional state and, with a rather strong reliability, its identity that are intrinsic variabilities.

Characterization of the effect of some of these specific variations, together with related techniques to improve ASR robustness is a major research topic.

As a first obvious theme, the speech signal is non-stationary. The power spectral density of speech varies over time according to the glottal signal (which for instance affect the pitch) and the configuration of the speech articulators (tongue, jaws, lips...). This signal is modeled, through Hidden Markov Models (HMMs), as a sequence of stationary random regimes. At a first stage of processing, most ASR processes analyze short signal frames (typically covering 30 ms of speech) on which stationarity is assumed. More subtle signal analysis techniques are being studied in the framework of ASR.

Compensation for noise degradation (additive noise)can be done at several levels: either by enhancing speech signal, or by training models on noisy databases, or by designing specific models for noise and speech, or by considering noise as missing information that can be marginalized in a statistical training of models by making hypotheses on the parametric distributions of noise and speech.

Known as convolution noise, degradations due to the channel come from its slowly varying spectral properties (or impulse response) that can be reduced by averaging speech features (Cepstral Mean Subtraction) or by evaluating the impulse response as missing data and combined with additive noise reduction.

Among intrinsic variabilities, modification of the speech production at the level of articulatory mechanisms under specific conditions plays a crucial role. Studies on the impact of coarticulation have yielded segment based, articulatory, as well as widely used context dependent (CD) modeling techniques. Even in carefully articulated speech, the production of a particular phoneme results from a continuous gesture of the articulators, coming from the configuration of the previous phoneme, and going to the configuration of the following phoneme. In different and more relaxed speaking styles, stronger pronunciation effects always appear. Some of these being particular to a language (and mostly unconscious). Other are related to regional origin, and are referred to as accents (or dialects for the linguistic counterpart) or to social groups and are referred to as sociolects. Although some of these phenomena may be mod-

eled appropriately by CD modeling techniques, their impact is rather characterized more simply at the pronunciation model level. At this stage, phonological knowledge may be helpful, especially in the case of strong effects like foreign accent. Fully data-driven techniques have also been proposed.

Following coarticulation and pronunciation effects, speaker related spectral characteristics (and gender) have been identified as another major dimension of speech variability. Specific models of frequency warping (based on vocal tract length differences) have been proposed, as well as more general features compensation and model adaptation techniques, relying on Maximum Likelihood or Maximum a Posteriori criteria. These model adaptation techniques provide a general formalism for re-estimation based on moderate amounts of speech data.

Besides these speaker specific properties outlined above, other extra-linguistic variabilities are admittedly affecting the signal and ASR systems. A person can change his voice to be louder, quieter, more tense or softer, or even a whisper; Also, some reflex effects exist, such as speaking louder when the environment is noisy.

Speaking faster or slower, also has influence on the speech signal. This impacts both temporal and spectral characteristics of the signal, both affecting the acoustic models. Obviously, faster speaking rates may also translate in more frequent and stronger pronunciation changes.

Speech also varies with age, due to both generational and physiological reasons. The two “extremes” of the range are generally put at a disadvantage due to the fact that research corpora, as well as corpora used for model estimation, are typically not designed to be representative of children and elderly speech. Some general adaptation techniques can however be applied to counteract this problem.

Emotions are also becoming a hot topic, as they can indeed have a negative effect on ASR; and also because added-value can emerge from applications that are able to identify the user emotional state (frustration due to compromised usability for instance).

Finally, research on recognition of spontaneous conversations has allowed to highlight the strong detrimental impact of this elocution style; and current studies are trying to better characterize pronunciation variation phenomena inherent in spontaneous speech.

This paper reviews some current advances related to these topics. It focuses on variations within the speech signal that make the ASR task difficult. Intrinsic variations to the speech signal affect the different levels of the ASR processing chain. Extrinsic variations have been more studied in the past but recent new approaches deserve a special report in this paper that summarizes the current literature without pretending to be exhaustive and highlights specific feature extraction or modeling weaknesses.

The paper is organized as follows. In a first section, intrinsic variability factors are reviewed individually according to the major trends identified in the literature. The section gathers information on the effect of variations on the structure of speech as well as the ASR performance. Typical modeling or engineering solutions that have been adopted at the different stages of the ASR chain are also introduced.

In general, this review further motivates research on the acoustic, phonetic and pronunciation limitations of speech recognition by machines. It is for instance acknowledged that pronunciation discrepancies is a major factor of reduced performance (in the case of accented and spontaneous speech). Section 3 reviews ongoing trends and possible breakthroughs in general feature extraction and modeling techniques that provides more resistance to speech production variability. It also includes recent techniques for noise/channel compensation. The issues that are being addressed include the fact that temporal representations/models may not match the structure of speech, as well as the fact that some analysis and modeling assumptions can be detrimental. General techniques such as compensation, adaptation, multiple models, additional acoustic cues and more accurate models are briefly surveyed.

2. Variation in speech

2.1. Speaker characteristics

Obviously, the speech signal not only conveys the linguistic information (the message) but also a lot of information about the speaker himself: gender, age, social and regional origin, health and emotional state and, with a rather strong reliability, its identity. Apart from the intra- speaker variability (emotion, health, age), it is commonly admitted that the speaker uniqueness results from a complex combination of physiological and cultural aspects [1]. While investigating the variability between speakers through statistical analysis methods, [2] found that the first two principal components correspond to the gender and accent respectively. Gender would then appear as the prime factor related to physiological differences, and accent would be one of the most important from the cultural point of view. This section deals mostly with physiological factors.

The complex shape of the vocal organs determines the unique ”timbre” of every speaker. The larynx which is the location of the source of the speech signal conveys the pitch and important speaker information. The vocal tract, can be modeled by a tube resonator [3]. The resonant frequencies (the formants) are structuring the global shape of the instantaneous voice spectrum and are mostly defining the phonetic content and quality of the vowels.

Standard feature extraction methods (PLP, MFCC) simply ignore the pitch component. On the other hand, the effect of the vocal tract shape on the intrinsic variability of the speech signal between different speakers has been widely studied and many solutions to compensate for its impact on ASR performance have been proposed: ”speaker independent” feature extraction, speaker normalization, speaker adaptation. The formant structure of vowel spectra has been the subject of early studies [4] that amongst other have established the standard view that the F1-F2 plane is the most descriptive, two-dimensional representation of the phonetic quality of spoken vowel sounds. On the other hand, similar studies underlined the speaker specificity of higher formants and spectral content above 2.5 kHz [5]. Other important studies [6] suggested that relative positions of the formant frequencies are rather constant for a given sound spoken by different speakers and, as a corollary, that absolute formant positions are speaker-specific. These observations are corroborated by the acoustic theory applied to the tube resonator model of the vocal tract which states that positions of the resonant frequencies are inversely proportional to the length of the vocal tract [7]. This observation is at the root of different techniques that increase the robustness of ASR systems to inter-speaker variability.

The preponderance of lower frequencies for carrying the linguistic information has been assessed by both perceptual and acoustical analysis and justify the success of the non-linear frequency scales such as Mel, Bark, Erb. Other approaches aim at building acoustic features invariant to the frequency warping [8, 9]. A direct application of the tube resonator model of

the vocal tract lead to the different vocal tract length normalization (VTLN) techniques: speaker-dependent formant mapping [10], transformation of the LPC pole modeling [11], frequency warping, either linear [12] or non-linear [13], all consists in modifying the position of the formants in order to get closer to an "average" canonical speaker. Incidentally, channel compensation techniques such as the cepstral mean subtraction or the RASTA filtering of spectral trajectories, also compensate for the speaker-dependent component of the long-term spectrum [14, 15].

On the other side, general adaptation techniques reduce speaker specificities and tends to further reduce the gap between speaker-dependent and speaker-independent ASR by adapting the acoustic models to a particular speaker [16, 17]

2.2. Foreign and regional accents

As introduced earlier, accent is one of the major components of interspeaker variability, as demonstrated in [2]. And indeed, compared to native speech recognition, performances degrades when recognizing accented speech and even more for non-native speech recognition [18]. In fact accented speech is associated to a shift within the feature space [19]. For native accents the shift is applied by large groups of speakers, is more or less important, more or less global, but overall acoustic confusability is not changed significantly. On the opposite, for foreign accents, the shift is very variable, is influenced by the native language, and depends also on the level of proficiency of the speaker.

Regional variants correspond to significantly different data, and enriched modelling is generally used to handle such variants. This can be achieved through the use of multiple acoustic models associated to large groups of speakers as in [20] or through the introduction of detailed pronunciation variants at the lexical level [21]. However adding too many systematic pronunciation variants may be harmful [22].

Non-native speech recognition is not properly handled by native speech models, no matter how much dialect data is included in the training [23]. This is due to the fact that non-native speakers can replace an unfamiliar phoneme in the target language, which is absent in their native language phoneme inventory, with the one considered as the closest in their native language phoneme inventory [24]. This behaviour makes the non-native alterations dependent on both the native language and the speaker. Some sounds may be replaced by other sounds, or inserted or omitted, and such insertion/omission behaviour cannot be handled by the usual triphone-based modelling [25]. In the specific context of speaker dependent recognition, adaptation techniques can be used [18]. For speaker independent systems this is not feasible. Introducing multiple phonetic transcriptions that handle alterations produced by non-native speakers is a usual approach, and is generally associated to a combination of phone models of the native language with phone models of the target language [26]. When a single foreign accent is handled, some accented data can be used for training or adapting the acoustic models [27]. Proper and foreign name processing is another topic strongly related with foreign accent [28].

Multilingual phone models are investigated since many years in the hope of achieving language independent units [29]. Language independent phone models are often useful when little or no data exists in a particular language and their use reduces the size of the phoneme inventory of multilingual speech recognition systems. The mapping between phoneme models of different languages can be derived from data [30] or determined from phonetic knowledge [31], but this is far from obvious as each language has his own characteristic set of phonetic units and associated distinctive features. Moreover, a phonemic distinguishing feature for a given language may hardly be audible to a native of another language.

Altough accent robustness is a desirable property of spoken language systems, accent classification is also studied since many years [32]. As a contiguous topic, speech recognition technology is also used in foreign language learning for rating the quality of the pronunciation [33].

2.3. Speaking rate and style

Speaking rate, expressed for instance in phonemes or syllables per second, is an important factor of intra-speaker variability. When speaking a a fast rate, the timing and acoustic realization of syllables are strongly affected due in part to the limitations of the articulatory machinery.

In automatic speech recognition, the significant performance degradations caused by speech rate variations stimulated many studies for modeling the spectral effects of speech rate variations. All the schemes presented in the literature make use of a speech rate estimator, based on different methods, providing the number of phones or syllables per second. The most common methods rely upon the evaluation of the frequency of phonemes or syllables in a sentence [34], through a preliminary segmentation of the test utterance; other approaches perform a normalization by dividing the measured phone duration by the average duration of the underlying phone [35]. Some approaches address the pronunciation correlates of fast speech. In [36], the authors rely upon an explicit modeling strategy, using different variants of pronunciation.

In casual situations or under time pressure, sluring pronunciations of certain phonemes indeed happen. Besides physiology, this builds on the speech redundancy and it has been hypothesized that this slurring affects more strongly sections that are more easily predicted. In contrast, speech portions where confusability is higher tend to be articulated more carefully [37].

In circumstances where the transmission and intelligibility of the message is at risk, a person can make use of an opposite articulatory behaviour, and for instance articulate more distinctly. Another related phenomenon happens in noisy environments where the speaker adapts s(maybe unconsciously) with the communicative purpose of increasing the intelligibility. This effect of augmented tension on the vocal folds as well as augmented loudness is known as the Lombard reflex [38].

These are crucial issues and research on speaking style specificities as well as spontaneous speech modeling is hence very active. Techniques to increase accuracy towards spontaneous speech have mostly focused on pronunciation studies[1]. Also, the strong dependy of pronunciation phenomena with respect to the syllable structure has been highlighted [39, 40]. As a consequence, extensions of acoustic modeling dependency to the phoneme position in a syllable and to the syllable position in word and sentences have been proposed [39].

Variations in spontaneous speech can also extend beyond the typical phonological alterations outlined previously. Phenomena called disfluencies can also be present, such as false starts, repetitions, hesitations and filled pauses. The reader will find useful information in [41, 42].

2.4. Age

Age is another major cause of variability and mismatch in speech recognition systems. The first reason is of physiological

[1] besides language modeling which is out of the scope of this paper

nature [43]. Children have shorter vocal tract and vocal folds compared with adults. This results in higher position of formants and fundamental frequency. The high fundamental frequency is reflected as a large distance between the harmonics, resulting in poor spectral resolution of voiced sounds. The difference in vocal tract size results in a non-linear increase of the formant frequencies.

In order to reduce this effect, previous studies have focused on the acoustic analysis of children speech [44, 45]. This work has put in evidence the challenges faced by Speech Recognition systems that will be developed to automatically recognize children speech. For example, it has been shown that children below the age of 10 exhibit a wider range of vowel durations relative to older children and adults, larger spectral and suprasegmental variations, and wider variability in formant locations and fundamental frequencies in the speech signal.

Obvisously, younger children may not have a correct pronunciation. Sometimes they have not yet learnt how to articulate specific phonemes [46]. Finally, children are using language in a different way. The vocabulary is smaller but may also contain words that don't appear in grown-up speech. The correct inflectional forms of certain words may not have been acquired fully, especially for those words that are exceptions to common rules. Spontaneous speech is also believed to be less grammatical than for adults. A number of different solutions have been proposed, modification of the pronunciation dictionary, and the use of language models which are customized for children speech have all been tried [47].

Several studies have attempted to address this problem by adapting the acoustic features of children speech to match that of acoustic models trained from adult speech [48]. Such approaches include vocal tract length normalization (VTLN) [49] as well as spectral normalization [50]. However, most of these studies point to lack of children acoustic data and resources to estimate speech recognition parameters relative to the over abundance of existing resources for adult speech recognition. Simply training a conventional speech recognizer on children speech is not sufficient to yield high accuracies, as demonstrated by Wilpon and Jacobsen [51]. Recently, corpora for children speech recognition have begun to emerge (for instance [52, 53] and [54] of the PF-STAR project).

2.5. Emotions

Similarly to the previously discussed speech intrinsic variations, emotional state is found to significantly influence the speech spectrum. It is recognized that a speaker mood change has a considerable impact on the features extracted from his speech, hence directly affecting the basis of all speech recognition systems.

Studies on speaker emotions is a fairly recent, emerging field and most of today literature that remotely deals with emotions in speech recognition is concentrated on attempting to classify a "stressed" speech signal into its correct emotion category. The purpose of these efforts is to further improve man-machine communication. The studies that interest us are different. Being interested in speech intrinsic variabilities, we focus our attention on the recognition of speech produced in different emotional states. The stressed speech categories studied are generally a collection of all the previously described intrinsic variabilities: loud, soft, Lombard, fast, angry, scared; and noise.

As Hansen formulates it in [55], approaches for robust recognition can be summarized under three areas: (i) better training methods, (ii) improved front-end processing, and (iii) improved back-end processing or robust recognition measures. A majority of work undertaken up to now revolves around inspecting the specific differences in the speech signal under the different stress conditions. Concerning the research specifically geared towards robust recognition, the first approach, based on improved training methods, comprises the following works: multi-style training [56], and simulated stress token generation [57]. As for all the improved training methods, recognition performance is increased only around the training conditions and degradation in results is observed as the test conditions drift from the original training data.

The second category of research is front-end processing, the goal being to devise feature extraction methods tailored for the recognition of stressed and non-stressed speech simultaneously [55, 58].

Finally, some interest has been focused on improving back-end processing as means of robust recognition. These techniques rely on adapting the model structure within the recognition system to account for the variability in the input signal. Consequently to the drawback of the "improved modeling" approach, one practice has been to bring the training and test conditions closer by space projection [59].

3. ASR techniques

In this section, we review methodologies towards improved ASR analysis/modeling accuracy and resistance towards variability sources.

3.1. Front-end techniques

An update on feature extraction front-end is proposed, particularly showing how to take advantage of techniques targeting the non-stationarity assumption. Also, the feature extraction stage can be the appropriate level to target some other variations, like the speaker spectral characteristics (through feature compensation [60] or else improved invariance [9]) and other dimensions of speech variability. Also, noise reduction can be achieved by feature compensation. Finally, techniques for combining estimation based on different features sets are reminded. This may also involve dimensionality reduction approaches.

3.1.1. Overcoming assumptions

Most of the Automatic Speech Recognition (ASR) acoustic features, such as Mel-Frequency Cepstral Coefficient (MFCC)[61] or Perceptual Linear Prediction (PLP) coefficient[62], are based on some sort of representation of the smoothed spectral envelope, usually estimated over fixed analysis windows of typically 20 ms to 30 ms [61, 63]. Such analysis is based on the assumption that the speech signal is quasi-stationary over these segment durations. However, it is well known that the voiced speech sounds such as vowels are quasi-stationary for 40ms-80ms while, stops and plosive are time-limited by less than 20ms [63]. Therefore, it implies that the spectral analysis based on a fixed size window of 20ms-30ms has some limitations, including:

- The frequency resolution obtained for quasi-stationary segments (QSS) longer than 20ms is quite low compared to what could be obtained using larger analysis windows.
- In certain cases, the analysis window can span the transition between two QSSs, thus blurring the spectral properties of the QSSs, as well as of the transitions. Indeed, in theory, Power Spectral Density (PSD) cannot even be defined for such non stationary segments [64]. Further-

more, on a more practical note, the feature vectors extracted from such transition segments do not belong to a single unique (stationary) class and may lead to poor discrimination in a pattern recognition problem .

In [65], the usual assumption is made that the piecewise quasi-stationary segments (QSS) of the speech signal can be modeled by a Gaussian AR process of a fixed order p as in [66, 67, 68]. The problem of detecting QSSs is then formulated using a Maximum Likelihood (ML) criterion, defining a QSSs as the longest segment that has most probably been generated by the same AR process.[2] Given a p^{th} order AR Gaussian QSS, the Minimum Mean Square Error (MMSE) linear prediction (LP) filter parameters $[a(1),\ a(2), ...\ a(p)]$ are the most "compact" representation of that QSS amongst all the p^{th} order all pole filters [64]. In other words, the normalized "coding error"[3] is minimum amongst all the p^{th} order LP filters. When erroneously analyzing two distinct p^{th} order AR Gaussian QSSs in the same non-stationary analysis window, it can be shown that the "coding error" will then always be greater than the ones resulting of QSSs analyzed individually in stationary windows[64]. Therefore, higher coding error is expected in the former case as compared to the optimal case when each QSS is analyzed in a stationary window. Once the "start" and the "end" points of a QSS are known, all the speech samples coming from this QSS are analyzed within that window, resulting in (variable-scale) acoustic vectors.

Another approach is proposed in [69], which described a temporal decomposition technique to represent the continuous variation of the LPC parameters as a linearly weighted sum of a number of discrete elementary components. These elementary components are designed such that they have the minimum temporal spread (highly localized in time) resulting in superior coding efficiency. However, the relationship between the optimization criterion of "the minimum temporal spread" and the quasi-stationarity is not obvious. Therefore, the discrete elementary components are not necessarily quasi-stationary and vice-versa.

Coifman et al [70] have described a minimum entropy basis selection algorithm to achieve the minimum information cost of a signal relative to the designed orthonormal basis. In [66], Svendsen et al have proposed a ML segmentation algorithm using a single fixed window size for speech analysis, followed by a clustering of the frames which were spectrally similar for sub-word unit design. More recently, Achan et al [71] have proposed a segmental HMM for speech waveforms which identifies waveform samples at the boundaries between glottal pulse periods with applications in pitch estimation and time-scale modifications.

As a complementary principle to developing features that "work around" the non-stationarity of speech, significant efforts have also been made to develop new speech signal representations which can better describe the non-stationarity inherent in the speech signal. Some representative examples are temporal patterns features[72], MLP and the several modulation spectrum related techniques[73, 74, 75, 76]. In this approach temporal trajectories of spectral energies in individual critical bands over windows as long as one second are used as features for pattern classification. Another methodology is to use the notion of the amplitude modulation (AM) and the frequency modulation (FM) [77]. In theory, the AM signal modulates a narrow-band carrier signal (specifically, a monochromatic sinusoidal signal). Therefore to be able to extract the AM signals of a wide-band signal such as speech (typically 4KHz), it is necessary to decompose the speech signal into narrow spectral bands. In [78], this approach is opposed to the previous use of the speech modulation spectrum [73, 74, 75, 76] which was derived by decomposing the speech signal into increasingly wider spectral bands (such as critical, Bark or Mel). Similar arguments from the modulation filtering point of view, were presented by Schimmel and Atlas[79]. In their experiment, they consider a wide-band filtered speech signal $x(t) = a(t)c(t)$, where $a(t)$ is the AM signal and $c(t)$ is the broad-band carrier signal. Then, they perform a low-pass modulation filtering of the AM signal $a(t)$ to obtain $a_{LP}(t)$. The low-pass filtered AM signal $a_{LP}(t)$ is then multiplied with the original carrier $c(t)$ to obtain a new signal $\tilde{x}(t)$. They show t hat the acoustic bandwidth of $\tilde{x}(t)$ is not necessarily less than that of the original signal $x(t)$. This unexpected result is a consequence of the signal decomposition into wide spectral bands that results in a broad-band carrier.

Finally, as extension to the "traditional" AR process (all-pole model) speech modeling, pole-zero transfer functions that are used for modeling the frequency response of a signal, have been well studied and understood [80]. Lately, Kumaresan et al.[81, 82] have proposed to model analytic signals using pole-zero models in the temporal domain. Along similar lines, Athineos et al.[83] have used the dual of the linear prediction in the frequency domain to improve upon the TRAP features.

3.1.2. Compensation and invariance

Simple models may exist that appropriately reflects the effect of a variability on speech features. This is for instance the case for long-term spectral characteristics, mostly referred to the Vocal Tract Length (VTL) of the speaker. Simple yet powerful techniques for normalizing (compensating) the features to the VTL are widely used [60].

An alternative to normalization is the generation of invariant features. For vocal tract length for instance, [8, 9] propose to exploit the fact that vocal-tract length variations can be approximated via linear frequency warping. In [8], the scale transform and the scale cepstrum have been introduced. Both transforms exhibit the interesting property that their magnitudes are invariant to linear frequency warping. In [9], the continuous wavelet transform has been used as a preprocessing step, in order to obtain a speech representation in which linear frequency scaling leads to a translation in the time-scale plane. In a second step, frequency-warping invariant features were generated. These include the auto- and cross-correlation of magnitudes of local wavelet spectra as well as linear and nonlinear transforms thereof. It could be shown that these features not only lead to better recognition scores than standard MFCCs, but that they are also more robust to mismatches between training and test conditions, such as training on male and testing on female data. The best results were obtained when MFCCs and the vocal tract length invariant features were combined, showing that both sets contain complementary information [9].

Normalization (compensation) or invariance with respect to other dimensions may also be useful (f.i. with respect to speaking rate).

When simple parametric models of the effect of the variability are not appropriate, feature compensation can be performed using more generic non-parametric transformation

[2] Equivalent to the detection of the transition point between the two adjoining QSSs.

[3] The power of the residual signal normalized by the number of samples in the window

schemes, including linear and non-linear transformation. This becomes a dual approach to model adaptation, which is the topic of Section 3.2.2.

3.1.3. Noise compensation

One of the most popular technique for increasing recognition accuracy in noise is the spectral subtraction [84, 85] where noise sprectrum is estimated during short pauses and subtracted from the spectrum of noisy speech. Although this method is not appropriate for non-stationary noise, slowly varying noise can be removed from the signal since noise spectrum is regularly updated. Two major drawbacks are the difficulty to detect pauses (non-speech) in low SNR and that the subtraction should be carefully controlled to avoid negative values for the "clean" speech spectrum that leads to the so-called musical noise effect [86, 87]. Also , the assumption that noisy speech power spectrum is the sum of noise power spectrum and the clean speech power spectrum is not correct (see more recent techniques where this hypothesis is relaxed).

Another noise reduction method consists in filtering speech with a high order adaptive FIR filter [88]. When no reference to an external noise source is available, A Wiener linear prediction filter may suppress interfering noise under the hypotheses of stationarity of input and noise and if noise spectrum is much wider than spectrum of the input. The main advantage of this method is that no noise reference source is required. In speech case, most hypotheses are not valid but for voiced speech, the signal can be seen as a sequence of sinusoids: interesting results can be demonstrated.

Statistical approaches for noise reduction have been reported in [89]. More recently, several new approaches like uncertainty decoding [90] and the SPLICE algorithm described by [91, 92] have raised a strong interest for these techniques that estimate simultaneously noise and clean speech making a priori hypotheses on their distributions.SPLICE works on spectral representation of speech. Another similar algorithm ALGONQUIN works on log-spectrra and has been described recently in Kristjansson's thesis [93]where the hypotheses of decorrelation between noise and clean speech are shown unnecessary. Also these authors deal with the convolution noise. Up to now, convolution noise that is usually varying slowly, can be low pass filtered out: this is achieved by removing cepstral mean from all feature vectors of the utterance [94, 95].

3.1.4. Additional cues and feature combinations

As a complementary perspective to improving or compensating single feature sets, one can also make use of several "streams" of features that rely on different underlying assumptions and exhibit different properties.

Intrinsic feature variability depends on the set of classes that feature have to discriminate. Given a set of acoustic measurements, algorithms have been described to select subsets of them that improve automatic classification of speech data into phonemes or phonetic features. Unfortunately, pertinent algorithms are computationally intractable with this types of classes as stated in [96], [97], where a sub-optimal solution is proposed. It consists in selecting a set of acoustic measurement that guarantees a high value of the mutual information between acoustic measurements and phonetic distinctive features.

Without attempting to find an optimal set of acoustic measurements, many recent automatic speech recognition systems combine streams of different acoustic measurements on the assumption that some characteristics that are de-emphasized by a particular feature are emphasized by another feature, and therefore the combined feature streams capture complementary information present in individual features. In [98], it is shown that log-linear combination provides good results when used for integrating probabilities provided by acoustic models.

In order to take into account different temporal behaviors in different bands, it has been proposed ([99, 100, 101]) to consider separate streams of features extracted in separate channels with different frequency bands. Other approaches integrate some specific parameters into a single stream of features. Examples of added parameters are:

- periodicity and jitter ([102]),
- voicing ([103], [104]),
- rate of speech and pitch ([105]).

To benefit from the strengths of both MLP-HMM and Gaussian-HMM techniques, the Tandem solution was proposed in [106], using posterior probability estimation obtained at MLP outputs as observations for a Gaussian-HMM. An error analysis of Tandem MLP features showed that the errors using MLP features are different from the errors using cepstral features. This motivates the combination of both feature styles. In ([107]), combination techniques were applied to increasingly more advanced systems showing the benefits of the MLP-based features. These features have been combined with TRAP features ([98, 108]). In ([109]), Gabor filters are proposed, in conjunction with MLP features, to model the characteristics of neurons in the auditory system as they do in the visual system. There is evidence that in primary auditory cortex each individual neuron is tuned to a specific combination of spectral and temporal modulation frequencies.

In [110], it is proposed to use mixture gaussians to represent presence and absence of features.

Additional features have also been considered as cues for speech recognition failures [111].

3.1.5. Dimensionality reduction and feature selection

Using additional features/cues as reviewed in the previous section, or simply extending the context by concatenating feature vectors from adjacent frames may yield very long feature vectors in which several features contain redundant information, thus requiring an additional dimension-reduction stage [112, 113] and/or improved training procedures [114].

The most common feature-reduction technique is the use of a linear transform $\boldsymbol{y} = A\boldsymbol{x}$ where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the original and the reduced feature vectors, respectively, and A is a $p \times n$ matrix with $p < n$ where n and p are the original and the desired number of features, respectively. The principal component analysis (PCA) [115, 116] is the most simple way of finding A. It allows for the best reconstruction of $\boldsymbol{x}$ from $\boldsymbol{y}$ in the sense of a minimal average squared Euclidean distance. However, it does not take the final classification task into account and is therefore only suboptimal for finding reduced feature sets. A more classification-related approach is the linear discriminant analysis (LDA), which is based on Fisher's ratio (F-ratio) of between-class and within-class covariances [115, 116]. Here the columns of matrix A are the eigenvectors belonging to the p largest eigenvalues of matrix $[S_w^{-1} S_b]$, where S_w and S_b are the within-class and between-class scatter matrices, respectively. Good results with LDA have been reported for small vocabulary speech recognition tasks, but for large-vocabulary speech recognition, results were mixed [112]. In [112] it was found that the LDA should best be trained on sub-phone units in order to serve

as a preprocessor for a continuous mixture density based recognizer. A limitation of LDA is that it cannot effectively take into account the presence of different within-class covariance matrices for different classes. Heteroscedastic discriminant analysis (HDA) [113] overcomes this problem, but the method usually requires the use of numerical optimization techniques to find the matrix A. An exception is the method in [117], which uses the Chernoff distance to measure between-class distances and leads to a straight forward solution for A. Finally, LDA and HDA can be combined with maximum likelihood linear transform (MLLT) [118], which is a special case of semi-tied covariance matrices (STC) [119]. Both aim at transforming the reduced features in such a way that they better fit with the diagonal covariance matrices that are applied in many HMM recognizers. It has been reported [120] that such a combination performs better than LDA or HDA alone. Also, HDA has been combined with minimum phoneme error (MPE) analysis [121]. Recently, the problem of finding optimal dimension-reducing feature transformations has been studied from the viewpoint of maximizing the mutual information between the obtained feature set and the corresponding phonetic class [96, 122].

A problem of the use of linear transforms for feature reduction is that the entire feature vector $\boldsymbol{x}$ needs to be computed before the reduced vector $\boldsymbol{y}$ can be generated. This may lead to a large computational cost for feature generation, although the final number of features may be relatively low. An alternative is the direct selection of feature subsets, which, expressed by matrix A, means that each row of A contains a single one while all other elements are zero. The question is then the one of which features to include and which to exclude. Because the elements of A have to be binary, simple algebraic solutions like with PCA or LDA cannot be found, and iterative strategies have been proposed. For example, in [123], the maximum entropy principle was used to decide on the best feature space.

3.2. Acoustic modeling techniques

Concerning acoustic modeling, good performance is generally achieved when the model is matched to the task, which can be obtained through adequate training data. Systems with stronger generalization capabilities can then be built through a so-called multi-style training. Estimating the parameters of a traditional modeling architecture in this way however has some limitation due to the inhomogeneity of the data, which increases the spread of the models, and hence negatively impacts accuracy compared to task-specific models. This is partly to be related to the inability of the framework to properly model long-term correlations of the speech signals.

Also, within the acoustic modeling framework, adaptation techniques provide a general formalism for reestimating optimal model parameters for given circumstances based on moderate amounts of speech data.

Then, the modeling framework can be extended to allow multiple specific models to cover the space of variation. These can be obtained through generalizations of the HMM modeling framework, or through explicit construction of multiple models build on knowledge-based or data-driven clusters of data.

In the following, extensions for modeling using additional cues and features is also reviewed.

3.2.1. Model compensation

In section 3.1.3, feature compensation techniques were reported for enhancing speech features. A dual approach is to apply acoustic model compensation. Two main techniques were proposed. In [124], Moore proposed MM decomposition where dynamic time warping was extended to a 3D-array where the additional dimension represents a noise reference and an optimal path has to be found in this 3D-domain. The major problenm was the definition of a local Probality for each box [125]. Existence of a single noise model is also a severe limitation.

This last difficulty was circumvented by a parallel model decomposition (PMC) [126, 127] where clean speech and noise are both modeled by HMM and where the local probabilities are combined at the level of linear spectrum: this implies that only additive noise can be taken into account.

Feature compensation methods seem to be more successful than model compensation. Howevere there is a strong relation between the two techniques that is particularly well illustrated by contrained MLLR (C-MLLR) [128]where the transformation matrix for the covariance matrices is the same as the matrix for the mean vectors. In that case for Gaussian distriutions it is trivial to observe that model compensation is strictly equal to feature transformation (this is no longer valid in case of compensation of state classes).

3.2.2. Adaptation

In Section 3.1.2, we have been illustrating techniques that can be used to compensated the features to speech variation. the dual approach is to adapt the ASR acoustic models.

In some cases, some variations in the speech signal could be considered as long term given the applicative scenario. For instance, a system embedded in a personal device and hence mainly designed to be used by a single person, or a system designed to transcribe and index spontaneous speech, or characterized by utilization in a particular environment. In these cases, it is often possible to adapt the models to these particular conditions, hence partially factoring out the detrimental effect of these. A popular technique is to estimate a linear transformation of the model parameters using a Maximum Likelihood (ML) criterion [17]. A Maximum a Posteriori (MAP) objective function may also be used.

Being able to perform this adaptation using limited amounts of condition-specific data would be a very desirable property for such adaptation methodologies, as this would reduce the cost and hassle of such adaptation phases. Such “rapid” (sometimes on-line) adaptation schemes have been proposed a few year ago, mostly based on speaker-space methods, such as eigenvoices and cluster-based adaptation [129, 130].

Intuitively, these techniques rest on the principle of acquiring knowledge from the training corpora that represent the prior distribution (or clusters) of model parameters given a variability factor under study. With these adaptation techniques, knowledge about the effect of the inter-speaker variabilities are gathered in the model. In the traditional approach, this knowledge is simply discarded, and, although all the speakers are used to build the model, and pdfs are modeled using mixtures of gaussians, the ties between particular mixture components across the several CD phonemes are not represented/used.

Recent publications have been extending and refining this class of techniques. In [131], rapid adaptation is further extended through a more accurate speaker space model, and an on-line algorithm is also proposed. In [132], the correlations between the means of mixture components of the different features are modeled using a Markov Random Field, which is then used to constrain the transformation matrix used for adaptation. Other publications include [132, 133, 134, 135, 136, 137]

Other forms of transformations for adaptation are also pro-

posed in [138], where the Maximum Likelihood criterion is used but the transformation are allowed to be nonlinear.

3.2.3. Multiple modeling

Instead of adapting the models to particular conditions, one may also train ensemble of models specialized to specific conditions or variations. These models may then be used within a selection, competition or else combination framework. Such techniques are the object of this section.

Acoustic models are estimated from speech corpora, and they provide their best recognition performances when the operating (or testing) conditions are consistent with the training conditions. Hence many adaptation procedures were studied to adapt generic models to specific tasks and conditions. When the speech recognition system have to handle various possible conditions, several speech corpora can be used together for estimating the acoustic models, leading to mixed models or hybrid systems [139, 140], which provide good performances in those various conditions (for example in both landline and wireless networks). However, merging too many heterogeneous data in the training corpus makes acoustic models less discriminant. Hence the numerous investigations along multiple modeling, that is the usage of several models for each unit, each model being train from a subset of the training data, defined according to a priori criteria such as gender, age, rate-of-speech (ROS) or through automatic clustering procedures. Ideally subsets should contain homogeneous data, and be large enough for making possible a reliable training of the acoustic models.

Gender information is one of the most often used criteria. It leads to gender-dependent models that are either directly used in the recognition process itself [141, 142] or used as a better seed for speaker adaptation [143]. Gender dependence is applied to whole word units, for example digits [144], or to context dependent phonetic units [142], as a result of an adequate splitting of the training data.

Age dependent modeling has been less investigated, may be due to the lack of large size children speech corpora. The results presented in [145] fail to demonstrate a significant improvement when using age dependent acoustic models, possibly due to the limited amount of training data for each class of age.

Speaking rate affects notably the recognition performances, thus speaking rate dependent models were studied [34]. It was also noticed that speaking rate dependent models are often getting less speaker-independent because the range of speaking rate shown by different speakers is not the same [146], and that training procedures robust to sparse data need to be used. Speaking rate can be estimated on line [146], or computed from a decoding result using a generic set of acoustic models, in which case a rescoring is applied for fast or slow sentences [147]; or the various rate dependent models may be used simultaneously during decoding [148, 149].

Signal-to-Noise Ratio (SNR) also impacts recognition performances, hence, besides or in addition to noise reduction techniques, SNR-dependent models have been investigated. In [150] multiple sets of models are trained according to several noise masking levels and the model set appropriate for the estimated noise level is selected automatically in recognition phase. On the opposite, in [151] acoustic models composed under various SNR conditions are run in parallel during decoding.

Automatic clustering techniques have also been used for elaborating several models per word for connected-digit recognition [152]. Clustering the trajectories deliver more accurate modeling for the different groups of speech samples [153]; and clustering training data at the utterance level provides the best performances [154].

Multiple modeling of phonetic units may be handled also through the usual triphone-based modeling approach by incorporating questions on some variability sources in the set of questions used for building the decision trees: gender information in [155]; syllable boundary and stress tags in [156]; and voice characteristics in [157].

When multiple modeling is available, all the available models may be used simultaneously during decoding, as done in many approaches, or the most adequate set of acoustic models may be selected from a priori knowledge (for example network or gender), or their combination may be handled dynamically by the decoder. This is the case of parallel hidden Markov models [158] where the acoustic densities are modulated depending on the probability of a master context HMM being in certain states. More recently Dynamic Bayesian Networks have been used to handle dependencies of the acoustic models with respect to auxiliary variables, such as local speaking rate [159], or hidden factors related to a clustering of the data [160, 161].

Multiple models can also be used in a parallel decoding framework [162]; then the final answer results from a "voting" process [163], or from the application of elaborated decision rules that take into account the recognized word hypotheses [164]. Multiple decoding is also useful for estimating reliable confidence measures [165].

At the pronunciation level, multiple pronunciations are generally used for the vocabulary words. Hidden model sequences offer a possible way of handling multiple realizations of phonemes [166] possibly depending on phone context. For handling hyper articulated speech where pauses may be inserted between syllables, ad hoc variants are necessary [161]. And, as detailed in section 2.2, adding more variants is usually required for handling foreign accents.

Also, if models of some of the factors affecting speech variation are known, adaptive training schemes can be developed, avoiding training data sparsity issues that could result from cluster-based techniques. This has been used for instance in the case of VTLN normalization, where a specific estimation of the vocal tract length (VTL) is associated to each speaker of the training data [60]. This allows to build "canonical" models based on appropriately normalized data. During recognition, a VTL is estimated in order to be able to normalize the feature stream before recognition. More general normalization schemes have also been investigated [167], based on associating transforms (mostly linear transforms) to each speaker, or more generally, to different cluster of the training data. These transforms can also be constrained to reside in an reduce-dimensionality eigenspace [130]. A technique for "factoring-in" selected transformations back in the canonical model is also proposed in [168], providing a flexible way of building factor-specific models, for instance multi-speaker models within a particular noise environment, or multi-environment models for a particular speaker.

3.2.4. Auxiliary parameters

Most of speech recognition systems rely on acoustic parameters that represent the speech spectrum, for example cepstral coefficients. However, these features are sensitive to auxiliary information such as pitch, energy, rate-of-speech, etc. Hence attempts have been made in taking into account this auxiliary information in the modeling and in the decoding processes.

Pitch and voicing parameters have been used since a long

time, but mainly for endpoint detection purposes [169] making it much more robust in noisy environments [170]. Many algorithms have been developed and tuned for computing these parameters, but are out of the scope of this paper.

For what concerns speech recognition itself, the most simple way of using such parameters (pitch and/or voicing) is their direct introduction in the feature vector, along with the cepstral coefficients, for example periodicity and jitter are used in [171] for connected digits and large vocabulary. Correlation between pitch and acoustic features is taken into account in [172] and a LDA is applied on the full set of features (i.e. energy, MFCC, voicing and pitch) in [173].

Pitch has to be taken into account for the recognition of tonal languages. Tone can be modeled separately through specific HMMs [174] or decision trees [175], or the pitch parameter can be included in the feature vector [176], or both information streams (acoustic features and tonal features) can be handled directly by the decoder, possibly with different optimized weights [177]. Various coding and normalization schemes of the pitch parameter are generally applied to make it less speaker dependent; the derivative of the pitch is the most useful feature [178], and pitch tracking and voicing are investigated in [179]. A comparison of various modeling approaches is available in [180]. For tonal languages, pitch modeling usually concerns the whole syllable; however limiting the modeling to the vowel seems sufficient [181].

Voicing has been used in the decoder to constraint the Viterbi decoding (when phoneme node characteristics are not consistent with the voiced/unvoiced nature of the segment, corresponding paths are not extended) making the system more robust to noise [182].

Pitch, energy and duration have also been used as prosodic parameters in speech recognition systems, or for reducing ambiguity in post-processing steps. These aspects are out of scope of this paper.

Dynamic Bayesian Networks (DBN) offer an integrated formalism for introducing dependence on auxiliary features. This approach is used in [105] with pitch and energy as auxiliary features. Other information can also be taken into account such as articulatory information in [183] where the DBN utilizes an additional variable for representing the state of the articulators. As mentioned in previous section, speaking rate is another factor that can be taken into account in such a framework. Most experiments deal with limited vocabulary sizes; extension to large vocabulary continuous speech recognition is proposed through an hybrid HMM/BN acoustic modeling in [184].

Another approach for handling heterogeneous features is the TANDEM approach used with pitch, energy or rate of speech in [185]. The TANDEM approach transforms the input features into posterior probabilities of sub-word units using artificial neural networks (ANNs), which are then processed to form input features for conventional speech recognition systems.

Finally, auxiliary parameters may be used to normalize spectral parameters, for example based on pitch value in [186], or used to modify the parameters of the densities (during decoding) through multiple regressions as with pitch and speaking rate in [187].

4. Conclusion

This paper gathers important literature references related to the endogenous variation of the speech signal and their importance in automatic speech recognition. Important references addressing specific individual speech variation sources are first been surveyed. This covers accent, speaking style, speaker physiology, age, emotions. Finally, the paper proposed an overview of general techniques for better handling intrinsic and extrinsic variation sources in ASR, mostly tackling the speech analysis and acoustic modeling aspect.

5. Acknowledgments

This work has been partly supported by the EU 6th Framework Programme, under contract number IST-2002-002034 (DIVINES project). The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

References

[1] F. Nolan, *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press, 1983.

[2] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou, "Analysis of speaker variability," in *Proc. of Eurospeech*, (Aalborg, Denmark), pp. 1377–1380, Sept. 2001.

[3] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: a guide to theory, algorithm, and system development*. New Jersey: Prentice-Hall, 2001.

[4] R. K. Potter and J. C. Steinberg, "Toward the specification of speech," *The Journal of the Acoustical Society of America*, vol. 22, pp. 807–820, 1950.

[5] L. C. W. Pols, L. J. T. V. der Kamp, and R. Plomp, "Perceptual and physical space of vowel sounds," *The Journal of the Acoustical Society of America*, vol. 46, pp. 458–467, 1969.

[6] T. M. Nearey, *Phonetic feature systems for vowels*. Bloomington, Indiana, USA: Indiana University Linguistics Club, 1978.

[7] D. O'Saughnessy, *Speech communication - human and machine*. Addison-Wesley, 1987.

[8] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 40–45, Jan. 1999.

[9] A. Mertins and J. Rademacher, "Vocal tract length invariant features for automatic speech recognition," in *Proc. of ASRU*, (Cancun, Mexico), Dec. 2005.

[10] M.-G. D. Benedetto and J.-S. Liénard, "Extrinsic normalization of vowel formant values based on cardinal vowels mapping," in *Proc. of ICSLP*, pp. 579–582, 1992.

[11] J. Slifka and T. R. Anderson, "Speaker modification with lpc pole analysis," in *Proc. of ICASSP*, (Detroit, MI), pp. 644–647, May 1995.

[12] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *Proc. of ICASSP*, (Munich, Germany), 1997.

[13] Y. Ono, H. Wakita, and Y. Zhao, "Speaker normalization using constrained spectra shifts in auditory filter domain," in *Proc. of Eurospeech*, pp. 355–358, 1993.

[14] S. Kajarekar, N. Malayath, and H. Hermansky, "Analysis of source of variability in speech," in *Proc. of Eurospeech*, (Budapest, Hungary), Sept. 1999.

[15] M. Westphal, "The use of cepstral means in conversational speech recognition," in *Proc. of Eurospeech*, (Rhodos, Greece), 1997.

[16] C. Lee, C. Lin, and B. Juang., "A study on speaker adaptation of the parameters of continuous density Hidden Markov Models," *IEEE Trans. Signal Processing.*, vol. 39, pp. 806–813, April 1991.

[17] C. Leggetter and P. Woodland., "Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models," *Computer, Speech and Language*, vol. 9, pp. 171–185, April 1995.

[18] F. Kubala, A. Anastasakos, J. Makhoul, L. Nguyen, R. Schwartz, and E. Zavaliagkos, "Comparative experiments on large vocabulary speech recognition," in *Proc. of ICASSP*, Apr. 1994.

[19] D. V. Compernolle, "Recognizing speech of goats, wolves, sheep and ... non-natives," in *Speech Communication*, pp. 71–79, Aug. 2001.

[20] D. V. Compernolle, J. Smolders, P. Jaspers, and T. Hellemans, "Speaker clustering for dialectic robustness in speaker independent speech recognition," in *Proc. of Eurospeech*, 1991.

[21] J. J. Humphries, P. C. Woodland, and D. Pearce, "Using accent-specific pronunciation modelling for robust speech recognition," in *Proc. of ICSLP*, 1996.

[22] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: a survey of the literature," in *Speech Communication*, pp. 225–246, Nov. 1999.

[23] V. Beattie, S. Edmondson, D. Miller, Y. Patel, and G. Talvola, "An integrated multidialect speech recognition system with optional speaker adaptation," in *Proc. of Eurospeech*, 1995.

[24] J. E. Flege, C. Schirru, and I. R. A. MacKay, "Interaction between the native and second language phonetic subsystems," in *Speech Communication*, pp. 467–491, 2003.

[25] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen, "What kind of pronunciation variation is hard for triphones to model?," in *Proc. of ICASSP*, (Salt Lake City, Utah), May 2001.

[26] K. Bartkova and D. Jouvet, "Language based phone model combination for asr adaptation to foreign accent," in *Proc. of ICPhS*, (San Francisco, USA), Aug. 1999.

[27] W. K. Liu and P. Fung, "MLLR-based accent model adaptation without accented data," in *Proc. of ICSLP*, (Beijing, China), 2000.

[28] K. Bartkova, "Generating proper name pronunciation variants for automatic speech recognition," in *Proc. of ICPhS*, (Barcelona, Spain), 2003.

[29] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. of ICSLP*, 1998.

[30] F. Weng, H. Bratt, L. Neumeyer, and A. Stomcke, "A study of multilingual speech recognition," in *Proc. of Eurospeech*, (Rhodes, Greece), 1997.

[31] U. Uebler, "Multilingual speech recognition in seven languages," in *Speech Communication*, pp. 53–69, Aug. 2001.

[32] L. Arslan and J. Hansen, "Language accent classification in american english," *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[33] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," in *Speech Communication*, pp. 121–130, Feb. 2000.

[34] N. Mirghafori, E. Fosler, and N. Morgan, "Towards robustness to fast speech in ASR," in *Proc. of ICASSP*, (Atlanta, Georgia), pp. 335–338, May 1996.

[35] M. Richardson, M. Hwang, A. Acero, and X. D. Huang, "Improvements on speech recognition for fast talkers," in *Proc. of Eurospeech*, (Budapest, Hungary), Sept. 1999.

[36] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word predictability on conversational pronunciations," *Speech Communication*, vol. 29, no. 2-4, pp. 137–158, 1999.

[37] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, "Effects of disfluencies, predictability, and utterance position on word form variation in english conversation," *The Journal of the Acoustical Society of America*, vol. 113, pp. 1001–1024, Feb. 2003.

[38] J.-C. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognisers," *JASA*, vol. 93, pp. 510–524, Jan. 1993.

[39] S. Greenberg and S. Chang, "Linguistic dissection of switchboard-corpus automatic speech recognition systems," in *Proc. of ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millenium*, (Paris, France), Sept. 2000.

[40] M. Adda-Decker, P. B. de Mareuil, G. Adda, and L. Lamel, "Investigating syllabic structures and their variation in spontaneous french," *Speech Communication*, vol. 46, pp. 119–139, June 2005.

[41] S. Furui, M. B. J. Hirschberg, S. Itahashi, T. Kawahara, S. Nakamura, and S. Narayanan, "Introduction to the special issue on spontaneous speech processing," *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 349–350, July 2004.

[42] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Z. Wei-Jin, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 420–435, July 2004.

[43] D. Elenius and M. Blomberg, "Comparing speech recognition for adults and children," in *Proceedings of FONETIK 2004*, (Stockholm, Sweden), 2004.

[44] G. Potamianos, S. Narayanan, and S. Lee, "Analysis of children speech: duration, pitch and formants," in *Proc. of Eurospeech*, (Rhodes, Greece), Sept. 1997.

[45] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children speech: developmental changes of temporal and spectral parameters," in *The Journal of the Acoustical Society of America*, (Vol.105), pp. 1455–1468, Mar. 1999.

[46] S. Schötz, "A perceptual study of speaker age," in *Working paper 49 (2001), 136-139*, (Lund University, Dept Of Linguistic), Nov. 2001.

[47] G. P. M. Eskenazi, "Pinpointing pronunciation errors in children speech: examining the role of the speech recognizer," in *Proceedings of the PMLA Workshop*, (Colorado, USA), Sept. 2002.

[48] D. Giuliani and M. Gerosa, "Investigating recognition of children speech," in *Proc. of ICASSP*, (Hong Kong), Apr. 2003.

[49] S. Das, D. Nix, and M. Picheny, "Improvements in children speech recognition performance," in *Proc. of ICASSP*, vol. 1, (Seattle, USA), pp. 433–436, May 1998.

[50] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. of ICASSP*, vol. 1, (Atlanta, Georgia), pp. 353–356, May 1996.

[51] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. of ICASSP*, vol. 1, (Atlanta, Georgia), pp. 349–352, May 1996.

[52] M. Eskenazi, "Kids: a database of children's speech," in *The Journal of the Acoustical Society of America*, (Vol.100, No. 4, Part 2), Dec. 1996.

[53] K. Shobaki, J.-P. Hosom, and R. Cole, "The ogi kids speech corpus and recognizers," in *Proc. of ICSLP*, (Beijing, China), Oct. 2000.

[54] M. Blomberg and D. Elenius, "Collection and recognition of children speech in the pf-star project," in *In Proc of Fonetik 2003 Umea University*, (Department of Philosophy and Linguistics PHONUM), 2003.

[55] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, 1996.

[56] R. P. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. of ICASSP*, 1987.

[57] S. E. Bou-Ghazale and J. L. H. Hansen, "Improving recognition and synthesis of stressed speech via feature perturbation in a source generator framework," in *ECSA-NATO Proc. Speech Under Stress Workshop, Lisbon, Portugal*, 1995.

[58] B. A. Hanson and T. Applebaum, "Robust speaker-independent word recognition using instantaneous, dynamic and acceleration features: experiments with Lombard and noisy speech," in *Proc. of ICASSP*, 1990.

[59] B. Carlson and M. Clements, "Speech recognition in noise using a projection-based likelihood measure for mixture density hmms," in *Proc. of ICASSP*, 1992.

[60] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive training by vocal tract normalization," *IEEE Trans. Speech Audio Process.*, vol. 10, pp. 415–426, Sept. 2002.

[61] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, pp. 357–366, August 1980.

[62] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, Apr. 1990.

[63] L. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, ch. 2, pp. 20–37. Englewoood Cliffs, NJ, USA: Prentice Hall PTR, 1993.

[64] S. Haykin, *Adaptive filter theory*. Prentice-Hall Publishers, N.J., USA., 1993.

[65] V. Tyagi, C. Wellekens, and H. Bourlard, "On variable-scale piecewise stationary spectral analysis of speech signals for ASR," in *Proc. of Eurospeech*, (Lisbon, Portugal), September 2005.

[66] T. Svendsen, "On the automatic segmentation of speech signals," in *Proc. of ICASSP*, 1987.

[67] T. Svendsen, K. K. Paliwal, E. Harborg, and P. O. Husoy, "An improved sub-word based speech recognizer," in *Proc. of ICASSP*, 1989.

[68] R. A. Obrect, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, January 1988.

[69] B. Atal, "Efficient coding of LPC parameters by temporal decomposition," in *Proc. of ICASSP*, (Boston, USA), 1983.

[70] R. R. Coifman and M. V. Wickerhauser, "Entropy based algorithms for best basis selection," *IEEE Trans. on Information Theory*, vol. 38, March 1992.

[71] K. Achan, S. Roweis, A. Hertzmann, and B. Frey, "A segmental HMM for speech waveforms," Tech. Rep. UTML Techical Report 2004-001, University of Toronto, Toronto, Canada, 2004.

[72] H. Hermansky and S. Sharma, "TRAPS: classifiers of temporal patterns," in *Proc. of ICSLP*, (Sydney, Australia), pp. 1003–1006, 1998.

[73] V. Tyagi, I. McCowan, H. Bourlard, and H. Misra, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR," in *Proc. of ASRU*, (St. Thomas, US Virgin Islands), 2003.

[74] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, August 1998.

[75] Q. Zhu and A. Alwan, "AM-demodualtion of speech spectra and its application to noise robust speech recognition," in *Proc. of ICSLP*, vol. 1, pp. 341–344, 2000.

[76] B. P. Milner, "Inclusion of temporal information into features for speech recognition," in *Proc. of ICSLP*, 1996.

[77] S. Haykin, *Communication systems*. New York, USA: John Wiley and Sons, 3 ed., 1994.

[78] V. Tyagi and C. Wellekens, "Fepstrum representation of speech," in *Proc. of ASRU*, (Cancun, Mexico), December 2005.

[79] S. Schimmel and L. Atlas, "Coherent envelope detection for modulation filtering of speech," in *Proc. of ICASSP*, (Philadephia, USA), 2005.

[80] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of IEEE*, vol. 63, April 1975.

[81] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech appications," *J. Acoust. Soc. Am*, vol. 105, March 1999.

[82] R. Kumaresan, "An inverse signal approach to computing the envelope of a real valued signal," *IEEE Signal Processing Letters*, vol. 5, October 1998.

[83] M. Athineos and D. Ellis, "Frequency domain linear prediction for temporal features," in *Proc. of ASRU*, (St. Thomas,US Virgin Islands, USA), December 2003.

[84] J.-C. Junqua and J.-P. Haton, *Robustness in automatic speech recognition*. Kluwer, 1996.

[85] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol. 27(2), 1979.

[86] M. Berouti, R.Schwartz, and J.Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. of ICASSP*, 1979.

[87] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden markov models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, 1992.

[88] V.Tyagi and C.J.Wellekens, "Least squares filtering of speech signals for robust asr," in *Proc. of MLMI*, 2005.

[89] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80(10), 1992.

[90] H. Liao and M. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Proc. of Interspeech*, 2005.

[91] L. Deng, J.Droppo, and A. Acero, "Recursive estimation of non-stationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. SAP*, vol. 11, 2003.

[92] L. Deng, J.Droppo, and A. Acero, "Dynamic compensation of hmm variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. SAP*, vol. 13, 2005.

[93] T. Kristjansson, *Speech Recognition in adverse environments: a probabilistic approach*. PhD thesis, University of Waterloo, Canada, 2002.

[94] L. Fissore, P. Laface, G. Micca, and G. Sperto, "Channel adaptation for a continuous speech recognizer," in *Proc. of ICSLP*, 1992.

[95] A.Anastasakos, F. Kubala, J.Makhoul, and R. Schwartz, "Adaptation to new microphones using tied-mixture normalization," in *Proc. of ICASSP*, 1994.

[96] M. K. Omar and M. Hasegawa-Johnson, "Maximum mutual information based acoustic features representation of phonological features for speech recognition," in *Proc. of ICASSP*, (Montreal, Canada), 2002.

[97] M. K. Omar, K. Chena, M. Hasegawa-Johnson, and Y. Bradman, "An evaluation on using mutual information for selection of acoustic features representation of phonemes for speech recognition," in *Proc. of ICSLP*, (Denver, CO), pp. 2129–2132, 2002.

[98] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. of ICASSP*, vol. I, (Philadelphia, PA), pp. 457–460, 2005.

[99] H. Bourlard and D. Dupont, "Sub-band based speech recognition," in *Proc. of ICASSP*, (Munich Germany), pp. 1251–1254, April, 1997.

[100] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," in *Proc. of ICASSP*, (Munich Germany), pp. 1255–1258, 1997.

[101] M. J. Tomlinson, M. J. Russell, R. K. Moore, A. P. Buckland, and M. A. Fawley, "Modelling asynchrony in speech using elementary single-signal decomposition," in *Proc. of ICASSP*, (Munich Germany), pp. 1247–1250, 1997.

[102] D. L. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition feature," in *Proc. of ICASSP*, vol. 1, (Seattle, WA), pp. 21 – 24, 1998.

[103] A. Zolnay, R. Schlüter, and H. Ney, "Robust speech recognition using a voiced-unvoiced feature," in *Proc. of ICSLP*, vol. 2, (Denver, CO), pp. 1065 – 1068, 2002.

[104] M. Graciarena, H. France, J. Zheng, D. Vergyri, and A. Stolcke, "Voicing feature integration in SRI's DECIPHE LVCSR system," in *Proc. of ICASSP*, (Montreal, Canada), 2004.

[105] T. A. Stephenson, M. M. Doss, and H. Bourlard, "Speech recognition with auxiliary information," *IEEE Trans. Speech Audio Process.*, vol. SAP-12, no. 3, pp. 189–203, 2004.

[106] D. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Proc. of ICASSP*, (Salt Lake City, USA), May 2001.

[107] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP features in LVCSR," in *Proc. of ICSLP*, (Jeju Island, Korea), 2004.

[108] N. Morgan, B. Chen, Q. Zhu, and A. Stolcke, "TRAPping conversational speech: extending trap/tandem approaches to conversational telephone speech recognition," in *Proc. of ICASSP*, (Montreal, Canada), 2004.

[109] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with gabor feature extraction," in *Proc. of ICSLP*, (Denver, Colorado), pp. 25–28, 2002.

[110] E. Eide, "Distinctive features for use in automatic speech recognition," in *Proc. of Eurospeech*, (Aalborg, Denmark), pp. 1613–1616, september 2001.

[111] D. Litman, J. Hirschberg, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 43, no. 1-2, pp. 155–175, 2004.

[112] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. of ICASSP*, (San Francisco), pp. 13–16, 1992.

[113] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283–297, 1998.

[114] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, vol. 34, no. 3, pp. 287–310, 2001.

[115] K. Fukunaga, *Introduction to statistical pattern recognition*. New York: Academic Press, 1972.

[116] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*. New York: Wiley, 1973.

[117] M. Loog and R. P. W. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, pp. 732–739, June 2004.

[118] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. of ICASSP*, 1998.

[119] M. J. F. Gales, "Semi-tied covariance matrices," in *Proc. of ICASSP*, 1998.

[120] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," in *icassp*, pp. 1129–1132, jun 2000.

[121] B. Zhang and S. Matsoukas, "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," in *Proc. of ICASSP*, vol. 1, pp. 925–928, Mar. 2005.

[122] M. Padmanabhan and S. Dharanipragada, "Maximizing information content in feature extraction," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 512–519, July 2005.

[123] Y. H. Abdel-Haleem, S. Renals, and N. D. Lawrence, "Acoustic space dimensionality selection and combination using the maximum entropy principle," in *Proc. of ICASSP*, (Montreal, Canada), May 2004.

[124] R. Moore, "Signal decomposition using markov modeling techniques," Tech. Rep. Memo no 3931, Royal Signal and Radar Establishment, Malvern, Worcs, UK, 1986.

[125] A.Varga and R.K.Moore, "Hidden markov decomposition of speech and noise," in *Proc. of ICASSP*, 1990.

[126] M. Gales and S.Young, "An improved approach to the hidden markov model decomposition of speech and noise," in *Proc. of ICASSP*, 1992.

[127] M.J.F.Gales and S.Young, "Cepstral parameter compensation for hmm recognition," *Speech Communication*, vol. 12(3), 1993.

[128] M. Gales, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer, Speech and Language*, vol. 9, 1998.

[129] P. Nguyen, R. Kuhn, J.-C. Junqua, N. Niedzielski, and C. Wellekens, "Eigenvoices : a compact representation of speakers in a model space," *Annales des Télécommunications*, vol. 55, March-April 2000.

[130] M. J. F. Gales, "Cluster adaptive training for speech recognition," in *Proc. of ICSLP*, pp. 1783–1786, 1998.

[131] D. K. Kim and N. S. Kim, "Rapid online adaptation using speaker space model evolution," *Speech Communication*, vol. 42, pp. 467–478, Apr. 2004.

[132] X. Wu and Y. Yan, "Speaker adaptation using constrained transformation," *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 168–174, Mar. 2004.

[133] B. Mak and R. Hsiao, "Improving eigenspace-based mllr adaptation by kernel PCA," in *Proc. of ICSLP*, (Jeju Island, Korea), Sept. 2004.

[134] B. Zhou and J. Hansen, "Rapid discriminative acoustic model based on eigenspace mapping for fast speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 554–564, July 2005.

[135] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 345–354, May 2005.

[136] S. Tsakalidis, V. Doumpiotis, and W. Byrne, "Discriminative linear transforms for feature normalization and speaker adaptation in HMM estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 367–376, May 2005.

[137] Y. Tsao, S.-M. Lee, and L.-S. Lee, "Segmental eigenvoice with delicate eigenspace for improved speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 399–411, May 2005.

[138] M. Padmanabhan and S. Dharanipragada, "Maximum-likelihood nonlinear transformation for acoustic adaptation," *IEEE Trans. Speech Audio Process.*, vol. 12, pp. 572–578, Nov. 2004.

[139] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin, and K. Bartkova, "Towards improving ASR robustness for PSN and GSM telephone applications," in *Speech Communication*, pp. 141–159, Oct. 1997.

[140] S. Das, D.Lubensky, and C. Wu, "Towards robust speech recognition in the telephony network environment - cellular and landline conditions," in *Proc. of Eurospeech*, pp. 1959–1962, 1999.

[141] Y.Konig and N.Morgan, "GDNN: a gender-dependent neural network for continuous speech recognition," in *Proc. of Int. Joint Conf. on Neural Networks*, pp. 332 – 337, June 1992.

[142] J. O. P.C. Woodland and, V. Valtchev, and S. Young, "Large vocabulary continuous speech recognition using HTK," in *Proc. of ICASSP*, pp. 125–128, Apr. 1994.

[143] C.-H. Lee and J.-L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," in *Proc. of ICASSP*, pp. 558–561, Apr. 1993.

[144] S. Gupta, F. Soong, and R. Haimi-Cohen, "High-accuracy connected digit recognition for mobile applications," in *Proc. of ICASSP*, pp. 57–60, May 1996.

[145] S. M. D. Arcy, L. P. Wong, and M. J. Russell, "Recognition of read and spontaneous children's speech using two new corpora," in *Proc. of ICSLP*, (Jeju Island, Korea), Sept. 2004.

[146] T. Pfau and G. Ruske, "Creating Hidden Markov Models for fast speech," in *Proc. of ICSLP*, p. 0255, 1998.

[147] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," in *Proc. of ICASSP*, pp. 725–728, 2002.

[148] C. Chesta, P. Laface, and F. Ravera, "Connected digit recognition using short and long duration models," in *Proc. of ICASSP*, pp. 557–560, Mar. 1999.

[149] J. Zheng, H. Franco, and A. Stolcke, "Effective acoustic modeling for rate-of-speech variation in large vocabulary conversational speech recognition," in *Proc. of ICSLP*, (Jeju Island, Korea), pp. 401–404, Sept. 2004.

[150] M. G. Song, H. Jung, K.-J. Shim, and H. S. Kim, "Speech recognition in car noise environments using multiple models according to noise masking levels," in *Proc. of ICSLP*, p. 1065, 1998.

[151] S. Sakauchi, Y. Yamaguchi, S. Takahashi, and S. Kobashikawa, "Robust speech recognition based on hmm composition and modified wiener filter," in *Proc. of Interspeech*, (Jeju Island, Korea), pp. 2053–2056, 2004.

[152] L. Rabiner, C. Lee, B. Juang, and J. Wilpon, "HMM clustering for connected word recognition," in *Proc. of ICASSP*, pp. 405–408, May 1989.

[153] F. Korkmazskiy, B.-H. Juang, and F. Soong, "Generalized mixture of HMMs for continuous speech recognition," in *Proc. of ICASSP*, pp. 144 –1446, Apr. 1997.

[154] T. Shinozaki and S. Furui, "Spontaneous speech recognition using a massively parallel decoder," in *Proc. of ICSLP*, (Jeju Island, Korea), pp. 1705–1708, Sept. 2004.

[155] C. Neti and S. Roukos, "Phone-context specific gender-dependent acoustic-models for continuous speech recognition," in *Proc. of ASRU*, pp. 192–198, Dec. 1997.

[156] D. Paul, "Extensions to phone-state decision-tree clustering: single tree and tagged clustering," in *Proc. of ICASSP*, pp. 1487–1490, Apr. 1997.

[157] H. Suzuki, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "Speech recognition using voice-characteristic-dependent acoustic models," in *Proc. of ICASSP*, pp. 740–743, Apr. 2003.

[158] F. Brugnara, R. D. Mori, D. Giuliani, and M. Omologo, "A family of parallel Hidden Markov Models," in *Proc. of ICASSP*, pp. 377–380, Mar. 1992.

[159] T. Shinozaki and S. Furui, "Hidden mode HMM using bayesian network for modeling speaking rate fluctuation," in *Proc. of ASRU*, (US Virgin Islands), pp. 417–422, Dec. 2003.

[160] F. Korkmazsky, M. Deviren, D. Fohr, and I. Illina, "Hidden factor dynamic bayesian networks for speech recognition," in *Proc. of ICSLP*, (Jeju Island, Korea), Sept. 2004.

[161] S. Matsuda, T. Jitsuhiro, K. Markov, and S. Nakamura, "Speeh recognition system robust to noise and speaking styles," in *Proc. of ICSLP*, (Jeju Island, Korea), Sept. 2004.

[162] Y. Zhang, C. Desilva, A. Togneri, M. Alder, and Y. Attikiouzel, "Speaker-independent isolated word recognition using multiple Hidden Markov Models," in *Proc. IEE Vision, Image and Signal Processing*, pp. 197–202, June 1994.

[163] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," in *Proc. of ASRU*, pp. 347–354, Dec. 1997.

[164] L. Barrault, R. de Mori, R. Gemello, F. Mana, and D. Matrouf, "Variability of automatic speech recognition systems using different features," in *Proc. of Interspeech*, (Lisboa, Portugal), pp. 221–224, 2005.

[165] T. Utsuro, T. Harada, H. Nishizaki, and S. Nakagawa, "A confidence measure based on agreement among multiple LVCSR models - correlation between pair of acoustic models and confidence," in *Proc. of ICSLP*, pp. 701–704, 2002.

[166] T. Hain and P. C. Woodland, "Dynamic HMM selection for continuous speech recognition," in *Proc. of Eurospeech*, pp. 1327–1330, 1999.

[167] M. J. F. Gales, "Multiple-cluster adaptive training schemes," in *Proc. of ICASSP*, 2001.

[168] M. J. F. Gales, "Acoustic factorization," in *Proc. of ASRU*, 2001.

[169] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," in *IEEE Trans. on Acoustics, Speech, and Signal Processing*, pp. 201–212, June 1976.

[170] A. M. . L. Mauuary, "Voicing parameter and energy-based speech/non-speech detection for speech recognition in adverse conditions," in *Proc. of Eurospeech*, (Geneva, Switzerland), pp. 3069–3072, Sept. 2003.

[171] D. L. Thomson and R. Chengalvarayan, "Use of voicing features in HMM-based speech recognition," in *Speech Communication*, pp. 197–211, July 2002.

[172] N. Kitaoka, D. Yamada, and S. Nakagawa, "Speaker independent speech recognition using features based on glottal sound source," in *Proc. of ICSLP*, (Denver, USA), pp. 2125–2128, Sept. 2002.

[173] A. Ljolje, "Speech recognition using fundamental frequency and voicing in acoustic modeling," in *Proc. of ICSLP*, (Denver, USA), pp. 2137–2140, Sept. 2002.

[174] W.-J. Yang, J.-C. Lee, Y.-C. Chang, and H.-C. Wang, "Hidden Markov Model for Mandarin lexical tone recognition," in *IEEE Trans. on Acoustics, Speech, and Signal Processing*, pp. 988–992, July 1988.

[175] P.-F. Wong and M.-H. Siu, "Decision tree based tone modeling for chinese speech recognition," in *Proc. of ICASSP*, pp. 905–908, May 2004.

[176] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New methods in continuous mandarin speech recognition," in *Proc. of Eurospeech*, pp. 1543–1546, 1997.

[177] Y. Y. Shi, J. Liu, and R. Liu, "Discriminative HMM stream model for Mandarin digit string speech recognition," in *Proc. of Int. Conf. on Signal Processing*, pp. 528–531, Aug. 2002.

[178] S. Liu, S. Doyle, A. Morris, and F. Ehsam, "The effect of fundamental frequency on mandarin speech recognition," in *Proc. of ICSLP*, (Sydney, Australia), pp. 2647–2650, Nov/Dec 1998.

[179] C.-H. H. Hank and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," in *Proc. of ICASSP*, pp. 1523–1526, June 2000.

[180] T. Demeechai and K. Mäkeläinen, "Recognition of syllables in a tone language," in *Speech Communication*, pp. 241–254, Feb. 2001.

[181] C. Chen, H. Li, L. Shen, and G. Fu, "Recognize tone languages using pitch information on the main vowel of each syllable," in *Proc. of ICASSP*, pp. 61–64, May 2001.

[182] D. O'Shaughnessy and H. Tolba, "Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision," in *Proc. of ICASSP*, pp. 413–416, Mar. 1999.

[183] T. A. Stephenson, H. Bourlard, S. Bengio, and A. C. Morris, "Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables," in *Proc. of ICSLP*, (Beijing, China), pp. 951–954, Oct. 2000.

[184] K. Markov and S. Nakamura, "Hybrid HMM/BN LVCSR system integrating multiple acoustic features," in *Proc. of ICASSP*, pp. 840–843, Apr. 2003.

[185] M. Magimai-Doss, T. A. Stephenson, S. Ikbal, and H. Bourlard, "Modelling auxiliary features in tandem systems," in *Proc. of ICSLP*, (Jeju Island, Korea), Sept. 2004.

[186] H. Singer and S. Sagayama, "Pitch dependent phone modelling for HMM based speech recognition," in *Proc. of ICASSP*, pp. 273–276, Mar. 1992.

[187] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression Hidden Markov Model," in *Proc. of ICASSP*, (Salt Lake City, USA), pp. 513–516, May 2001.