Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 763–786
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Automatic speech recognition and speech variability: A review

M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont *, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens

Multitel, Parc Initialis, Avenue Copernic, B-7000 Mons, Belgium

Received 14 April 2006; received in revised form 30 January 2007; accepted 6 February 2007

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 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. Also, some applications, like directory assistance, particularly stress the core recognition technology due to the very high active vocabulary (application perplexity). There are actually many factors affecting the speech realization: regional, sociolinguistic, or related to the environment or the speaker herself. These create a wide range of variations that may not be modeled correctly (speaker, gender, speaking rate, vocal effort, regional accent, speaking style, non-stationarity, etc.), especially when resources for system training are scarce. This paper outlines current advances related to these topics.

Keywords: Speech recognition; Speech analysis; Speech modeling; Speech intrinsic variations

1. Introduction

It is well-known that 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, his identity. Beside intra-speaker variability (emotion, health, age), it is also commonly admitted that the speaker uniqueness results from a complex combination of physiological and cultural aspects (Garvin and Ladefoged, 1963; Nolan, 1983).

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 source signal, which is the glottal signal for voiced sounds, in which case it affects the pitch, the configuration of the speech articulators (tongue, jaw, 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 front-ends analyze short signal frames (typically covering 30 ms of speech) on which stationarity is assumed. Also, more subtle signal analysis techniques are being studied in the framework of ASR.

The effects of coarticulation have motivated studies on segment based, articulatory, 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 phonemes, going to the configuration of the following phonemes (coarticulation effects may indeed stretch over more than one phoneme). In different and more relaxed speaking styles, stronger

* Corresponding author. Tel.: +32 65 374770; fax: +32 65 374729.
E-mail address: dupont@multitel.be (S. Dupont).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.02.006

pronunciation effects may appear, often lead to reduced articulation. 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 modeled appropriately by CD modeling techniques, their impact may be more simply characterized 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 feature 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, as reported in Lombard (1911).

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 result 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 poor usability for instance).

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

This paper reviews current advances related to these topics. It focuses on variations within the speech signal that make the ASR task difficult. These variations are intrinsic to the speech signal and affect the different levels of the ASR processing chain. For different causes of speech variation, the paper summarizes the current literature and highlights specific feature extraction or modeling weaknesses.

The paper is organized as follows. In a first section, 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.

Methodologies that can help analyzing and diagnose the weaknesses of ASR technology can also be useful. These diagnosis methodologies are the object of Section 3. A specific methodology consists in performing comparisons between man and machine recognition. This provides an absolute reference point and a methodology that can help pinpointing the level of interest. Man–machine comparison also strengthens interdisciplinary insights from fields such as audiology and speech technology.

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 variation is a major factor of reduced performance (in the case of accented and spontaneous speech). Section 4 reviews ongoing trends and possible breakthroughs in general feature extraction and modeling techniques that provides more resistance to speech production variability. 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 surveyed.

2. Speech variability sources

Prior to reviewing the most important causes of intrinsic variation of speech, it is interesting to briefly look into the effects. Indeed, improving ASR systems regarding sources of variability will mostly be a matter of counteracting the effects. Consequently, it is likely that most of the *variability-proof* ASR techniques actually address several causes that produce similar modifications of the speech.

We can roughly consider three main classes of effects; first, the fine structure of the voice signal is affected, the color and the quality of the voice are modified by physiological or behavioral factors. The individual physical characteristics, the smoking habit, a disease, the environmental context that make you soften your voice or, on the contrary, tense it, etc. are such factors. Second, the long-term modulation of the voice may be modified, intentionally – to transmit high level information such as emphasizing or questioning – or not-to convey emotions. This effect is an integral part of the human communication and is therefore very important. Third, the word pronunciation is altered. The acoustic realization in terms of the core spoken language components, the phonemes, may be deeply affected, going from variations due to coarticulation, to substitutions (accents) or suppressions (spontaneous speech).

As we will further observe in the following sections, some variability sources can hence have multiple effects, and several variability sources obviously produce effects that belong to the same category. For instance, foreign accents, speaking style, rate of speech, or children speech all cause pronunciation alterations with respect to the "standard form". The actual alterations that are produced are however dependent on the source of variability, and on the different factors that characterize it.

Although this is outside the scope of this paper, we should add a fourth class of effects that concerns the grammatical and semantic structure of the language. Sociological factors, partial knowledge of the language (non-nativeness, childhood, etc.), may lead to important deviations from the canonical language structure.

2.1. Foreign and regional accents

While investigating the variability between speakers through statistical analysis methods, Huang et al. (2001) found that the first two principal components of variation correspond to the gender (and related to physiological properties) and accent respectively. Indeed, compared to native speech recognition, performance degrades when recognizing accented speech and non-native speech (Kubala et al., 1994; Lawson et al., 2003). In fact accented speech is associated with a shift within the feature space (VanCompernolle, 2001). Good classification results between regional accents are reported in Draxler and Burger (1997) for human listeners on German SpeechDat data, and in Lin and Simske (2004) for automatic classification between American and British accents which demonstrates that regional variants correspond to significantly different data. 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. In contrast, 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.

Non-native speech recognition is not properly handled by speech models estimated using native speech data. This issue remains no matter how much dialect data is included in the training (Beattie et al., 1995). 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 sound considered as the closest in their native language phoneme inventory (Flege et al., 2003). This behavior 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 behavior cannot be handled by the usual triphone-based modeling (Jurafsky et al., 2001).

Accent classification is also studied since many years (Arslan and Hansen, 1996), based either on phone models (Kumpf and King, 1996; Teixeira et al., 1996) or specific acoustic features (Fung and Liu, 1999).

Speech recognition technology is also used in foreign language learning for rating the quality of the pronunciation (Eskenazi, 1996; Franco et al., 2000; Neumeyer et al., 1996; Townshend et al., 1998). Experiments showed that the provided rating is correlated with human expert ratings (Cucchiarini et al., 2000; Neumeyer et al., 2000; Witt and Young, 2000) when sufficient amount of speech is available.

Proper and foreign name processing is another topic strongly related with foreign accent. Indeed, even if speakers are not experts in all foreign languages, neither are they linguistically naive, hence they may use different systems or sub-systems of rules to pronounce unknown names which they perceive to be non-native (Fitt, 1995). Foreign names are hard to pronounce for speakers who are not familiar with the names and there are no standardized methods for pronouncing proper names (Gao et al., 2001). Native phoneme inventories are enlarged with some phonemes of foreign languages in usual pronunciations of foreign names, especially in some languages (Eklund and Lindström, 2001). Determining the ethnic origin of a word improves pronunciation models (Llitjos and Black, 2001) and is useful in predicting additional pronunciation variants (Bartkova, 2003; Maison, 2003).

2.2. Speaker physiology

Beside the regional origin, another speaker-dependent property that is conveyed through the speech signal results from the shape of the vocal apparatus which determines the range within which the parameters of a particular speaker's voice may vary. From this point of view, a very detailed study of the speech-speaker dichotomy can be found in Mokhtari (1998).

The impact of inter-speaker variability on the automatic speech recognition performance has been acknowledged for years. In Huang and Lee (1991), Lee et al. (1991), Schwartz et al. (1989), the authors mention error rates two to three times higher for speaker-independent ASR systems compared with speaker-dependent systems. Methods that aims at reducing this gap in performance are now part of state-of-the-art commercial ASR systems.

Speech production can be modeled by the so-called source-filter model (Fant, 1960) where the "source" refers to the air stream generated by the lungs through the larynx and the "filter" refers to the vocal tract, which is composed of the different cavities situated between the glottis and the lips. Both of the components are inherently time-varying and assumed to be independent of each other.

The complex shape of the vocal organs determines the unique "timbre" of every speaker. The glottis at the larynx is the source for voiced phonemes and shapes the speech signal in a speaker characteristic way. Aside from the long-term F0 statistics (Carey et al., 1996; Iivonen et al., 2003; Markel et al., 1977) which are probably the most perceptually relevant parameters (the pitch), the shape of glottal pulse will affect the long-term overall shape of the

power spectrum (spectral tilt) (Nolan, 1983) and the tension of vocal folds will affect the voice quality. The vocal tract, can be modeled by a tube resonator (Fant, 1960; Laver, 1994). 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.

Modeling of the glottal flow is a difficult problem and very few studies attempt to precisely decouple the source-tract components of the speech signal (Blomberg, 1991; Bozkurt et al., 2005; Plumpe et al., 1999). Standard feature extraction methods (PLP, MFCC) simply ignore the pitch component and roughly compensate for the spectral tilt by applying a pre-emphasis filter prior to spectral analysis or by applying band-pass filtering in the cepstral domain (the cepstral liftering) (Juang et al., 1987).

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 (Peterson and Barney, 1952; Pols et al., 1969; Potter and Steinberg, 1950) 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 (Pols et al., 1969; Saito and Itakura, 1983). Another important observation (Ladefoged and Broadbent, 1957; Nearey, 1978; Peterson and Barney, 1952; Potter and Steinberg, 1950) 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 (Flanagan, 1972; O'Saughnessy, 1987). This observation is at the root of different techniques that increase the robustness of ASR systems to inter-speaker variability (cf. 4.1.2 and 4.2.1).

2.3. Speaking style and spontaneous speech

In spontaneous casual speech, or under time pressure, reduction of pronunciations of certain phonemes, or syllables often happen. It has been suggested that this "slurring" affects more strongly sections that convey less information. In contrast, speech portions where confusability (given phonetic, syntactic and semantic cues) is higher tend to be articulated more carefully, or even hyperarticulated. Some references to such studies can be found in Bard et al. (2001), Janse (2004), Lindblom (1990), Sotillo and Bard (1998), and possible implications to ASR in Bell et al. (2003).

This dependency of casual speech slurring on identified factors holds some promises for improving recognition of spontaneous speech, possibly by further extending the context dependency of phonemes to measures of such perplexity, with however very few research ongoing to our knowledge, except maybe in the use of phonetic transcription for multi-word compounds or user formulation (Colibro et al., 2005) (cf. 4.3).

Research on spontaneous speech modeling is nevertheless very active. Several studies have been carried out on using the Switchboard spontaneous conversations corpus. An appealing methodology has been proposed in Weintraub et al. (1996), where a comparison of ASR accuracy on the original Switchboard test data and on a reread version of it is proposed. Using modeling methodologies that had been developed for read speech recognition, the error rate obtained on the original corpus was twice the error rate observed on the read data.

Techniques to increase accuracy towards spontaneous speech have mostly focused on pronunciation studies.[1] As a fundamental observation, the strong dependency of pronunciation phenomena with respect to the syllable structure has been highlighted in Adda-Decker et al. (2005), Greenberg and Chang (2000). 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. This class of approaches is sometimes referred to as long-units (Messina and Jouvet, 2004).

Variations in spontaneous speech can also extend beyond the typical phonological alterations outlined previously. Disfluencies, such as false starts, repetitions, hesitations and filled pauses, need to be considered. The reader will find useful information in the following papers: (Byrne et al., 2004; Furui et al., 2004).

There are also regular workshops specifically addressing the research activities related to spontaneous speech modeling and recognition (Disfluency in spontaneous speech (diss'05), 2005). Regarding the topic of pronunciation variation, the reader should also refer to (ESCA, 1998).

2.4. Rate of speech

Rate-of-speech (ROS) is considered as an important factor which makes the mapping process between the acoustic signal and the phonetic categories more complex.

Timing and acoustic realization of syllables are affected due in part to the limitations of the articulatory machinery, which may affect pronunciation through phoneme reductions (typical to fast spontaneous speech), time compression/expansion, changes in the temporal patterns, as well as smaller-scale acoustic–phonetic phenomena.

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

In Janse (2004), production studies on normal and fast-rate speech are reported. They have roughly quantified the way people compress some syllables more than others. Note also that the study reports on a series of experiments investigating how speakers produce and listeners perceive fast speech. The main research question is how the perception of naturally produced fast speech compares to the perception of artificially time-compressed speech, in terms of intelligibility.

Several studies also reported that different phonemes are affected differently by ROS. For example, compared to consonants, the duration of vowels is significantly more reduced from slow to fast speech (Kuwabara, 1997).

The relationship between speaking rate variation and different acoustic correlates are usually not well taken into account in the modeling of speech rate variation for automatic speech recognition, where it is typical that the higher the speaking rate is, the higher the error rate is. Usually, slow speaking rate does not affect performance; however, when people hyperarticulate, and make pauses among syllables, speech recognition performance can also degrade a lot.

In automatic speech recognition, the significant performance degradations (Martinez et al., 1997; Mirghafori et al., 1995; Siegler and Stern, 1995) caused by speaking rate variations stimulated many studies for modeling the spectral effects of speaking rate variations. The schemes presented in the literature generally make use of ROS (rate of speech) estimators. Almost all existing ROS measures are based on the same principle which is how to compute the number of linguistic units (usually phonemes or syllables) in the utterance. So, usually, a speaking rate measure based on manually segmented phones or syllables is used as a reference to evaluate a new ROS measure. Current ROS measures can be divided into (1) *lexically-based measures* and (2) *acoustically-based measures*. The *lexically-based measures* estimate the ROS by counting the number of linguistic units per second using the inverse of mean duration (Siegler and Stern, 1995), or *mean* of *m* (Mirghafori et al., 1995). To reduce the dependency on the phone type, a normalization scheme by the expected phone duration (Martinez et al., 1997) or the use of phone duration percentile (Siegler, 1995) are introduced. These kinds of measures are effective if the segmentation of the speech signal provided by a speech recognizer is reliable. In practice this is not the case since the recognizer is usually trained with normal speech. As an alternative technique, acoustically-based measures are proposed. These measures estimate the ROS directly from the speech signal without recourse to a preliminary segmentation of the utterance. In Morgan and Fosler-Lussier (1998), the authors proposed the *mrate* measure (short for *multiple rate*). It combines three independent ROS measures, i.e., (1) the energy rate or *enrate* (Morgan et al., 1997), (2) a simple peak counting algorithm performed on the wideband energy envelope and (3) a sub-band based module that computes a trajectory that is the average product over all pairs of compressed sub-band energy trajectories. A modified version of the *mrate* is also proposed in Beauford (1999). In Tuerk and Young (1999), the authors found that successive feature vectors are more dependent (correlated) for slow speech than for fast speech. An Euclidean distance is used to estimate this dependency and to discriminate between slow and fast speech. In Falthauser et al. (2000), speaking rate dependent GMMs are used to classify speech spurts into slow, medium and fast speech. The output likelihoods of these GMMs are used as input to a neural network whose targets are the actual phonemes. The authors made the assumption that ROS does not affect the temporal dependencies in speech, which might not be true.

It has been shown that speaking rate can also have a dramatic impact on the degree of variation in pronunciation (Fosler-Lussier and Morgan, 1999; Greenberg and Fosler-Lussier, 2000), for the presence of deletions, insertions, and coarticulation effects.

In Section 4, different technical approaches to reduce the impact of the speaking rate on the ASR performance are discussed. They basically all rely on a good estimation of the ROS. Practically, since fast speech and slow speech have different effects (for example fast speech increases deletion as well as substitution errors and slow speech increases insertion errors (Martinez et al., 1997; Nanjo and Kawahara, 2004)), several ROS estimation measures are combined in order to use appropriate compensation techniques.

2.5. Children speech

Children automatic speech recognition is still a difficult problem for conventional automatic speech recognition systems. Children speech represents an important and still poorly understood area in the field of computer speech recognition. The impact of children voices on the performance of standard ASR systems is illustrated in Elenius and Blomberg (2004), Hagen et al., 2003, Traunmüller (1997). The first one is mostly related to physical size. Children have shorter vocal tract and vocal folds compared to adults. This results in higher positions of formants and fundamental frequency. The high fundamental frequency is reflected in 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 these effects, previous studies have focused on the acoustic analysis of children speech (Lee et al., 1999; Potamianos et al., 1997). This work demonstrates the challenges faced by speech recognition systems 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. 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 (Das et al., 1998; Giuliani and Gerosa, 2003; Potamianos and Narayanan, 2003; Potamianos et al., 1997). Such Approaches included vocal tract length normalization (VTLN) (Das et al., 1998) as well as spectral normalization (Lee and Rose, 1996).

A second problem is that younger children may not have a correct pronunciation. Sometimes they have not yet learned how to articulate specific phonemes (Schötz, 2001). Finally, a third source of difficulty is linked to the way children are using language. The vocabulary is smaller but may also contain words that do not 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 to the second and third source of difficulty have been proposed, modification of the pronunciation dictionary, and the use of language models which are customized for children speech have all been tried. In Eskenazi and Pelton (2002), the number of tied-states of a speech recognizer was reduced to compensate for data sparsity. Recognition experiments using acoustic models trained from adult speech and tested against speech from children of various ages clearly show performance degradation with decreasing age. On average, the word error rates are two to five times worse for children speech than for adult speech. Various techniques for improving ASR performance on children speech are reported.

Although several techniques have been proposed to improve the accuracy of ASR systems on children voices, a large shortfall in performance for children relative to adults remains. Eskenazi (1996), Wilpon and Jacobsen (1996) report ASR performance to be around 100% higher, in average, for children speech than for adults. The difference increases with decreasing age. Many papers report a larger variation in recognition accuracy among children, possibly due to their larger variability in pronunciation. Most of these studies point to lack of children acoustic data and resources to estimate speech recognition parameters relative to the abundance of existing resources for adult speech recognition.

2.6. Emotional state

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 (Cowie and Cornelius, 2003; Scherer, 2003).

Studies on speaker emotions is a fairly recent, emerging field and most of today's literature that remotely deals with emotions in speech recognition is concentrated on attempting to classify a "stressed" or "frustrated" speech signal into its correct emotion category (Ang et al., 2002). The purpose of these efforts is to further improve man–machine communication. Being interested in speech intrinsic variabilities, we will rather focus our attention on the recognition of speech produced in different emotional states. The stressed speech categories studied generally are a collection of all the previously described intrinsic variabilities: loud, soft, Lombard, fast, angry, scared, and noise. Nevertheless, note that emotion recognition might play a role, for instance in a framework where the system could select during operation the most appropriate model in an ensemble of more specific acoustic models (cf. Section 4.2.2).

As Hansen formulates in Hansen (1996), 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. As an example, the phonetic features have been examined in the case of task stress or emotion (Bou-Ghazale and Hansen, 1995; Hansen, 1989; Hansen, 1993; Hansen, 1995; Murray and Arnott, 1993). The robust ASR approaches are covered by Section 4.

2.7. And more...

Many more sources of variability affect the speech signal and this paper can probably not cover all of them. Let us cite pathologies affecting the larynx or the lungs, or even the discourse (dysphasia, stuttering, cerebral vascular accident, etc.), long-term habits as smoking, singing, etc., speaking styles like whispering, shouting, etc. physical activity causing breathlessness, fatigue, etc.

The impact of those factors on the ASR performance has been little studied and very few papers have been published that specifically address them.

3. ASR diagnosis

3.1. ASR performance analysis and diagnosis

When devising a novel technique for automatic speech recognition, the goal is to obtain a system whose ASR performance on a specific task will be superior to that of existing methods.

The mainstream aim is to formulate an objective measure for the comparison of a novel system to either similar ASR systems, or humans (cf. Section 3.2). For this purpose, the general evaluation is the word error rate, measuring the global incorrect word recognition in the total recognition task. As an alternative, the error rate is also measured in smaller units such as phonemes or syllables. Further assessments put forward more detailed errors: insertion, deletion and substitution rates.

Besides, detailed studies are found to identify recognition results considering different linguistic or phonetic

properties of the test cases. In such papers, the authors report their systems outcome in the various categories in which they divide the speech samples. The general categories found in the literature are acoustic–phonetic classes, for example: vocal/non-vocal, voiced/unvoiced, nasal/non-nasal (Chollet et al., 1981; Hunt, 2004). Further groupings separate the test cases according to the physical differences of the speakers, such as male/female, children/adult, or accent (Huang et al., 2001). Others, finally, study the linguistic variations in detail and devise more complex categories such as 'VCV' (Vowel–Consonant–Vowel) and 'CVC' (Consonant–Vowel–Consonant) and all such different variations (Greenberg and Chang, 2000). Alternatively, other papers report confidence scores to measure the performance of their recognizers (Zhou et al., 2002; Williams, 1999).

It is however more challenging to find reports on the actual diagnosis of the individual recognizers rather than on the abstract semantics of the recognition sets. In Greenberg and Chang (2000), the authors perform a diagnostic evaluation of several ASR systems on a common database. They provide error patterns for both phoneme- and word-recognition and then present a decision-tree analysis of the errors providing further insight of the factors that cause the systematic recognition errors. Steeneken et al. present their diagnosis method in Steeneken and van Velden (1989) where they establish recognition assessment by manipulating speech, examining the effect of speech input level, noise and frequency shift, on the output of the recognizers. In another approach, Eide et al. display recognition errors as a function of word type and length (Eide et al., 1995). They also provide a method of diagnostic trees to scrutinize the contributions and interactions of error factors in recognition tasks. Alongside, the ANOVA (Analysis of Variance) method (Kajarekar et al., 1999; Kajarekar et al., 1999; Sun and Deng, 1995) allows a quantification of the multiple sources of error acting in the overall variability of the speech signals. It offers the possibility to calculate the relative significance of each source of variability as they affect the recognition. On the other hand, Doddington (2003) introduces time alignment statistics to reveal systematic ASR scoring errors.

The second, subsequent, difficulty is in discovering research that attempts to actually predict the recognition errors rather than simply giving a detailed analysis of the flaws in the ASR systems. This aspect would give us useful insight by providing generalization to unseen test data. Finally, Fosler-Lussier et al. (2005) provides a framework for predicting recognition errors in unseen situations through a collection of lexically confusable words established during training. This work follows former studies on error prediction (Deng et al., 2003; Hirschberg et al., 2004; Printz and Olsen, 2002) and assignment of error liability (Chase, 1997) and is adjacent to the research on confusion networks (Goel et al., 2004; Hetherington, 1995; Mangu et al., 2000; National Institute of Standards and Technology, 2001; Schaaf and Kemp, 1997).

3.2. Man–machine comparison

A few years ago, a publication (Lippmann, 1997) gathered results from both human and machine speech recognition, with the goal of stimulating the discussion on research directions and contributing to the understanding of what has still to be done to reach close-to-human performance. In the reported results, and although problems related to noise can be highlighted, one of the most striking observation concerns the fact that the human listener far outperforms (in relative terms) the machine in tasks characterized by a quiet environment and where no long term grammatical constraints can be used to help disambiguate the speech. This is the case for instance in digits, letters and nonsense sentences where human listeners can in some cases outperform the machine by more than an order of magnitude. We can thus interpret that the gap between machine performance and human performance (10% vs. 1% word error rate on the WSJ large vocabulary continuous speech task in a variety of acoustic conditions) is by a large amount related to acoustico-phonetic aspects. The deficiencies probably come from a combination of factors. First, the feature representations used for ASR may not contain all the useful information for recognition. Then, the modeling assumptions may not be appropriate. Third, the applied features extraction and the modeling approaches may be too sensitive to intrinsic speech variabilities, amongst which are: speaker, gender, age, dialect, accent, health condition, speaking rate, prosody, emotional state, spontaneity, speaking effort, articulation effort.

In Sroka and Braida (2005), consonant recognition within different degradation conditions (high-pass and low-pass filtering, as well as background noise) is compared between human and automatic systems. Results are presented globally in terms of recognition accuracy, and also in more details in terms of confusion matrices as well as information transfer of different phonetic features (voicing, place, frication, sibilance). Although the test material is not degraded in the exact same fashion for the comparison tests, results clearly indicate different patterns of accuracy for human and machines, with weaker machine performance on recognizing some phonological features, such as voicing, especially under noise conditions. This happens despite the fact that the ASR system training provides acoustic models that are almost perfectly matched to the test conditions, using the same speakers, same material (CVCs) and same conditions (noise added to the training set to match the test condition).

In Wesker et al. (2005) (experiments under way), this line of research is extended with the first controlled comparison of human and machine on speech after removing high-level knowledge (lexical, syntactic, etc.) sources, complementing the analysis of phoneme identification scores with the impact of intrinsic variabilities (rather than high-pass/low-pass filters and noise in the previous literature, etc.) Another goal of the research is to extend the scope of previous research (which was for instance mostly related

to English) and address some procedures that can sometimes be questioned in previous research (for instance the difference of protocols used for human and machine tests).

Besides simple comparisons in the form of human intelligibility versus ASR accuracy, specific experimental designs can also provide some relevant insights in order to pinpoint possible weaknesses (with respect to humans) at different stages of processing of the current ASR recognition chain. This is summarized in the next subsection.

3.2.1. Specific methodologies

Some references are given here, revolving around the issue of feature extraction limitations (in this case the presence or absence of phase information) vs. modeling limitations.

It has been suggested (Demuynck et al., 2004; Leonard, 1984; Peters et al., 1999) that conventional cepstral representation of speech may destroy important information by ignoring the phase (power spectrum estimation) and reducing the spectral resolution (Mel filter bank, LPC, cepstral liftering, etc.).

Phase elimination is justified by some evidence that humans are relatively insensitive to the phase, at least in steady-state contexts, while resolution reduction is mostly motivated by practical modeling limitations. However, natural speech is far from being constituted of steady-state segments. In Liu et al. (1997), the authors clearly demonstrate the importance of the phase information for correctly classifying stop consonants, especially regarding their voicing property. Moreover, in Schroeder and Strube (1986), it is demonstrated that vowel-like sounds can be artificially created from flat spectrum signal by adequately tuning the phase angles of the waveform.

In order to investigate a possible loss of crucial information, reports of different experiments have been surveyed in the literature. In these experiments, humans were asked to recognize speech reconstructed from the conventional ASR acoustic features, hence with no phase information and no fine spectral representation.

Experiments conducted by Leonard and reported by Lippmann (1997), seems to show that ASR acoustic analysis (LPC in that case) has little effect on human recognition, suggesting that most of the ASR weaknesses may come from the acoustic modeling limitations and little from the acoustic analysis (i.e. front-end or feature extraction portion of the ASR system) weaknesses. Those experiments have been carried out on sequences of digits recorded in a quiet environment.

In their study, Demuynck et al. re-synthesized speech from different steps of the MFCC analysis, i.e. power spectrum, Mel spectrum and Mel cepstrum (Demuynck et al., 2004). They come to the conclusion that re-synthesized speech is perfectly intelligible given that an excitation signal based on pitch analysis is used, and that the phase information is not required. They emphasize that their experiments are done on clean speech only.

Experiments conducted by Peters et al. (1999) demonstrate that these conclusions are not correct in case of noisy speech recordings. He suggests that information lost by the conventional acoustic analysis (phase and fine spectral resolution) may become crucial for intelligibility in case of speech distortions (reverberation, environment noise, etc.). These results show that, in noisy environment, the degradation of the speech representation affects the performance of the human recognition almost in the same order as the machine. More particularly, ignoring the phase leads to a severe drop of human performance (from almost perfect recognition to 8.5% sentence error rate) suggesting that the insensitivity of human to the phase is not that true in adverse conditions.

In Paliwal and Alsteris (2003), the authors perform human perception experiments on speech signals reconstructed either from the magnitude spectrum or from the phase spectrum and conclude that phase spectrum contribute as much as amplitude to speech intelligibility if the shape of the analysis window is properly selected.

Finally, experiments achieved at Oldenburg demonstrated that the smearing of the temporal resolution of conventional acoustic features affects human intelligibility for modulation cut-off frequencies lower than 32 Hz on a phoneme recognition task. Also, they conclude that neglecting the phase causes approximately 5% error rate in phoneme recognition of human listeners.

4. ASR techniques

In this section, we review methodologies towards improved ASR analysis/modeling accuracy and robustness against the intrinsic variability of speech. Similar techniques have been proposed to address different sources of speech variation. This section will introduce both the general ideas of these approaches and the specific usage regarding variability sources.

4.1. Front-end techniques

An update on feature extraction front-ends 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 the effects of some other variations, like the speaker physiology (through feature compensation (Welling et al., 2002) or else improved invariance (Mertins and Rademacher, 2005)) and other dimensions of speech variability. Finally, techniques for combining estimation based on different features sets are reviewed. This also involves dimensionality reduction approaches.

4.1.1. Overcoming assumptions

Most of the automatic speech recognition (ASR) acoustic features, such as Mel-frequency cepstral coefficients (MFCC) (Davis and Mermelstein, 1980) or perceptual linear prediction (PLP) coefficients (Hermansky, 1990),

are based on some sort of representation of the smoothed spectral envelope, usually estimated over fixed analysis windows of typically 20 ms–30 ms (Davis and Mermelstein, 1980; Rabiner and Juang, 1993).[2] 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 40 ms–80 ms, while stops and plosive are time-limited by less than 20 ms (Rabiner and Juang, 1993). Therefore, it implies that the spectral analysis based on a fixed size window of 20 ms–30 ms has some limitations, including:

- The frequency resolution obtained for quasi-stationary segments (QSS) longer than 20 ms 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 (Haykin, 1993). Furthermore, 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 Tyagi et al. (2005), the usual assumption is made that the piecewise quasi-stationary segments (QSS) of the speech signal can be modeled by a Gaussian autoregressive (AR) process of a fixed order p as in Andre-Obrecht (1988), Svendsen et al. (1989), Svendsen and Soong (1987). The problem of detecting QSSs is then formulated using a maximum likelihood (ML) criterion, defining a QSS as the longest segment that has most probably been generated by the same AR process.[3]

Another approach is proposed in Atal (1983), which describes 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 and Wickerhauser (1992) have described a minimum entropy basis selection algorithm to achieve the minimum information cost of a signal relative to the designed orthonormal basis. Svendsen and Soong (1987) 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 subword unit design. More recently, Achan et al. (2004) 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 (TRAPs) features (Hermansky and Sharma, 1998), MLPs and several modulation spectrum related techniques (Kingsbury et al., 1998; Milner, 1996; Tyagi et al., 2003; Zhu and Alwan, 2000). 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) (Haykin, 1994). 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 4 KHz), it is necessary to decompose the speech signal into narrow spectral bands. In Tyagi and Wellekens (2005), this approach is opposed to the previous use of the speech modulation spectrum (Kingsbury et al., 1998; Milner, 1996; Tyagi et al., 2003; Zhu and Alwan, 2000) 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 (2005). 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 $a_{\mathrm{LP}}(t)$. The low-pass filtered AM signal $a_{\mathrm{LP}}(t)$ is then multiplied with the original carrier $c(t)$ to obtain a new signal $\tilde{x}(t)$. They show that 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 (Makhoul, 1975). Lately, Kumaresan and Rao (1999), Kumaresan (1998) have proposed to model analytic signals using pole-zero models in the temporal domain. Along similar lines, Athineos and Ellis (2003) have used the dual of the

[2] Note that these widely used ASR front-end techniques make use of frequency scales that are inspired by models of the human auditory system. An interesting critical contribution to this has however been provided in Hunt (1999), where it is concluded that so far, there is little evidence that the study of the human auditory system has contributed to advances in automatic speech recognition.

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

linear prediction in the frequency domain to improve upon the TRAP features.

Another strong assumption that has been addressed in recent papers, concern the worthlessness of the phase for speech intelligibility. We already introduced in Section 3.2.1 the conclusions of several studies that reject this assumption. A few papers have tried to reintroduce the phase information into the ASR systems. In Paliwal and Atal (2003), the authors introduce the instantaneous frequency which is computed from the phase spectrum. Experiments on vowel classification show that these features contain meaningful information. Other authors are proposing features derived from the group delay (Bozkurt and Couvreur, 2005; Hegde et al., 2004; Zhu and Paliwal, 2004) which presents a formant-like structure with a much higher resolution than the power spectrum. As the group delay in inherently very noisy, the approaches proposed by the authors mainly aims at smoothing the estimation. ASR experiments show interesting performance in noisy conditions.

4.1.2. Compensation and invariance

For other sources of speech variability (besides non-stationarity), a simple model may exist that appropriately reflects and compensate its effect on the speech features.

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, etc. Similarly, in Hermansky (1990), the PLP parameters present a fair robustness to inter-speaker variability, thanks to the low order (5th) linear prediction analysis which only models the two main peaks of the spectral shape, typically the first two formants. Other approaches aim at building acoustic features invariant to the frequency warping.

In Umesh et al. (1999), the authors define the "scale transform" and the "scale cepstrum"of a signal spectrum whose magnitude is invariant to a scaled version of the original spectrum. In Mertins and Rademacher (2005), 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 non-linear 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 the sets contain complementary information (Mertins and Rademacher, 2005).

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 (Di Benedetto and Liénard, 1992; Wakita, 1977), transformation of the LPC pole modeling (Slifka and Anderson, 1995), frequency warping, either linear (Eide and Gish, 1996; Lee and Rose, 1996; Tuerk and Robinson, 1993; Zhan and Westphal, 1997) or non-linear (Ono et al., 1993), all consist of modifying the position of the formants in order to get closer to an "average" canonical speaker. Simple yet powerful techniques for normalizing (compensating) the features to the VTL are widely used (Welling et al., 2002). Note that VTLN is often combined with an adaptation of the acoustic model to the canonical speaker (Eide and Gish, 1996; Lee and Rose, 1996) (cf. Section 4.2.1). The potential of using piece-wise linear and phoneme-dependent frequency warping algorithms for reducing the variability in the acoustic feature space of children have also been investigated (Das et al., 1998).

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 (Kajarekar et al., 1999; Westphal, 1997).

Similarly, some studies attempted to devise feature extraction methods tailored for the recognition of stressed and non-stressed speech simultaneously. In his paper (Chen, 1987), Chen proposed a Cepstral Domain Compensation when he showed that simple transformations (shifts and tilts) of the cepstral coefficients occur between the different types of speech signals studied. Further processing techniques have been employed for more robust speech features (Hansen, 1996; Hermansky and Morgan, 1994; Hunt and Lefebvre, 1989) and some researchers simply assessed the better representations from the existing pool of features (Hanson and Applebaum, 1990).

When simple parametric models of the effect of the variability are not appropriate, feature compensation can be performed using more generic non-parametric transformation schemes, including linear and non-linear transformations. This becomes a dual approach to model adaptation, which is the topic of Section 4.2.1.

4.1.3. Additional cues and multiple feature streams

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 features 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 these types of classes as stated in Kamal Omar and Hasegawa-Johnson (2002), Kamal Omar et al. (2002), 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 order to take into account different temporal behavior in different bands, it has been proposed (Bourlard and Dupont, 1997; Tibrewala and Hermansky, 1997; Tomlinson et al., 1997) to consider separate streams of features extracted in separate channels with different frequency bands. Inspired by the multi-stream approach, examples of acoustic measurement combination are:

- Multi-resolution spectral/time correlates (Hariharan et al., 2001; Vaseghi et al., 1997),
- segment and frame-based acoustic features (Hon and Wang, 1999),
- MFCC, PLP and an auditory feature (Jiang and Huang, 1999),
- spectral-based and discriminant features (Benitez et al., 2001),
- acoustic and articulatory features (Kirchhoff, 1998; Tolba et al., 2002),
- LPC based cepstra, MFCC coefficients, PLP coefficients, energies and time-averages (Kamal Omar et al., 2002; Kamal Omar and Hasegawa-Johnson, 2002), MFCC and PLP (Zolnay et al., 2005),
- full band non-compressed root cepstral coefficients (RCC), Full band PLP 16 kHz, Telephone band PLP 8 kHz (Kingsbury et al., 2002),
- PLP, MFCC and wavelet features (Gemello et al., 2006),
- joint features derived from the modified group-delay function (Hegde et al., 2005),
- combinations of frequency filtering (FF), MFCC, RASTA-FF, (J)RASTA-PLP (Pujol et al., 2005).

Other approaches integrate some specific parameters into a single stream of features. Examples of added parameters are:

- periodicity and jitter (Thomson and Chengalvarayan, 1998),
- voicing (Graciarena et al., 2004; Zolnay et al., 2002),
- rate of speech and pitch (Stephenson et al., 2004).

To benefit from the strengths of both MLP–HMM and Gaussian-HMM techniques, the Tandem solution was proposed in Ellis et al. (2001), 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 Zhu et al. (2004), 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 (Morgan et al., 2004). In Kleinschmidt and Gelbart (2002), Gabor filters are proposed, in conjunction with MLP features, to model the characteristics of neurons in the auditory system as is done for 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 Eide (2001), 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 (Hirschberg et al., 2004).

This section introduced several works where several streams of acoustic representations of the speech signal were successfully combined in order to improve the ASR performance. Different combination methods have been proposed and can roughly be classified as:

- direct feature combination/transformation such as PCA, LDA, HDA, etc. or selection of the best features will be discussed in Section 4.1.4;
- combination of acoustic models trained on different feature sets will be discussed in Section 4.2.2.

4.1.4. 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 (Haeb-Umbach and Ney, 1992; Kumar and Andreou, 1998) and/or improved training procedures.

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) (Duda and Hart, 1973; Fukunaga, 1972) 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 (Duda and Hart, 1973; Fukunaga, 1972). 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 (Haeb-Umbach and Ney, 1992). In Haeb-Umbach and

Ney (1992) 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) (Kumar and Andreou, 1998) overcomes this problem, and is actually a generalization of LDA. The method usually requires the use of numerical optimization techniques to find the matrix A. An exception is the method in Loog and Duin (2004), 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) (Gopinath, 1998), which is identical to semi-tied covariance matrices (STC) (Gales, 1999). 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 (cf. (Pitz, 2005), Section 2.1). It has been reported (Saon et al., 2000) that such a combination performs better than LDA or HDA alone. Also, HDA has been combined with minimum phoneme error (MPE) analysis (Zhang and Matsoukas, 2005). 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 (Kamal Omar and Hasegawa-Johnson, 2002; Padmanabhan and Dharanipragada, 2005).

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 Abdel-Haleem et al. (2004), the maximum entropy principle was used to decide on the best feature space.

4.2. 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 (see also Section 4.4). 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 built on knowledge-based or data-driven clusters of data.

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

4.2.1. Adaptation

In Section 4.1.2, we have been reviewing techniques that can be used to compensate for speech variation at the feature extraction level. A 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 application. 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 (Leggetter and Woodland, 1995). A maximum a posteriori (MAP) objective function may also be used (Chesta et al., 1999; Zavaliagkos et al., 1996).

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 "fast" (sometimes on-line) adaptation schemes have been proposed a few years ago, based on the clustering of the speakers into sets of speakers which have similar voice characteristics. Inferred acoustic models present a much smaller variance than speaker-independent systems (Naito et al., 1998; Padmanabhan et al., 1996). The eigenvoice approach (Gales, 1998; Nguyen et al., 2000) takes from this idea by building a low dimension eigenspace in which any speaker is located and modeled as a linear combination of "eigenvoices".

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 Kim and Kim (2004), rapid adaptation is further extended through a more accurate speaker space model, and an on-line algorithm is also proposed. In Wu and Yan (2004), 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 (Kenny et al., 2005; Mak and Hsiao, 2004; Tsakalidis et al., 2005; Tsao et al., 2005; Wu and Yan, 2004; Zhou and Hansen, 2005).

Other forms of transformations for adaptation are also proposed in (Padmanabhan and Dharanipragada, 2004), where the Maximum Likelihood criterion is used but the transformations are allowed to be nonlinear. Let us also mention alternate non-linear speaker adaptation paradigms based on connectionist networks (Abrash et al., 1996; Watrous, 1993).

Speaker normalization algorithms that combine frequency warping and model transformation have been proposed to reduce acoustic variability and significantly improve ASR performance for children speakers (by 25–45% under various model training and testing conditions) (Potamianos and Narayanan, 2003; Potamianos et al., 1997). ASR on emotional speech has also benefited from techniques relying on adapting the model structure within the recognition system to account for the variability in the input signal. One practice has been to bring the training and test conditions closer by space projection (Carlson and Clements, 1992; Mansour and Juang, 1989). In Kubala et al. (1994), it is shown that acoustic model adaptation can be used to reduce the degradation due to non-native dialects. This has been observed on an English read speech recognition task (Wall Street Journal), and the adaptation was applied at the speaker level to obtain speaker dependent models. For speaker independent systems this may not be feasible however, as this would require adaptation data with a large coverage of non-native speech.

4.2.2. Multiple modeling

Instead of adapting the models to particular conditions, one may also train an 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 has to handle various possible conditions, several speech corpora can be used together for estimating the acoustic models, leading to mixed models or hybrid systems (Das et al., 1999; Mokbel et al., 1997), 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 trained from a subset of the training data, defined according to a priori criteria such as gender, accent, 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 (Odell et al., 1994; Konig and Morgan, 1992) or used as a better seed for speaker adaptation (Lee and Gauvain, 1993). Gender dependence is applied to whole word units, for example digits (Gupta et al., 1996), or to context dependent phonetic units (Odell et al., 1994), as a result of an adequate splitting of the training data.

In many cases, most of the regional variants of a language are handled in a blind way through a global training of the speech recognition system using speech data that covers all of these regional variants, and enriched modeling is generally used to handle such variants. This can be achieved through the use of multiple acoustic models associated with large groups of speakers as in Beattie et al. (1995), VanCompernolle et al. (1991). These papers showed that it was preferable to have models only for a small number of large speaker populations than for many small groups. When a single foreign accent is handled, some accented data can be used for training or adapting the acoustic models (Aalburg and Hoege, 2004; He and Zhao, 2003; Liu and Fung, 2000; Uebler and Boros, 1999).

Age dependent modeling has been less investigated, may be due to the lack of large size children speech corpora. The results presented in D'Arcy et al. (2004) 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. Simply training a conventional speech recognizer on children speech is not sufficient to yield high accuracies, as demonstrated by Wilpon and Jacobsen (1996). Recently, corpora for children speech recognition have begun to emerge. In Eskenazi (1996) a small corpus of children speech was collected for use in interactive reading tutors and led to a complete children speech recognition system. In Shobaki et al. (2000), a more extensive corpus consisting of 1100 children, from kindergarten to grade 10, was collected and used to develop a speech recognition system for isolated word and finite state grammar vocabularies for US English.

Speaking rate notably affects the recognition performances, thus ROS dependent models were studied (Mirghafori et al., 1996). It was also noticed that ROS dependent models are often getting less speaker-independent because the range of speaking rate shown by different

speakers is not the same (Pfau and Ruske, 1998), and that training procedures robust to sparse data need to be used. In that sense, comparative studies have shown that rate-adapted models performed better than rate-specific models (Wrede et al., 2001). Speaking rate can be estimated on line (Pfau and Ruske, 1998), 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 (Nanjo and Kawahara, 2002); or the various rate dependent models may be used simultaneously during decoding (Chesta et al., 1999; Zheng et al., 2004).

The 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 Song et al. (1998) 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. In contrast, in Sakauchi et al. (2004) acoustic models composed under various SNR conditions are run in parallel during decoding.

The same way, speech variations due to stress and emotions has been addressed by the multi-style training (Lippmann et al., 1987; Paul, 1987), and simulated stress token generation (Bou-Ghazale and Hansen, 1994, 1995). 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 away from the original training data.

Automatic clustering techniques have also been used for elaborating several models per word for connected-digit recognition (Rabiner et al., 1989). Clustering the trajectories (or sequences of speech observations assigned to some particular segment of the speech, like word or subword units) deliver more accurate modeling for the different groups of speech samples (Korkmazskiy et al., 1997); and clustering training data at the utterance level provided the best performances in Shinozaki and Furui (2004).

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 Neti and Roukos (1997), syllable boundary and stress tags in Paul (1997), and voice characteristics in Suzuki et al. (2003).

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 for parallel hidden Markov models (Brugnara et al., 1992) where the acoustic densities are modulated depending on the probability of a master context HMM being in certain states. In Zolnay et al. (2005), it is shown that log-linear combination provides good results when used for integrating probabilities provided by acoustic models based on different acoustic feature sets. 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 (Shinozaki and Furui, 2003), or hidden factors related to a clustering of the data (Korkmazsky et al., 2004; Matsuda et al., 2004).

Multiple models can also be used in a parallel decoding framework (Zhang et al., 1994); then the final answer results from a "voting" process (Fiscus, 1997), or from the application of elaborated decision rules that take into account the recognized word hypotheses (Barrault et al., 2005). Multiple decoding is also useful for estimating reliable confidence measures (Utsuro et al., 2002).

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 VTL normalization, where a specific estimation of the vocal tract length (VTL) is associated with each speaker of the training data (Welling et al., 2002). 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. The estimation of the VTL factor can either be perform by a maximum likelihood approach (Lee and Rose, 1996; Zhan and Waibel, 1997) or from a direct estimation of the formant positions (Eide and Gish, 1996; Lincoln et al., 1997). More general normalization schemes have also been investigated (Gales, 2001), based on associating transforms (mostly linear transforms) to each speaker, or more generally, to different clusters of the training data. These transforms can also be constrained to reside in an reduced-dimensionality eigenspace (Gales, 1998). A technique for "factoring-in" selected transformations back in the canonical model is also proposed in Gales (2001), 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.

4.2.3. Auxiliary acoustic features

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 inherent in the speech signal 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, voicing and formant parameters have been used since a long time, but mainly for endpoint detection purposes (Atal and Rabiner, 1976) making it much more robust in noisy environments (Martin and Mauuary, 2003). 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, formants

and/or voicing) is their direct introduction in the feature vector, along with the cepstral coefficients, for example periodicity and jitter are used in Thomson and Chengalvarayan (2002) and formant and auditory-based acoustic cues are used together with MFCC in Holmes et al. (1997), Selouani et al. (2002). Correlation between pitch and acoustic features is taken into account in Kitaoka et al. (2002) and an LDA is applied on the full set of features (i.e. energy, MFCC, voicing and pitch) in Ljolje (2002). In de Wet et al. (2004), the authors propose a 2-dimension HMM to extract the formant positions and evaluate their potential on a vowel classification task. In Garner and Holmes (1998), the authors integrate the formant estimations into the HMM formalism, in such a way that multiple formant estimate alternatives weighted by a confidence measure are handled. In Tolba et al. (2003), a multi-stream approach is used to combine MFCC features with formant estimates and a selection of acoustic cues such as acute/grave, open/close, tense/lax, etc.

Pitch has to be taken into account for the recognition of tonal languages. Tone can be modeled separately through specific HMMs (Yang et al., 1988) or decision trees (Wong and Siu, 2004), or the pitch parameter can be included in the feature vector (Chen et al., 1997), or both information streams (acoustic features and tonal features) can be handled directly by the decoder, possibly with different optimized weights (Shi et al., 2002). 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 (Liu et al., 1998), and pitch tracking and voicing are investigated in Huank and Seide (2000). A comparison of various modeling approaches is available in Demeechai and Mäkeläinen (2001). For tonal languages, pitch modeling usually concerns the whole syllable; however, limiting the modeling to the vowel seems sufficient (Chen et al., 2001).

Voicing has been used in the decoder to constrain 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 (O'Shaughnessy and Tolba, 1999).

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 Stephenson et al. (2004) with pitch and energy as auxiliary features. Other information can also be taken into account such as articulatory information in Stephenson et al. (2000) where the DBN utilizes an additional variable for representing the state of the articulators by direct measurement (note that these experiments require a very special X-ray microbeam database). 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 Markov and Nakamura (2003).

Another approach for handling heterogeneous features is the TANDEM approach used with pitch, energy or rate of speech in Magimai-Doss et al. (2004). 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 measured pitch (Singer and Sagayama, 1992), or used to modify the parameters of the densities (during decoding) through multiple regressions as with pitch and speaking rate in Fujinaga et al. (2001).

4.3. Pronunciation modeling techniques

As mentioned in the introduction of Section 2, some speech variations, like foreign accent or spontaneous speech, affect the acoustic realization to the point that their effect may be better described by substitutions and deletion of phonemes with respect to canonical (dictionary) transcriptions.

As a complementary principle to multiple acoustic modeling approaches reviewed in Section 4.2.2, multiple pronunciations are generally used for the vocabulary words. Hidden model sequences offer a possible way of handling multiple realizations of phonemes (Hain and Woodland, 1999) possibly depending on phone context. For handling hyper articulated speech where pauses may be inserted between syllables, ad hoc variants are necessary (Matsuda et al., 2004). And adding more variants is usually required for handling foreign accents.

Modern approaches attempt to build in rules underlying pronunciation variation, using representations frameworks such as FSTs (Hazen et al., 2005; Seneff and Wang, 2005), based on phonological knowledge, data and recent studies on the syllabic structure of speech, for instance in English (Greenberg and Chang, 2000) or French (Adda-Decker et al., 2005).

In Adda-Decker et al. (2005), an experimental study of phoneme and syllable reductions is reported. The study is based on the comparison of canonical and pronounced phoneme sequences, where the latter are obtained through a forced alignment procedure (whereas (Greenberg and Chang, 2000) was based on fully manual phonetic annotation). Although results following this methodology are affected by ASR errors (in addition to "true" pronunciation variants), they present the advantage of being able to benefit from analysis of much larger and diverse speech corpora. In the alignment procedure, the word representations are defined to allow the dropping of any phoneme and/or syllable, in order to avoid limiting the study to pre-defined/already know phenomena. The results are

presented and discussed so as to study the correlation of reduction phenomena with respect to the position of the phoneme in the syllable, the syllable structure and the position of the syllable within the word. Within-word and cross-word resyllabification (frequent in French but not in English) is also addressed. The results reinforce previous studies (Greenberg and Chang, 2000) and suggest further research in the use of more elaborate contexts in the definition of ASR acoustic models. Context-dependent phonemes could be conditioned not only on neighboring phones but also on the contextual factors described in this study. Such approaches are currently being investigated (Lamel and Gauvain, 2005; Messina and Jouvet, 2004). These rely on the modeling capabilities of acoustic models that can implicitly model some pronunciation effect (Dupont et al., 2005; Hain, 2005; Jurafsky et al., 2001), provided that they are represented in the training data. In Hain (2005), several phone sets are defined within the framework of triphone models, in the hope of improving the modeling of pronunciation variants affected by the syllable structure. For instance, an extended phone set that incorporates syllable position is proposed. Experimental results with these novel phone sets are not conclusive however. The good performance of the baseline system could (at least partly) be attributed to implicit modeling, especially when using large amounts of training data resulting in increased generalization capabilities of the used models. Also it should be considered that "continuous" (or "subtle") pronunciation effects are possible (e.g. in spontaneous speech), where pronunciations cannot be attributed to a specific phone from the phone set anymore, but might cover "mixtures" or transitional realizations between different phones. In this case, approaches related to the pronunciation lexicon alone will not be sufficient.

The impact of regional and foreign accents may also be handled through the introduction of detailed pronunciation variants at the phonetic level (Adda-Decker and Lamel, 1999; Humphries et al., 1996). Introducing multiple phonetic transcriptions that handle alterations produced by non-native speakers is a usual approach, and is generally associated with a combination of phone models of the native language with phone models of the target language (Bartkova and Jouvet, 1999; Bonaventura et al., 1998; Witt and Young, 1999). However adding too many systematic pronunciation variants may be harmful (Strik and Cucchiarini, 1999).

Alteration rules can be defined from phonetic knowledge or estimated from some accented data (Livescu and Glass, 2000). Deriving rules using only native speech of both languages is proposed in Goronzy et al. (2004). Raux (2004) investigates the adaptation of the lexicon according to preferred phonetic variants. When dealing with various foreign accents, phone models of several languages can be used simultaneously with the phone models of the target language (Bartkova and Jouvet, 2004), multilingual units can be used (Uebler and Boros, 1999) or specialized models for different speaker groups can be elaborated (Cincarek et al., 2004). Multilingual phone models have been investigated for many years in the hope of achieving language independent units (Bonaventura et al., 1997; Dalsgaard et al., 1998; Köhler, 1996; Schultz and Waibel, 1998). Unfortunately language independent phone models do not provide as good results as language dependent phone models when the latter are trained on enough speech data, but language independent phone models are 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 (Weng et al., 1997) or determined from phonetic knowledge (Uebler, 2001), 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.

As mentioned in Section 2.4, variations of the speaking rate may deeply affect the pronunciation. Regarding this source of variability, some approaches relying upon an explicit modeling strategy using different variants of pronunciation have been proposed; a multi-pass decoding enables the use of a dynamically adjusted lexicon employed in a second pass (Fosler-Lussier and Morgan, 1999). The acoustic changes, such as coarticulation, are modeled by directly adapting the acoustic models (or a subset of their parameters, i.e. weights and transition probabilities) to the different speaking rates (Bard et al., 2001; Martinez et al., 1998; Morgan et al., 1997; Shinozaki and Furui, 2003; Zheng et al., 2000). Most of the approaches are based on a separation of the training material into discrete speaking rate classes, which are then used for the training of rate dependent models. During the decoding, the appropriate set of models is selected according to the measured speaking rate. Similarly, to deal with changes in phone duration, as it is the case for instance for variation of the speaking rate, alteration schemes of the transition probabilities between HMM states are proposed (Martinez et al., 1997; Mirghafori et al., 1995; Morgan et al., 1997). The basic idea is to put high/low transition probability (exit probability) for fast slow/speech. These compensation techniques require *a priori* ROS estimation using one of the measures described in Section 2.4. In Zheng et al. (2000), the authors proposed a compensation technique that does not require ROS estimation. This technique used a set of parallel rate-specific acoustic and pronunciation models. Rate switching is permitted at word boundaries to allow within-sentence speaking rate variation.

The reader should also explore the publications from (ISCA Tutorial and Research Workshop, 2002).

4.4. Larger and diverse training corpora

Driven by the availability of computational resources, there is a still ongoing trend in trying to build bigger and

hopefully better systems, that attempt to take advantage of increasingly large amounts of training data.

This trend seems in part to be related to the perception that overcoming the current limited generalization abilities as well as modeling assumptions should be beneficial. This however implies more accurate modeling whose parameters can only be reliably estimated through larger data sets.

Several studies follow that direction. In Nguyen et al. (2003), 1200 h of training data have been used to develop acoustic models for the English broadcast news recognition task, with significant improvement over the previous 200 h training set. It is also argued that a vast body of speech recognition algorithms and mathematical machinery is aimed at smoothing estimates toward accurate modeling with scant amounts of data.

More recently, in Lamel and Gauvain (2005), up to 2300 h of speech have been used. This has been done as part of the EARS project, where training data of the order of 10,000 h has been put together. It is worth mentioning that the additional very large amounts of training data are usually either untranscribed or automatically transcribed. As a consequence, unsupervised or lightly supervised approaches (e.g. using closed captions) are essential here.

Research towards making use of larger sets of speech data are also involving schemes for training data selection, semi-supervised learning, as well as active learning (Venkataraman et al., 2004). These allow to minimize the manual intervention required while preparing a corpus for model training purposes.[4]

A complementary perspective to making use of more training data consists in using knowledge gathered on speech variations in order to synthesize large amounts of acoustic training data (Girardi et al., 1998).

Finally, another approach is proposed in Dupont et al. (2005), with discriminant non-linear transformations based on MLPs (multi-layer perceptrons) that present some form of genericity across several factors. The transformation parameters are estimated based on a large pooled corpus of several languages, and hence presents unique generalization capabilities. Language and domain specific acoustic models are then built using features transformed accordingly, allowing language and task specificity if required, while also bringing the benefit of detailed modeling and robustness to any tasks and language. A important study of the robustness of similarly obtained MLP-based acoustic features to domains and languages is also reported in Stolcke et al. (2006).

5. Conclusion

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

[4] Tur et al. (2005) combining active and semi-supervised learning for spoken language understanding. Methods of similar inspiration are also used in the framework of training models for spoken language understanding.

Acknowledgements

This review 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

Aalburg, S., Hoege, H., 2004. Foreign-accented speaker-independent speech recognition. In: Proceedings of ICSLP, Jeju Island, Korea, pp. 1465–1468.

Abdel-Haleem, Y.H., Renals, S., Lawrence, N.D., 2004. Acoustic space dimensionality selection and combination using the maximum entropy principle. In: Proceedings of ICASSP, Montreal, Canada, pp. 637–640.

Abrash, V., Sankar, A., Franco, H., Cohen, M., 1996. Acoustic adaptation using nonlinear transformations of HMM parameters. In: Proceedings of ICASSP, Atlanta, GA, pp. 729–732.

Achan, K., Roweis, S., Hertzmann, A., Frey, B., 2004. A segmental HMM for speech waveforms. Technical Report UTML Techical Report 2004-001, University of Toronto, Toronto, Canada.

Adda-Decker, M., Boula de Mareuil, P., Adda, G., Lamel, L., 2005. Investigating syllabic structures and their variation in spontaneous french. Speech Communication 46 (2), 119–139.

Adda-Decker, M., Lamel, L., 1999. Pronunciation variants across system configuration, language and speaking style. Speech Communication 29 (2), 83–98.

Andre-Obrecht, R., 1988. A new statistical approach for the automatic segmentation of continuous speech signals. IEEE Transactions on Acoustics, Speech and Signal Processing 36 (1), 29–40.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human–computer. Proceedings of ICSLP. Denver, Colorado, pp. 2037–2040.

Arslan, L.M., Hansen, J.H.L., 1996. Language accent classification in american english. Speech Communication 18 (4), 353–367.

Atal, B., 1983. Efficient coding of LPC parameters by temporal decomposition. In: Proceedings of ICASSP, Boston, USA, pp. 81–84.

Atal, B., Rabiner, L., 1976. A pattern recognition approach to voiced–unvoiced–silence classification with applications to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 24 (3), 201–212.

Athineos, M., Ellis, D., 2003. Frequency domain linear prediction for temporal features. Proceedings of ASRU. St. Thomas, US Virgin Islands, USA, pp. 261–266.

Bard, E.G., Sotillo, C., Kelly, M.L., Aylett, M.P., 2001. Taking the hit: leaving some lexical competition to be resolved post-lexically. Language and Cognitive Processes 15 (5–6), 731–737.

Barrault, L., de Mori, R., Gemello, R., Mana, F., Matrouf, D., 2005. Variability of automatic speech recognition systems using different features. In: Proceedings of Interspeech, Lisboa, Portugal, pp. 221–224.

Bartkova, K., 2003. Generating proper name pronunciation variants for automatic speech recognition, In: Proceedings of ICPhS. Barcelona, Spain.

Bartkova, K., Jouvet, D., 1999. Language based phone model combination for ASR adaptation to foreign accent. In: Proceedings of ICPhS, San Francisco, USA, pp. 1725–1728.

Bartkova, K., Jouvet, D., 2004. Multiple models for improved speech recognition for non-native speakers. In: Proceedings of SPECOM, Saint Petersburg, Russia.

Beattie, V., Edmondson, S., Miller, D., Patel, Y., Talvola, G., 1995. An integrated multidialect speech recognition system with optional speaker adaptation. In: Proceedings of Eurospeech, Madrid, Spain, pp. 1123–1126.

Beauford, J.Q., 1999. Compensating for variation in speaking rate, PhD thesis, Electrical Engineering, University of Pittsburgh.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., Gildea, D., 2003. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. The Journal of the Acoustical Society of America 113 (2), 1001–1024.

Benitez, C., Burget, L., Chen, B., Dupont, S., Garudadri, H., Hermansky, H., Jain, P., Kajarekar, S., Sivadas, S., 2001. Robust ASR front-end using spectral based and discriminant features: experiments on the aurora task. In: Proceedings of Eurospeech, Aalborg, Denmark, pp. 429–432.

Blomberg, Mats, 1991. Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references. Speech Communication 10 (5–6), 453–461.

Bonaventura, P., Gallochio, F., Mari, J., Micca, G., 1998. Speech recognition methods for non-native pronunciation variants. In: Proceedings ISCA Workshop on modelling pronunciation variations for automatic speech recognition, Rolduc, Netherlands, pp. 17–23.

Bonaventura, P., Gallochio, F., Micca, G., 1997. Multilingual speech recognition for flexible vocabularies. In: Proceedings of Eurospeech, Rhodes, Greece, pp. 355–358.

Bou-Ghazale, S.E., Hansen, J.L.H., 1994. Duration and spectral based stress token generation for HMM speech recognition under stress. In: Proceedings of ICASSP, Adelaide, Australia, pp. 413–416.

Bou-Ghazale, S.E., Hansen, J.L.H., 1995. Improving recognition and synthesis of stressed speech via feature perturbation in a source generator framework. In: ECSA-NATO Proceedings Speech Under Stress Workshop, Lisbon, Portugal, pp. 45–48.

Bourlard, H., Dupont, D., 1997. Sub-band based speech recognition. In: Proceedings of ICASSP, Munich, Germany, pp. 1251–1254.

Bozkurt, B., Couvreur, L., 2005. On the use of phase information for speech recognition. In: Proceedings of Eusipco, Antalya, Turkey.

Bozkurt, B., Doval, B., d'Alessandro, C., Dutoit, T., 2005. Zeros of z-transform representation with application to source-filter separation in speech. IEEE Signal Processing Letters 12 (4), 344–347.

Brugnara, F., De Mori, R., Giuliani, D., Omologo, M., 1992. A family of parallel Hidden Markov Models. In: Proceedings of ICASSP, vol. 1. pp. 377–380.

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajic, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., Wei-Jin, Z., 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Transactions on Speech and Audio Processing 12 (4), 420–435.

Carey, M., Parris, E., Lloyd-Thomas, H., Bennett, S., 1996. Robust prosodic features for speaker identification. In: Proceedings of ICSLP, Philadelphia, Pennsylvania, USA, pp. 1800–1803.

Carlson, B., Clements, M., 1992. Speech recognition in noise using a projection-based likelihood measure for mixture density HMMs. In: Proceedings of ICASSP, San Francisco, CA, pp. 237–240.

Chase, L., 1997. Error-responsive feedback mechanisms for speech recognizers. PhD thesis, Carnegie Mellon University.

Chen, C.J., Gopinath, R.A., Monkowski, M.D., Picheny, M.A., Shen, K., 1997. New methods in continuous mandarin speech recognition. In: Proceedings of Eurospeech, pp. 1543–1546.

Chen, C.J., Li, H., Shen, L., Fu, G., 2001. Recognize tone languages using pitch information on the main vowel of each syllable. In: Proceedings of ICASSP, vol. 1. pp. 61–64.

Chen, Y., 1987. Cepstral domain stress compensation for robust speech recognition. In: Proceedings of ICASSP, Dallas, TX, pp. 717–720.

Chesta, C., Laface, P., Ravera, F., 1999. Connected digit recognition using short and long duration models. In: Proceedings of ICASSP, vol. 2. pp. 557–560.

Chesta, C., Siohan, O., Lee, C.-H., 1999. Maximum a posteriori linear regression for Hidden Markov Model adaptation. In: Proceedings of Eurospeech, Budapest, Hungary, pp. 211–214.

Chollet, G.F., Astier, A.B.P., Rossi, M., 1981. Evaluating the performance of speech recognizers at the acoustic–phonetic level. In: in Proceedings of ICASSP, Atlanta, USA, pp. 758–761.

Cincarek, T., Gruhn, R. Nakamura, S., 2004. Speech recognition for multiple non-native accent groups with speaker-group-dependent acoustic models. In: Proceedings of ICSLP, Jeju Island, Korea, pp. 1509–1512.

Coifman, R.R., Wickerhauser, M.V., 1992. Entropy based algorithms for best basis selection. IEEE Transactions on Information Theory 38 (2), 713–718.

Colibro, D., Fissore, L., Popovici, C., Vair, C., Laface, P., 2005. Learning pronunciation and formulation variants in continuous speech applications. In: Proceedings of ICASSP, Philadelphia, PA, pp. 1001–1004.

Cowie, R., Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. Speech Communication Special Issue on Speech and Emotions 40 (1–2), 5–32.

Cucchiarini, C., Strik, H., Boves, L., 2000. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. Speech Communication 30 (2–3), 109–119.

Dalsgaard, P., Andersen, O., Barry, W., 1998. Cross-language merged speech units and their descriptive phonetic correlates. In: Proceedings of ICSLP, Sydney, Australia, pp. 482–485.

D'Arcy, S.M., Wong, L.P., Russell, M.J., 2004. Recognition of read and spontaneous children's speech using two new corpora. In: Proceedings of ICSLP, Jeju Island, Korea.

Das, S., Lubensky, D., Wu, C., 1999. Towards robust speech recognition in the telephony network environment – cellular and landline conditions. In: Proceedings of Eurospeech, Budapest, Hungary, pp. 1959–1962.

Das, S., Nix, D., Picheny, M., 1998. Improvements in children speech recognition performance. In: Proceedings of ICASSP, vol. 1. Seattle, USA, pp. 433–436.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing 28, 357–366.

de Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S., Bourlard, H., 2004. Evaluation of formant-like features on an automatic vowel classification task. The Journal of the Acoustical Society of America 116 (3), 1781–1792.

Demeechai, T., Mäkeläinen, K., 2001. Recognition of syllables in a tone language. Speech Communication 33 (3), 241–254.

Demuynck, K., Garcia, O., Van Compernolle, D., 2004. Synthesizing speech from speech recognition parameters. In: Proceedings of ICSLP'04, Jeju Island, Korea.

Deng, Y., Mahajan, M., Acero, A., 2003. Estimating speech recognition error rate without acoustic test data. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 929–932.

Di Benedetto, M.-G., Liénard, J.-S., 1992. Extrinsic normalization of vowel formant values based on cardinal vowels mapping. In: Proceedings of ICSLP, Alberta, USA, pp. 579–582.

Disfluency in spontaneous speech (diss'05). 2005. Aix-en-Provence, France.

Doddington, G., 2003. Word alignment issues in ASR scoring. In: Proceedings of ASRU, US Virgin Islands, pp. 630–633.

Draxler, C., Burger, S., 1997. Identification of regional variants of high german from digit sequences in german telephone speech. In: Proceedings of Eurospeech, pp. 747–750.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis. Wiley, New York.

Dupont, S., Ris, C., Couvreur, L., Boite, J.-M., 2005. A study of implicit anf explicit modeling of coarticulation and pronunciation variation. In: Proceedings of Interspeech, Lisboa, Portugal, pp. 1353–1356.

Dupont, S., Ris, C., Deroo, O., Poitoux, S., 2005. Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents. In: Proceedings of ASRU, San Juan, Puerto-Rico, pp. 29–34.

Eide, E., 2001. Distinctive features for use in automatic speech recognition. In: Proceedings of Eurospeech, Aalborg, Denmark, pp. 1613–1616.

Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization. In: Proceedings of ICASSP, Atlanta, GA, pp. 346–348.

Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization. In: Proceedings of ICASSP, Atlanta, GA, pp. 346–349.

Eide, E., Gish, H., Jeanrenaud, P., Mielke, A., 1995. Understanding and improving speech recognition performance through the use of diagnostic tools. In: Proceedings of ICASSP, Detroit, Michigan, pp. 221–224.

Eklund, R., Lindström, A., 2001. Xenophones: an investigation of phone set expansion in swedish and implications for speech recognition and speech synthesis. Speech Communication 35 (1–2), 81–102.

Elenius, D., Blomberg, M., 2004. Comparing speech recognition for adults and children. In: Proceedings of FONETIK, Stockholm, Sweden, pp. 156–159.

Ellis, D., Singh, R., Sivadas, S., 2001. Tandem acoustic modeling in large-vocabulary recognition. In: Proceedings of ICASSP, Salt Lake City, USA, pp. 517–520.

ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, 1998.

Eskenazi, M., 1996. Detection of foreign speakers' pronunciation errors for second language training-preliminary results. In: Proceedings of ICSLP, Philadelphia, PA, pp. 1465–1468.

Eskenazi, M., 1996. Kids: a database of children's speech. The Journal of the Acoustical Society of America, 2759.

Eskenazi, M., Pelton, G., 2002. Pinpointing pronunciation errors in children speech: examining the role of the speech recognizer. In: Proceedings of the PMLA Workshop, Colorado, USA.

Falthauser, R., Pfau, T., Ruske, G., 2000. On-line speaking rate estimation using gaussian mixture models. In: Proceedings of ICASSP, Istanbul, Turkey, pp. 1355–1358.

Fant, G., 1960. Acoustic Theory of Speech Production. Mouton, The Hague.

Fiscus, J.G., 1997. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In: Proceedings of ASRU, pp. 347–354.

Fitt, S., 1995. The pronunciation of unfamiliar native and non-native town names. In: Proceedings of Eurospeech, Madrid, Spain, pp. 2227–2230.

Flanagan, J., 1972. Speech Analysis and Synthesis and Perception. Springer-Verlag, Berlin–Heidelberg–New York.

Flege, J.E., Schirru, C., MacKay, I.R.A., 2003. Interaction between the native and second language phonetic subsystems. Speech Communication 40, 467–491.

Fosler-Lussier, E., Amdal, I., Kuo, H.-K.J., 2005. A framework for predicting speech recognition errors. Speech Communication 46 (2), 153–170.

Fosler-Lussier, E., Morgan, N., 1999. Effects of speaking rate and word predictability on conversational pronunciations. Speech Communication 29 (2–4), 137–158.

Franco, H., Neumeyer, L., Digalakis, V., Ronen, O., 2000. Combination of machine scores for automatic grading of pronunciation quality. Speech Communication 30 (2–3), 121–130.

Fujinaga, K., Nakai, M., Shimodaira, H., Sagayama, S., 2001. Multiple-regression Hidden Markov Model. In: Proceedings of ICASSP, vol. 1. Salt Lake City, USA, pp. 513–516.

Fukunaga, K., 1972. Introduction to Statistical Pattern Recognition. Academic Press, New York.

Fung, P., Liu, W.K., 1999. Fast accent identification and accented speech recognition. In: Proceedings of ICASSP, Phoenix, Arizona, USA, pp. 221–224.

Furui, S., Beckman, M., Hirschberg, J.B., Itahashi, S., Kawahara, T., Nakamura, S., Narayanan, S., 2004. Introduction to the special issue on spontaneous speech processing. IEEE Transactions on Speech and Audio Processing 12 (4), 349–350.

Gales, M.J.F., 1998. Cluster adaptive training for speech recognition. In: Proceedings of ICSLP, Sydney, Australia, pp. 1783–1786.

Gales, M.J.F., 1999. Semi-tied covariance matrices for hidden markov models. IEEE Transactions on Speech and Audio Processing 7, 272–281.

Gales, M.J.F., 2001. Acoustic factorization. In: Proceedings of ASRU, Madona di Campiglio, Italy.

Gales, M.J.F., 2001. Multiple-cluster adaptive training schemes. In: Proceedings of ICASSP, Salt Lake City, Utah, USA, pp. 361–364.

Gao, Y., Ramabhadran, B., Chen, J., Erdogan, H., Picheny, M., 2001. Innovative approaches for large vocabulary name recognition. In: Proceedings of ICASSP, Salt Lake City, Utah, pp. 333–336.

Garner, P., Holmes, W., 1998. On the robust incorporation of formant features into hidden markov models for automatic speech recognition. In: Proceedings of ICASSP, pp. 1–4.

Garvin, P.L., Ladefoged, P., 1963. Speaker identification and message identification in speech recognition. Phonetica 9, 193–199.

Gemello, R., Mana, F., Albesano, D., De Mori, R., 2006. Multiple resolution analysis for robust automatic speech recognition. Computer, Speech and Language 20, 2–21.

Girardi, A., Shikano, K., Nakamura, S., 1998. Creating speaker independent HMM models for restricted database using straight-tempo morphing. In: Proceedings of ICSLP, Sydney, Australia, pp. 687–690.

Giuliani, D., Gerosa, M., 2003. Investigating recognition of children speech. In: Proceedings of ICASSP, Hong Kong, pp. 137–140.

Goel, V., Kumar, S., Byrne, W., 2004. Segmental minimum Bayes-risk decoding for automatic speech recognition. Transactions of IEEE Speech and Audio Processing 12 (3), 234–249.

Gopinath, R.A., 1998. Maximum likelihood modeling with gaussian distributions for classification. In: Proceedings of ICASSP, Seattle, WA, pp. 661–664.

Goronzy, S., Rapp, S., Kompe, R., 2004. Generating non-native pronunciation variants for lexicon adaptation. Speech Communication 42 (1), 109–123.

Graciarena, M., France, H., Zheng, J., Vergyri, D., Stolcke, A., 2004. Voicing feature integration in SRI's DECIPHER LVCSR system. In: Proceedings of ICASSP, Montreal, Canada, pp. 921–924.

Greenberg, S., Chang, S., 2000. Linguistic dissection of switchboard-corpus automatic speech recognition systems. In: Proceedings of ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millenium, Paris, France.

Greenberg, S., Fosler-Lussier, E., 2000. The uninvited guest: information's role in guiding the production of spontaneous speech. In: in Proceedings of the Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling. Kloster Seeon, Germany.

Gupta, S.K., Soong, F., Haimi-Cohen, R., 1996. High-accuracy connected digit recognition for mobile applications. In: Proceedings of ICASSP, vol. 1, pp. 57–60.

Haeb-Umbach, R., Ney, H., 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: Proceedings of ICASSP, San Francisco, CA, pp. 13–16.

Hagen, A., Pellom, B., Cole, R., 2003. Children speech recognition with application to interactive books and tutors. Proceedings of ASRU. St. Thomas, US Virgin Islands, pp. 186–191.

Hain, T., 2005. Implicit modelling of pronunciation variation in automatic speech recognition. Speech Communication 46 (2), 171–188.

Hain, T., Woodland, P.C., 1999. Dynamic HMM selection for continuous speech recognition. In: Proceedings of Eurospeech, Budapest, Hungary, pp. 1327–1330.

Hansen, J.H.L., 1989. Evaluation of acoustic correlates of speech under stress for robust speech recognition. In: IEEE Proceedings 15th Northeast Bioengineering Conference, Boston, MA. Boston, Mass, pp. 31–32.

Hansen, J.H.L., 1993. Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments. In: Proceedings of ICASSP, Minneapolis, Minnesota, pp. 95–98.

Hansen, J.H.L., 1995. A source generator framework for analysis of acoustic correlates of speech under stress. part i: pitch, duration, and intensity effects. The Journal of the Acoustical Society of America.

Hansen, J.H.L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Communications, Special Issue on Speech Under Stress 20 (2), 151–170.

Hanson, B.A., Applebaum, T., 1990. Robust speaker-independent word recognition using instantaneous, dynamic and acceleration features: experiments with Lombard and noisy speech. In: Proceedings of ICASSP, Albuquerque, New Mexico, pp. 857–860.

Hariharan, R., Kiss, I., Viikki, O., 2001. Noise robust speech parameterization using multiresolution feature extraction. IEEE Transactions on Speech and Audio Processing 9 (8), 856–865.

Haykin, S., 1993. Adaptive Filter Theory. Prentice-Hall Publishers, NJ, USA.

Haykin, S., 1994. Communication Systems, third ed. John Wiley and Sons, New York, USA.

Hazen, T.J., Hetherington, I.L., Shu, H., Livescu, K., 2005. Pronunciation modeling using a finite-state transducer representation. Speech Communication 46 (2), 189–203.

He, X., Zhao, Y., 2003. Fast model selection based speaker adaptation for nonnative speech. IEEE Transactions on Speech and Audio Processing 11 (4), 298–307.

Hegde, R.M., Murthy, H.A., Gadde, V.R.R., 2004. Continuous speech recognition using joint features derived from the modified group delay function and MFCC. In: Proceedings of ICSLP, Jeju, Korea, pp. 905–908.

Hegde, R.M., Murthy, H.A., Rao, G.V.R., 2005. Speech processing using joint features derived from the modified group delay function. In: Proceedings of ICASSP, vol. I. Philadelphia, PA, pp. 541–544.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America 87 (4), 1738–1752.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing 2 (4), 578–589.

Hermansky, H., Sharma, S., 1998. TRAPS: classifiers of temporal patterns. In: Proceedings of ICSLP, Sydney, Australia, pp. 1003–1006.

Hetherington, L., 1995. New words: Effect on recognition performance and incorporation issues. In: Proceedings of Eurospeech, Madrid, Spain, pp. 1645–1648.

Hirschberg, J., Litman, D., Swerts, M., 2004. Prosodic and other cues to speech recognition failures. Speech Communication 43 (1–2), 155–175.

Holmes, J.N., Holmes, W.J., Garner, P.N., 1997. Using formant frequencies in speech recognition. In: Proceedings of Eurospeech, Rhodes, Greece, pp. 2083–2086.

Hon, H.W., Wang, K., 1999. Combining frame and segment based models for large vocabulary continuous speech recognition. Proceedings of ASRU. Keystone, Colorado.

Huang, C., Chen, T., Li, S., Chang, E., Zhou, J., 2001. Analysis of speaker variability. In: Proceedings of Eurospeech, Aalborg, Denmark, pp. 1377–1380.

Huang, X., Lee, K., 1991. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. In: Proceedings of ICASSP, Toronto, Canada, pp. 877–880.

Huank, H.C.-H., Seide, F., 2000. Pitch tracking and tone features for Mandarin speech recognition. In: Proceedings of ICASSP, vol. 3. pp. 1523–1526.

Humphries, J.J., Woodland, P.C., Pearce, D., 1996. Using accent-specific pronunciation modelling for robust speech recognition. In: Proceedings of ICSLP, Rhodes, Greece, pp. 2367–2370.

Hunt, M.J., 1999. Spectral signal processing for ASR. Proceedings of ASRU. Keystone, Colorado.

Hunt, M.J., 2004. Speech recognition, syllabification and statistical phonetics. In: Proceedings of ICSLP, Jeju Island, Korea.

Hunt, M.J., Lefebvre, C., 1989. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In: Proceedings of ICASSP, Glasgow, UK, pp. 262–265.

Iivonen, A., Harinen, K., Keinanen, L., Kirjavainen, J., Meister, E., Tuuri, L., 2003. Development of a multiparametric speaker profile for speaker recognition. In: Proceedings of ICPhS, Barcelona, Spain, pp. 695–698.

ISCA Tutorial and Research Workshop, 2002. Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA-2002).

Janse, E., 2004. Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. Speech Communication 42 (2), 155–173.

Jiang, K., Huang, X., 1999. Acoustic feature selection using speech recognizers. Proceedings of ASRU. Keystone, Colorado.

Juang, B.-H., Rabiner, L.R., Wilpon, J.G., 1987. On the use of bandpass liftering in speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 35, 947–953.

Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y., Sen, Z., 2001. What kind of pronunciation variation is hard for triphones to model? In: Proceedings of ICASSP, Salt Lake City, Utah, pp. 577–580.

Kajarekar, S., Malayath, N., Hermansky, H., 1999. Analysis of sources of variability in speech. In: Proceedings of Eurospeech, Budapest, Hungary, pp. 343–346.

Kajarekar, S., Malayath, N., Hermansky, H., 1999. Analysis of speaker and channel variability in speech. Proceedings of ASRU. Keystone, Colorado.

Kenny, P., Boulianne, G., Dumouchel, P., 2005. Eigenvoice modeling with sparse training data. IEEE Transactions on Speech and Audio Processing 13 (3), 345–354.

Köhler, J., 1996. Multilingual phonemes recognition exploiting acoustic–phonetic similarities of sounds. In: Proceedings of ICSLP, Philadelphia, PA, pp. 2195–2198.

Konig, Y., Morgan, N., 1992. GDNN: a gender-dependent neural network for continuous speech recognition. In: Proceedings of Int. Joint Conf. on Neural Networks, vol. 2. Baltimore, Maryland, pp. 332–337.

Kim, D.K., Kim, N.S., 2004. Rapid online adaptation using speaker space model evolution. Speech Communication 42 (3–4), 467–478.

Kingsbury, B., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Communication 25 (1–3), 117–132.

Kingsbury, B., Saon, G., Mangua, L., Padmanabhan, M., Sarikaya, R., 2002. Robust speech recognition in noisy environments: the 2001 IBM SPINE evaluation system. In: Proceedings of ICASSP, vol. I. Orlando, FL, pp. 53–56.

Kirchhoff, K., 1998. Combining articulatory and acoustic information for speech recognition in noise and reverberant environments. In: Proceedings of ICSLP, Sydney, Australia, pp. 891–894.

Kitaoka, N., Yamada, D., Nakagawa, S., 2002. Speaker independent speech recognition using features based on glottal sound source. In: Proceedings of ICSLP, Denver, USA, pp. 2125–2128.

Kleinschmidt, M., Gelbart, D., 2002. Improving word accuracy with gabor feature extraction. In: Proceedings of ICSLP, Denver, Colorado, pp. 25–28.

Korkmazskiy, F., Juang, B.-H., Soong, F., 1997. Generalized mixture of HMMs for continuous speech recognition. In: Proceedings of ICASSP, vol. 2. pp. 1443–1446.

Korkmazsky, F., Deviren, M., Fohr, D., Illina, I., 2004. Hidden factor dynamic bayesian networks for speech recognition. In: Proceedings of ICSLP, Jeju Island, Korea.

Kubala, F., Anastasakos, A., Makhoul, J., Nguyen, L., Schwartz, R., Zavaliagkos, E., 1994. Comparative experiments on large vocabulary

speech recognition. In: Proceedings of ICASSP, Adelaide, Australia, pp. 561–564.
Kumar, N., Andreou, A.G., 1998. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Communication 26 (4), 283–297.
Kumaresan, R., 1998. An inverse signal approach to computing the envelope of a real valued signal. IEEE Signal Processing Letters 5 (10), 256–259.
Kumaresan, R., Rao, A., 1999. Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications. The Journal of the Acoustical Society of America 105 (3), 1912–1924.
Kumpf, K., King, R.W., 1996. Automatic accent classification of foreign accented australian english speech. In: Proceedings of ICSLP, Philadelphia, PA, pp. 1740–1743.
Kuwabara, H., 1997. Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate. In: Proceedings of Eurospeech, Rhodes, Greece, pp. 1003–1006.
Ladefoged, P., Broadbent, D.E., 1957. Information conveyed by vowels. The Journal of the Acoustical Society of America 29, 98–104.
Lamel, L., Gauvain, J.-L., 2005. Alternate phone models for conversational speech. In: Proceedings of ICASSP, Philadelphia, Pennsylvania, pp. 1005–1008.
Laver, J., 1994. Principles of Phonetics. Cambridge University Press, Cambridge.
Lawson, A.D., Harris, D.M., Grieco, J.J., 2003. Effect of foreign accent on speech recognition in the NATO N-4 corpus. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 1505–1508.
Lee, C., Lin, C., Juang, B., 1991. A study on speaker adaptation of the parameters of continuous density Hidden Markov Models. IEEE Transactions Signal Processing 39 (4), 806–813.
Lee, C.-H., Gauvain, J.-L., 1993. Speaker adaptation based on MAP estimation of HMM parameters. In: Proceedings of ICASSP, vol. 2. pp. 558–561.
Lee, L., Rose, R.C., 1996. Speaker normalization using efficient frequency warping procedures. In: Proceedings of ICASSP, vol. 1. Atlanta, Georgia, pp. 353–356.
Lee, S., Potamianos, A., Narayanan, S., 1999. Acoustics of children speech: developmental changes of temporal and spectral parameters. The Journal of the Acoustical Society of America 105, 1455–1468.
Leggetter, C., Woodland, P., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models. Computer, Speech and Language 9 (2), 171–185.
Leonard, R.G., 1984. A database for speaker independent digit recognition. In: Proceedings of ICASSP, San Diego, US, pp. 328–331.
Lin, X., Simske, S., 2004. Phoneme-less hierarchical accent classification. In: Proceedings of Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, vol. 2. Pacific Grove, CA, pp. 1801–1804.
Lincoln, M., Cox, S.J., Ringland, S., 1997. A fast method of speaker normalisation using formant estimation. In: Proceedings of Eurospeech, Rhodes, Greece, pp. 2095–2098.
Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H& H theory. In: Hardcastle, W.J., Marchal, A. (Eds.), Speech Production and Speech Modelling. Kluwer Academic Publishers.
Lippmann, R.P., 1997. Speech recognition by machines and humans. Speech Communication 22 (1), 1–15.
Lippmann, R.P., Martin, E.A., Paul, D.B., 1987. Multi-style training for robust isolated-word speech recognition. In: Proceedings of ICASSP, Dallas, TX, pp. 705–708.
Liu, L., He, J., Palm, G., 1997. Effects of phase on the perception of intervocalic stop consonants. Speech Communication 22 (4), 403–417.
Liu, S., Doyle, S., Morris, A., Ehsani, F., 1998. The effect of fundamental frequency on mandarin speech recognition. In: Proceedings of ICSLP, vol. 6. Sydney, Australia, pp. 2647–2650.
Liu, W.K., Fung, P., 2000. MLLR-based accent model adaptation without accented data. In: Proceedings of ICSLP, vol. 3. Beijing, China, pp. 738–741.
Livescu, K., Glass, J., 2000. Lexical modeling of non-native speech for automatic speech recognition. In: Proceedings of ICASSP, vol. 3. Istanbul, Turkey, pp. 1683–1686.
Ljolje, A., 2002. Speech recognition using fundamental frequency and voicing in acoustic modeling. In: Proceedings of ICSLP, Denver, USA, pp. 2137–2140.
Llitjos, A.F., Black, A.W., 2001. Knowledge of language origin improves pronunciation accuracy of proper names. In: Proceedings of Eurospeech, Aalborg, Denmark.
Lombard, E., 1911. Le signe de l'élévation de la voix. Ann. Maladies Oreille, Larynx, Nez, Pharynx 37.
Loog, M., Duin, R.P.W., 2004. Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion. IEEE Transactions Pattern Analysis and Machine Intelligence 26 (6), 732–739.
Magimai-Doss, M., Stephenson, T.A., Ikbal, S., Bourlard, H., 2004. Modelling auxiliary features in tandem systems. In: Proceedings of ICSLP, Jeju Island, Korea.
Maison, B., 2003. Pronunciation modeling for names of foreign origin. In: Proceedings of ASRU, US Virgin Islands, pp. 429–434.
Mak, B., Hsiao, R., 2004. Improving eigenspace-based MLLR adaptation by kernel PCA. In: Proceedings of ICSLP, Jeju Island, Korea.
Makhoul, J., 1975. Linear prediction: a tutorial review. In: Proceedings of IEEE, vol. 63(4) pp. 561–580.
Mangu, L., Brill, E., Stolcke, A., 2000. Finding consensus in speech recognition: Word-error minimization and other applications of confusion networks. Computer Speech and Language 14 (4), 373–400.
Mansour, D., Juang, B.H., 1989. A family of distortion measures based upon projection operation for robust speech recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 37, 1659–1671.
Markel, J., Oshika, B., Gray, A.H., 1977. Long-term feature averaging for speaker recognition. IEEE Transactions on Acoustics, Speech and Signal Processing 25, 330–337.
Markov, K., Nakamura, S., 2003. Hybrid HMM/BN LVCSR system integrating multiple acoustic features. In: Proceedings of ICASSP, vol. 1. pp. 840–843.
Martin, A., Mauuary, L., 2003. Voicing parameter and energy-based speech/non-speech detection for speech recognition in adverse conditions. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 3069–3072.
Martinez, F., Tapias, D., Alvarez, J., 1998. Towards speech rate independence in large vocabulary continuous speech recognition. In: Proceedings of ICASSP, Seattle, Washington, pp. 725–728.
Martinez, F., Tapias, D., Alvarez, J., Leon, P., 1997. Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In: Proceedings of Eurospeech, Rhodes, Greece, pp. 469–472.
Matsuda, S., Jitsuhiro, T., Markov, K., Nakamura, S., 2004. Speech recognition system robust to noise and speaking styles. In: Proceedings of ICSLP, Jeju Island, Korea.
Mertins, A., Rademacher, J., 2005. Vocal tract length invariant features for automatic speech recognition. In: Proceedings of ASRU, Cancun, Mexico, pp. 308–312.
Messina, R., Jouvet, D., 2004. Context dependent long units for speech recognition. In: Proceedings of ICSLP, Jeju Island, Korea.
Milner,. B.P., 1996. Inclusion of temporal information into features for speech recognition. In: Proceedings of ICSLP, Philadelphia, PA, pp. 256–259.
Mirghafori, N., Fosler, E., Morgan, N., 1995. Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes. In: Proceedings of Eurospeech, Madrid, Spain, pp. 491–494.
Mirghafori, N., Fosler, E., Morgan, N., 1996. Towards robustness to fast speech in ASR. In: Proceedings of ICASSP, Atlanta, Georgia, pp. 335–338.
Mokbel, C., Mauuary, L., Karray, L., Jouvet, D., Monné, J., Simonin, J., Bartkova, K., 1997. Towards improving ASR robustness for PSN and

GSM telephone applications. Speech Communication 23 (1–2), 141–159.

Mokhtari, P., 1998. An acoustic–phonetic and articulatory study of speech-speaker dichotomy. PhD thesis, The University of New South Wales, Canberra, Australia.

Morgan, N., Chen, B., Zhu, Q., Stolcke, A., 2004. TRAPping conversational speech: extending TRAP/tandem approaches to conversational telephone speech recognition. In: Proceedings of ICASSP, vol. 1. Montreal, Canada, pp. 536–539.

Morgan, N., Fosler, E., Mirghafori, N., 1997. Speech recognition using on-line estimation of speaking rate. In: Proceedings of Eurospeech, vol. 4. Rhodes, Greece, pp. 2079–2082.

Morgan, N., Fosler-Lussier, E., 1998. Combining multiple estimators of speaking rate. In: Proceedings of ICASSP, Seattle, pp. 729–732.

Murray, I.R., Arnott, J.L., 1993. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. The Journal of the Acoustical Society of America 93 (2), 1097–1108.

Masaki Naito, Y.S., LiDeng, 1998. Speaker clustering for speech recognition using the parameters characterizing vocal-tract dimensions. In: Proceedings of ICASSP, Seattle, WA, pp. 1889–1893.

Nanjo, H., Kawahara, T., 2002. Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. In: Proceedings of ICASSP, vol. 1. Orlando, FL, pp. 725–728.

Nanjo, H., Kawahara, T., 2004. Language model and speaking rate adaptation for spontaneous presentation speech recognition. IEEE Transactions on Speech and Audio Processing 12 (4), 391–400.

National Institute of Standards and Technology, 2001. SCLITE scoring software. ftp://jaguar.ncls.nist.gov/pub/sctk-1.2.tar.Z.

Nearey, T.M., 1978. Phonetic feature systems for vowels. Indiana University Linguistics Club, Bloomington, Indiana, USA.

Neti, C., Roukos, S., 1997. Phone-context specific gender-dependent acoustic-models for continuous speech recognition. In: Proceedings of ASRU, Santa Barbara, CA, pp. 192–198.

Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M., 2000. Automatic scoring of pronunciation quality. Speech Communication 30 (2–3), 83–93.

Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. 1996. Automatic text-independent pronunciation scoring of foreign language student speech. In: Proceedings of ICSLP, Philadelphia, PA, pp. 1457–1460.

Nguyen, P., Kuhn, R., Junqua, J.-C., Niedzielski, N., Wellekens, C., 2000. Eigenvoices: a compact representation of speakers in a model space. Annales des Télécommunications 55 (3–4).

Nguyen, P., Rigazio, L., Junqua, J.-C., 2003. Large corpus experiments for broadcast news recognition. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 1837–1840.

Nolan, F., 1983. The Phonetic Bases of Speaker Recognition. Cambridge University Press, Cambridge.

Kamal Omar, M., Chen, K., Hasegawa-Johnson, M., Bradman, Y., 2002. An evaluation of using mutual information for selection of acoustic features representation of phonemes for speech recognition. In: Proceedings of ICSLP, Denver, CO, pp. 2129–2132.

Kamal Omar, M., Hasegawa-Johnson, M., 2002. Maximum mutual information based acoustic features representation of phonological features for speech recognition. In: Proceedings of ICASSP, vol. 1. Montreal, Canada, pp. 81–84.

Odell, J.J., Woodlandand, P.C., Valtchev, V., Young, S.J., 1994. Large vocabulary continuous speech recognition using HTK. In: Proceedings of ICASSP, vol. 2. Adelaide, Australia, pp. 125–128.

Ono, Y., Wakita, H., Zhao, Y., 1993. Speaker normalization using constrained spectra shifts in auditory filter domain. In: Proceedings of Eurospeech, Berlin, Germany, pp. 355–358.

O'Saughnessy, D., 1987. Speech Communication – Human and Machine. Addison-Wesley.

O'Shaughnessy, D., Tolba, H., 1999. Towards a robust/fast continuous speech recognition system using a voiced–unvoiced decision. In: Proceedings of ICASSP, vol. 1. Phoenix, Arizona, pp. 413–416.

Padmanabhan, M., Bahl, L., Nahamoo, D., Picheny, M., 1996. Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems. In: Proceedings of ICASSP, Atlanta, GA, pp. 701–704.

Padmanabhan, M., Dharanipragada, S., 2004. Maximum-likelihood nonlinear transformation for acoustic adaptation. IEEE Transactions on Speech and Audio Processing 12 (6), 572–578.

Padmanabhan, M., Dharanipragada, S., 2005. Maximizing information content in feature extraction. IEEE Transactions on Speech and Audio Processing 13 (4), 512–519.

Paliwal, K.K., Alsteris, L., 2003. Usefulness of phase spectrum in human speech perception. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 2117–2120.

Paliwal, K.K., Atal, B.S., 2003. Frequency-related representation of speech. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 65–68.

Paul, D.B., 1987. A speaker-stress resistant HMM isolated word recognizer. In: Proceedings of ICASSP, Dallas, Texas, pp. 713–716.

Paul, D.B., 1997. Extensions to phone-state decision-tree clustering: single tree and tagged clustering. In: Proceedings of ICASSP, vol. 2. Munich, Germany, pp. 1487–1490.

Peters, S.D., Stubley, P., Valin, J.-M., 1999. On the limits of speech recognition in noise. In: Proceedings of ICASSP'99. Phoenix, Arizona, pp. 365–368.

Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of the vowels. The Journal of the Acoustical Society of America 24, 175–184.

Pfau, T., Ruske, G., 1998. Creating Hidden Markov Models for fast speech. In: Proceedings of ICSLP, Sydney, Australia.

Michael Pitz, 2005. Investigations on Linear Transformations for Speaker Adaptation and Normalization. PhD thesis, RWTH Aachen University.

Plumpe, M., Quatieri, T., Reynolds, D., 1999. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Transactions on Speech and Audio Processing 7 (5), 569–586.

Pols, L.C.W., Van der Kamp, L.J.T., Plomp, R., 1969. Perceptual and physical space of vowel sounds. The Journal of the Acoustical Society of America 46, 458–467.

Potamianos, G., Narayanan, S., 2003. Robust recognition of children speech. IEEE Transactions on Speech and Audio Processing 11, 603–616.

Potamianos, G., Narayanan, S., Lee, S., 1997. Analysis of children speech: duration, pitch and formants. In: Proceedings of Eurospeech, Rhodes, Greece, pp. 473–476.

Potamianos, G., Narayanan, S., Lee, S., 1997. Automatic speech recognition for children. In: Proceedings of Eurospeech, Rhodes, Greece, pp. 2371–2374.

Potter, R.K., Steinberg, J.C., 1950. Toward the specification of speech. The Journal of the Acoustical Society of America 22, 807–820.

Printz, H., Olsen, P.A., 2002. Theory and practice of acoustic confusability. Computer Speech and Language 16 (1), 131–164.

Pujol, P., Pol, S., Nadeu, C., Hagen, A., Bourlard, H., 2005. Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system. IEEE Transactions on Speech and Audio Processing SAP-13 (1), 14–22.

Rabiner, L., Juang, B.H., 1993. Fundamentals of speech recognition. Prentice Hall PTR, Englewoood Cliffs, NJ, USA (pp. 20–37, Chapter 2).

Rabiner, L.R., Lee, C.H., Juang, B.H., Wilpon, J.G., 1989. HMM clustering for connected word recognition. In: Proceedings of ICASSP, vol. 1. Glasgow, Scotland, pp. 405–408.

Raux, A., 2004. Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition. In: Proceedings of ICSLP, Jeju Island, Korea.

Saito, S., Itakura, F., 1983. Frequency spectrum deviation between speakers. Speech Communication 2, 149–152.

Sakauchi, S., Yamaguchi, Y., Takahashi, S., Kobashikawa, S., 2004. Robust speech recognition based on HMM composition and modified

Wiener filter. In: Proceedings of Interspeech. Jeju Island, Korea, pp. 2053–2056.
Saon, G., Padmanabhan, M., Gopinath, R., Chen, S., 2000. Maximum likelihood discriminant feature spaces. In: Proceedings of ICASSP, pp. 1129–1132.
Schaaf, T., Kemp, T., 1997. Confidence measures for spontaneous speech recognition. In: Proceedings of ICASSP, Munich, Germany, pp. 875–878.
Scherer, K.R., 2003. Vocal communication of emotion: A review of research paradigms. Speech Communication Special Issue on Speech and Emotions 40 (1–2), 227–256.
Schimmel, S., Atlas, L., 2005. Coherent envelope detection for modulation filtering of speech. In: Proceedings of ICASSP, vol. 1. Philadephia, USA, pp. 221–224.
Schroeder, M.R., Strube, H.W., 1986. Flat-spectrum speech. The Journal of the Acoustical Society of America 79 (5), 1580–1583.
Schötz, S., 2001. A perceptual study of speaker age. In: Working paper 49, Lund University, Dept of Linguistic, pp. 136–139.
Schultz, T., Waibel, A., 1998. Language independent and language adaptive large vocabulary speech recognition. In: Proceedings of ICSLP, vol. 5. Sydney, Australia, pp. 1819–1822.
Schwartz, R., Barry, C., Chow, Y.-L., Deft, A., Feng, M.-W., Kimball, O., Kubala, F., Makhoul, J., Vandegrift, J., 1989. The BBN BYBLOS continuous speech recognition system. In: Proceedings of Speech and Natural Language Workshop. Philadelphia, Pennsylvania, pp. 21–23.
Selouani, S.-A., Tolba, H., O'Shaughnessy, D., 2002. Distinctive features, formants and cepstral coefficients to improve automatic speech recognition. In: Conference on Signal Processing, Pattern Recognition and Applications, IASTED. Crete, Greece, pp. 530–535.
Seneff, S., Wang, C., 2005. Statistical modeling of phonological rules through linguistic hierarchies. Speech Communication 46 (2), 204–216.
Shi, Y.Y., Liu, J., Liu, R.S., 2002. Discriminative HMM stream model for Mandarin digit string speech recognition. In: Proceedings of Int. Conf. on Signal Processing, vol. 1. Beijing, China, pp. 528–531.
Shinozaki, T., Furui, S., 2003. Hidden mode HMM using bayesian network for modeling speaking rate fluctuation. In: Proceedings of ASRU. US Virgin Islands, pp. 417–422.
Shinozaki, T., Furui, S., 2004. Spontaneous speech recognition using a massively parallel decoder. In: Proceedings of ICSLP, Jeju Island, Korea, pp. 1705–1708.
Shobaki, K., Hosom, J.-P., Cole, R., 2000. The OGI kids speech corpus and recognizers. In: Proceedings of ICSLP, Beijing, China, pp. 564–567.
Siegler, M.A., 1995. Measuring and compensating for the effects of speech rate in large vocabulary continuous speech recognition. PhD thesis, Carnegie Mellon University.
Siegler, M.A., Stern, R.M., 1995. On the effect of speech rate in large vocabulary speech recognition system. In: Proceedings of ICASSP, Detroit, Michigan, pp. 612–615.
Singer, H., Sagayama, S., 1992. Pitch dependent phone modelling for HMM based speech recognition. In: Proceedings of ICASSP, vol. 1. San Francisco, CA, pp. 273–276.
Slifka, J., Anderson, T.R., 1995. Speaker modification with LPC pole analysis. In: Proceedings of ICASSP, Detroit, MI, pp. 644–647.
Song, M.G., Jung, H.I., Shim, K.-J., Kim, H.S., 1998. Speech recognition in car noise environments using multiple models according to noise masking levels. In: Proceedings of ICSLP.
Sotillo, C., Bard, E.G., 1998. Is hypo-articulation lexically constrained?. In: Proceedings of SPoSS. Aix-en-Provence, pp. 109–112.
Sroka, J.J., Braida, L.D., 2005. Human and machine consonant recognition. Speech Communication 45 (4), 401–423.
Steeneken, H.J.M., van Velden, J.G., 1989. Objective and diagnostic assessment of (isolated) word recognizers. In: Proceedings of ICASSP, vol. 1. Glasgow, UK, pp. 540–543.
Stephenson, T.A., Bourlard, H., Bengio, S., Morris, A.C., 2000. Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables. In: Proceedings of ICSLP, vol. 2. Beijing, China, pp. 951–954.
Stephenson, T.A., Doss, M.M., Bourlard, H., 2004. Speech recognition with auxiliary information. IEEE Transactions on Speech and Audio Processing SAP-12 (3), 189–203.
Stolcke, A., Grezl, F., Hwang, M.-Y., Morgan, N., Vergyri, D., 2006. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In: Proceedings of ICASSP, vol. 1. Toulouse, France, pp. 321–324.
Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: a survey of the literature. Speech Communication 29 (2–4), 225–246.
Sun, D.X., Deng, L., 1995. Analysis of acoustic–phonetic variations in fluent speech using Timit. In: Proceedings of ICASSP, Detroit, Michigan, pp. 201–204.
Suzuki, H., Zen, H., Nankaku, Y., Miyajima, C., Tokuda, K., Kitamura, T., 2003. Speech recognition using voice-characteristic-dependent acoustic models. In: Proceedings of ICASSP, vol. 1. Hong-Kong (canceled), pp. 740–743.
Svendsen, T., Paliwal, K.K., Harborg, E., Husoy, P.O., 1989. An improved sub-word based speech recognizer. In: Proceedings of ICASSP, Glasgow, UK, pp. 108–111.
Svendsen, T., Soong, F., 1987. On the automatic segmentation of speech signals. In: Proceedings of ICASSP, Dallas, Texas, pp. 77–80.
Teixeira, C., Trancoso, I., Serralheiro, A., 1996. Accent identification. In: Proceedings of ICSLP, vol. 3. Philadelphia, PA, pp. 1784–1787.
Thomson, D.L., Chengalvarayan, R., 1998. Use of periodicity and jitter as speech recognition feature. In: Proceedings of ICASSP, vol. 1. Seattle, WA, pp. 21–24.
Thomson, D.L., Chengalvarayan, R., 2002. Use of voicing features in HMM-based speech recognition. Speech Communication 37 (3–4), 197–211.
Tibrewala, S., Hermansky, H., 1997. Sub-band based recognition of noisy speech. In: Proceedings of ICASSP, Munich Germany, pp. 1255–1258.
Tolba, H., Selouani, S.A., O'Shaughnessy, D., 2002. Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm. In: Proceedings of ICASSP, Orlando, FL, pp. 837–840.
Tolba, H., Selouani, S.A., O'Shaughnessy, D., 2003. Comparative experiments to evaluate the use of auditory-based acoustic distinctive features and formant cues for robust automatic speech recognition in low-snr car environments. In: Proceedings of Eurospeech, Geneva, Switzerland, pp. 3085–3088.
Tomlinson, M.J., Russell, M.J., Moore, R.K., Buckland, A.P., Fawley, M.A., 1997. Modelling asynchrony in speech using elementary single-signal decomposition. In: Proceedings of ICASSP, Munich Germany, pp. 1247–1250.
Townshend, B., Bernstein, J., Todic, O., Warren, E., 1998. Automatic text-independent pronunciation scoring of foreign language student speech. In: Proceedings of STiLL-1998, Stockholm, pp. 179–182.
Traunmüller, H., 1997. Perception of speaker sex, age and vocal effort. Technical Report, Institutionen för lingvistik, Stockholm Universitet.
Tsakalidis, S., Doumpiotis, V., Byrne, W., 2005. Discriminative linear transforms for feature normalization and speaker adaptation in HMM estimation. IEEE Transactions on Speech and Audio Processing 13 (3), 367–376.
Tsao, Y., Lee, S.-M., Lee, L.-S., 2005. Segmental eigenvoice with delicate eigenspace for improved speaker adaptation. IEEE Transactions on Speech and Audio Processing 13 (3), 399–411.
Tuerk, A., Young, S., 1999. Modeling speaking rate using a between frame distance metric. In: Proceedings of Eurospeech, vol. 1. Budapest, Hungary, pp. 419–422.
Tuerk, C., Robinson, T., 1993. A new frequency shift function for reducing inter-speaker variance. In: Proceedings of Eurospeech, Berlin, Germany, pp. 351–354.
Tur, G., Hakkani-Tür, D., Schapire, R.E., 2005. Combining active and semi-supervised learning for spoken language understanding. Speech Communication 45 (2), 171–186.

Tyagi, V., McCowan, I., Bourlard, H., Misra, H., 2003. Mel-cepstrum modulation spectrum (MCMS) features for robust ASR. Proceedings of ASRU. St. Thomas, US Virgin Islands, pp. 381–386.

Tyagi, V., Wellekens, C., 2005. Cepstrum representation of speech. In: Proceedings of ASRU, Cancun, Mexico.

Tyagi, V., Wellekens, C., Bourlard, H., 2005. On variable-scale piecewise stationary spectral analysis of speech signals for ASR. In: Proceedings of Interspeech, Lisbon, Portugal, pp. 209–212.

Uebler, U., 2001. Multilingual speech recognition in seven languages. Speech Communication 35 (1–2), 53–69.

Uebler, U., Boros, M., 1999. Recognition of non-native german speech with multilingual recognizers. In: Proceedings of Eurospeech, vol. 2. Budapest, Hungary, pp. 911–914.

Umesh, S., Cohen, L., Marinovic, N., Nelson, D., 1999. Scale transform in speech analysis. IEEE Transactions on Speech and Audio Processing 7 (1), 40–45.

Utsuro, T., Harada, T., Nishizaki, H., Nakagawa, S., 2002. A confidence measure based on agreement among multiple LVCSR models – correlation between pair of acoustic models and confidence. In: Proceedings of ICSLP, Denver, Colorado, pp. 701–704.

VanCompernolle, D., 2001. Recognizing speech of goats, wolves, sheep and ... non-natives. Speech Communication 35 (1–2), 71–79.

VanCompernolle, D., Smolders, J., Jaspers, P., Hellemans, T., 1991. Speaker clustering for dialectic robustness in speaker independent speech recognition. In: Proceedings of Eurospeech, Genova, Italy, pp. 723–726.

Vaseghi, S.V., Harte, N., Miller, B., 1997. Multi resolution phonetic/segmental features and models for HMM-based speech recognition. In: Proceedings of ICASSP, Munich Germany, pp. 1263–1266.

Venkataraman, A., Stolcke, A., Wangal, W., Vergyri, D., Ramana Rao Gadde, V., Zheng, J., 2004. An efficient repair procedure for quick transcriptions. In: Proceedings of ICSLP, Jeju Island, Korea.

Wakita, H., 1977. Normalization of vowels by vocal-tract length and its application to vowel identification. IEEE Transactions on Acoustics, Speech and Signal Processing 25, 183–192.

Watrous, R., 1993. Speaker normalization and adaptation using second-order connectionist networks. IEEE Transactions Neural Networks 4 (1), 21–30.

Mitch Weintraub, Kelsey Taussig, Kate Hunicke-Smith, Amy Snodgrass. 1996. Effect of speaking style on LVCSR performance. In: Proceedings Addendum of ICSLP, Philadelphia, PA, USA.

Welling, L., Ney, H., Kanthak, S., 2002. Speaker adaptive modeling by vocal tract normalization. IEEE Transactions on Speech and Audio Processing 10 (6), 415–426.

Weng, F., Bratt, H., Neumeyer, L., Stomcke, A., 1997. A study of multilingual speech recognition. In: Proceedings of Eurospeech, vol. 1. Rhodes, Greece, pp. 359–362.

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., Kollmeier, B., 2005. Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. In: Proceedings of Interspeech. Lisboa, Portugal, pp. 1273–1276.

Westphal, M., 1997. The use of cepstral means in conversational speech recognition. In: Proceedings of Eurospeech, vol. 3. Rhodes, Greece, pp. 1143–1146.

Williams, D.A.G., 1999. Knowing what you don't know: Roles for confidence measures in automatic speech recognition. PhD thesis, University of Sheffield.

Wilpon, J.G., Jacobsen, C.N., 1996. A study of speech recognition for children and the elderly. In: Proceedings of ICASSP, vol. 1. Atlanta, Georgia, pp. 349–352.

Witt, S.M., Young, S.J., 1999. Off-line acoustic modelling of non-native accents. In: Proceedings of Eurospeech, vol. 3. Budapest, Hungary, pp. 1367–1370.

Witt, S.M., Young, S.J., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Communication 30 (2–3), 95–108.

Wong, P.-F., Siu, M.-H., 2004. Decision tree based tone modeling for chinese speech recognition. In: Proceedings of ICASSP, vol. 1. Montreal, Canada, pp. 905–908.

Wrede, B., Fink, G.A., Sagerer, G., 2001. An investigation of modeling aspects for rate-dependent speech recognition. In: Proceedings of Eurospeech, Aalborg, Denmark.

Wu, X., Yan, Y., 2004. Speaker adaptation using constrained transformation. IEEE Transactions on Speech and Audio Processing 12 (2), 168–174.

Yang, W.-J., Lee, J.-C., Chang, Y.-C., Wang, H.-C., 1988. Hidden Markov Model for Mandarin lexical tone recognition. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36. pp. 988–992.

Zavaliagkos, G., Schwartz, R., McDonough, J., 1996. Maximum a posteriori adaptation for large scale HMM recognizers. In: Proceedings of ICASSP, Atlanta, Georgia, pp. 725–728.

Zhang, B., Matsoukas, S., 2005. Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition. In: Proceedings of ICASSP, vol. 1. Philadelphia, PA, pp. 925–928.

Zhan, P., Waibel, A., 1997. Vocal tract length normalization for large vocabulary continuous speech recognition. Technical Report CMU-CS-97-148, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Zhan, P., Westphal, M., 1997. Speaker normalization based on frequency warping. In: Proceedings of ICASSP, vol. 2. Munich, Germany, pp. 1039–1042.

Zhang, Y., Desilva, C.J.S., Togneri, A., Alder, M., Attikiouzel, Y., 1994. Speaker-independent isolated word recognition using multiple Hidden Markov Models. In: Proceedings IEE Vision, Image and Signal Processing, vol. 141(3). pp. 197–202.

Zheng, J., Franco, H., Stolcke, A., 2000. Rate of speech modeling for large vocabulary conversational speech recognition. In: Proceedings of ISCA tutorial and research workshop on automatic speech recognition: challenges for the new Millenium. Paris, France, pp. 145–149.

Zheng, J., Franco, H., Stolcke, A., 2004. Effective acoustic modeling for rate-of-speech variation in large vocabulary conversational speech recognition. In: Proceedings of ICSLP, Jeju Island, Korea, pp. 401–404.

Zhou, B., Hansen, J.H.L., 2005. Rapid discriminative acoustic model based on eigenspace mapping for fast speaker adaptation. IEEE Transactions on Speech and Audio Processing 13 (4), 554–564.

Zhou, G., Deisher, M.E., Sharma, S., 2002. Causal analysis of speech recognition failure in adverse environments. In: Proceedings of ICASSP, vol. 4. Orlando, Florida, pp. 3816–1819.

Zhu, Q., Alwan, A., 2000. AM-demodualtion of speech spectra and its application to noise robust speech recognition. In: Proceedings of ICSLP, vol. 1. Beijing, China, pp. 341–344.

Zhu, D., Paliwal, K.K., 2004. Product of power spectrum and group delay function for speech recognition. In: Proceedings of ICASSP, pp. 125–128.

Zhu, Q., Chen, B., Morgan, N., Stolcke, A., 2004. On using MLP features in LVCSR. In: Proceedings of ICSLP, Jeju Island, Korea.

Zolnay, A., Schlüter, R., Ney, H., 2002. Robust speech recognition using a voiced–unvoiced feature. In: Proceedings of ICSLP, vol. 2. Denver, CO, pp. 1065–1068.

Zolnay, A., Schlüter, R., Ney, H., 2005. Acoustic feature combination for robust speech recognition. In: Proceedings of ICASSP, vol. I. Philadelphia, PA, pp. 457–460.