



Automatic Speech Recognition and Speech Variability: a Review

M. Benzeguiba, Renato de Mori, O. Deroo, S. Dupon, T. Erbes, D. Jouvét, L. Fissore, P. Laface, A. Mertins, C. Ris, et al.

► To cite this version:

M. Benzeguiba, Renato de Mori, O. Deroo, S. Dupon, T. Erbes, et al.. Automatic Speech Recognition and Speech Variability: a Review. *Speech Communication*, 2007, 49 (10-11), pp.763. <10.1016/j.specom.2007.02.006>. <hal-00499180>

HAL Id: hal-00499180

<https://hal.science/hal-00499180v1>

Submitted on 9 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Accepted Manuscript

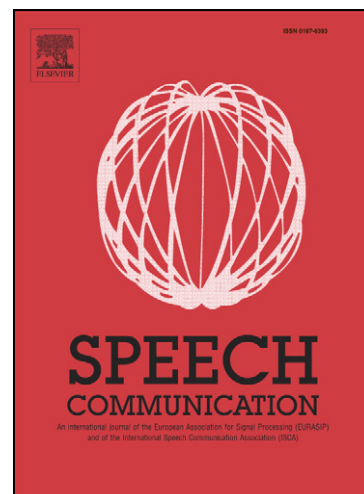
Automatic Speech Recognition and Speech Variability: a Review

M. Benzeguiba, R. De Mori, O. Deroo, S. Dupon, T. Erbes, D. Jouvét, L. Fissore,
P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens

PII: S0167-6393(07)00040-4
DOI: [10.1016/j.specom.2007.02.006](https://doi.org/10.1016/j.specom.2007.02.006)
Reference: SPECOM 1623

To appear in: *Speech Communication*

Received Date: 14 April 2006
Revised Date: 30 January 2007
Accepted Date: 6 February 2007



Please cite this article as: Benzeguiba, M., Mori, R.D., Deroo, O., Dupon, S., Erbes, T., Jouvét, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C., Automatic Speech Recognition and Speech Variability: a Review, *Speech Communication* (2007), doi: [10.1016/j.specom.2007.02.006](https://doi.org/10.1016/j.specom.2007.02.006)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Automatic Speech Recognition and Speech Variability: a Review

*M. Benzeguiba, 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*

contact: dupont@multitel.be

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...), especially when resources for system training are scarce. This paper outlines current advances related to these topics.

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 [91, 210].

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, and 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, and 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 pre-

vious phonemes, and 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 pronunciation effects may appear, and 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 [176].

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, ... 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 alter-

ations 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, ...), 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, [125] 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 [148, 158]. In fact accented speech is associated to a shift within the feature space [295]. Good classification results between regional accents are reported in [58] for human listeners on German SpeechDat data, and in [165] 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 [18]. 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 [77]. 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 [136].

Accent classification is also studied since many years [9], based either on phone models [152, 274] or specific acoustic features [83].

Speech recognition technology is also used in foreign language learning for rating the quality of the pronunciation [69, 80, 207, 281]. Experiments showed that the provided rating is correlated with human expert ratings [46, 206, 309] 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 [75]. 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 [89]. Native phoneme inventories are enlarged with some phonemes of foreign languages in usual pronunciations of foreign names, especially in some languages [66]. Determining the ethnic origin of a word improves pronunciation models [175] and is useful in predicting additional pronunciation variants [15, 179].

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 [196].

The impact of inter-speaker variability on the automatic speech recognition performance has been acknowledged for years. In [126, 159, 250], the authors mention error rates two to three times higher for speaker-independent ASR systems compared with speaker-dependent systems. Methods that aim 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 [73] 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 [33, 132, 184] 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) [210] and the tension of vocal folds will affect the voice quality. The vocal tract, can be modeled by a tube resonator [73, 157]. 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 [23, 30, 229]. 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) [135].

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 [226, 231, 235] 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 [231, 242]. Another important observation [155, 204, 226, 235] 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 [76, 215]. 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 [13, 133, 167, 263], and possible implications to ASR in [20].

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 [44] (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 [301], 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 studies¹. As a fundamental observation, the strong dependency of pronunciation phenomena with respect to the syllable structure has been highlighted in [5, 99]. 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 [191].

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: [32, 84].

There are also regular workshops specifically addressing the research activities related to spontaneous speech modeling and recognition [56]. Regarding the topic of pronunciation variation, the reader should also refer to [241].

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.

In [133], 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

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

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 [153].

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 [188, 193, 259] 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 [259], or *mean of m* [193]. To reduce the dependency on the phone type, a normalization scheme by the expected phone duration [188] or the use of phone duration percentile [258] 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 [199], the authors proposed the *mrates* measure (short for *multiple rate*). It combines three independent ROS measures, i.e., (1) the energy rate or *enrate* [198], (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 *mrates* is also proposed in [19]. In [285], 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 [72], 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 [79, 100], 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 dis-

cussed. 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 [188, 203]), 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 [67, 103, 282]. 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 [162, 233]. 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 [50, 94, 232, 234]. Such Approaches included vocal tract length normalization (VTLN) [50] as well as spectral normalization [161].

A second problem is that younger children may not have a correct pronunciation. Sometimes they have not yet learned how to articulate specific phonemes [251]. 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 don't appear in grown-up speech. The correct inflectional forms of certain words may not have been acquired fully, especially for those words that are exceptions to common rules. Spontaneous speech is also believed to be less grammatical than for adults. A number of different solutions 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 [71], 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. [70, 307] 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 [45, 246].

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 [8]. 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 (cfr Section 4.2.2).

As Hansen formulates in [109], 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 [27, 106, 107, 108, 200]. The robust ASR approaches are covered by chapter 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's cite pathologies affecting the larynx or the lungs, or even the discourse (dysphasia, stuttering, cerebral vascular accident, ...), long-term habits as smoking, singing, ..., speaking styles like whispering, shouting, ... physical activity causing breathlessness, fatigue, ...

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 (cfr. 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 [41] [130]. Further groupings separate the test cases according to the physical differences of the speakers, such as male/female, children/adult, or accent [125]. 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 [99]. Alternatively, other papers report confidence scores to measure the performance of their recognizers [306, 323].

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 [99], 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 [265] 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 [65]. 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 [137, 138, 270] 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 [57] 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, [78] 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 [55, 122, 236] and assignment of error liability [35] and is adjacent to the research on confusion networks [95, 121, 182, 211, 245].

3.2. Man-machine comparison

A few years ago, a publication [168] 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 con-

straints 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 [264], 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 [304] (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...) 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...) 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 [54, 164, 225] 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, ...).

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 [170], the authors clearly demonstrate the importance of the phase information for correctly classifying stop consonants, especially regarding their voicing property. Moreover, in [248], 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 [168], 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 [54]. 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 [225] 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, ...). 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 [220], 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 [302] or else improved invariance [190]) 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)[51] or Perceptual Linear Prediction (PLP) coefficients[118], are based on some sort of representation of the smoothed spectral envelope, usually estimated over fixed analysis windows of typically 20 ms to 30 ms [51, 238]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 [238]. 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 [112]. 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 [290], 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 [7, 272, 273]. 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 [10], 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 ele-

mentary components are not necessarily quasi-stationary and vice-versa.

Coifman et al [43] have described a minimum entropy basis selection algorithm to achieve the minimum information cost of a signal relative to the designed orthonormal basis. In [273], Svendsen et al have proposed a ML segmentation algorithm using a single fixed window size for speech analysis, followed by a clustering of the frames which were spectrally similar for sub-word unit design. More recently, Achan et al [4] 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[120], MLPs and several modulation spectrum related techniques[141, 192, 288, 325]. 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) [113]. In theory, the AM signal modulates a narrow-band carrier signal (specifically, a monochromatic sinusoidal signal). Therefore to be able to extract the AM signals of a wide-band signal such as speech (typically 4KHz), it is necessary to decompose the speech signal into narrow spectral bands. In [289], this approach is opposed to the previous use of the speech modulation spectrum [141, 192, 288, 325] 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[247]. In their experiment, they consider a wide-band filtered speech signal $x(t) = a(t)c(t)$, where $a(t)$ is the AM signal and $c(t)$ is the broad-band carrier signal. Then, they perform a low-pass modulation filtering of the AM signal $a(t)$ to obtain $a_{LP}(t)$. The low-pass filtered AM signal $a_{LP}(t)$ is then multiplied with the original carrier $c(t)$ to obtain a new signal $\tilde{x}(t)$. They show 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 [181]. Lately, Kumaresan et al.[150, 151] have proposed to model analytic signals using pole-zero models in the temporal domain. Along similar lines, Athineos et al.[12] have used the dual of the 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 [221], 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 [29, 116, 324] which presents a formant-like structure with a much higher resolution than the

²Note 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 [129], 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.

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

power spectrum. As the group delay is inherently very noisy, the approaches proposed by the authors mainly aim 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 compensates 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 justifies the success of the non-linear frequency scales such as Mel, Bark, Erb, ... Similarly, in [118], 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 [293], 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 [190], 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 [190].

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 [21, 299], transformation of the LPC pole modeling [261], frequency warping, either linear [63, 161, 286, 317] or non-linear [214], 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 [302]. Note that VTLN is often combined with an adaptation of the acoustic model to the canonical speaker [63, 161] (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 [50].

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 [138, 305].

Similarly, some studies attempted to devise feature extraction methods tailored for the recognition of stressed and non-stressed speech simultaneously. In his paper [38], 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 [109, 119, 131] and some researchers simply assessed the better representations from the existing pool of features [110].

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 [213], [212], 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 ([28, 277, 280]) 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 ([297], [111]),
- segment and frame-based acoustic features ([124]),
- MFCC, PLP and an auditory feature ([134]),
- spectral-based and discriminant features ([22]),
- acoustic and articulatory features ([143, 278]),
- LPC based cepstra, MFCC coefficients, PLP coefficients, energies and time-averages ([213],[212]), MFCC and PLP ([328]),
- full band non-compressed root cepstral coefficients (RCC), Full band PLP 16kHz, Telephone band PLP 8 kHz ([142]),
- PLP, MFCC and wavelet features ([92]),
- joint features derived from the modified group-delay function ([117]),
- combinations of frequency filtering (FF), MFCC, RASTA-FF, (J)RASTA-PLP ([237]).

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

- periodicity and jitter ([275]),
- voicing ([327], [98]),
- rate of speech and pitch ([267]).

To benefit from the strengths of both MLP-HMM and Gaussian-HMM techniques, the Tandem solution was proposed in [68], 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 ([326]), 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 ([197]). In ([145]), 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 [62], 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 [122].

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, ... 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
- combination of recognition system based on different acoustic features will be discussed in section ??

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 [102, 149] and/or improved training procedures.

The most common feature-reduction technique is the use of a linear transform $\mathbf{y} = \mathbf{A}\mathbf{x}$ where \mathbf{x} and \mathbf{y} are the original and the reduced feature vectors, respectively, and \mathbf{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) [59, 82] is the most simple way of finding \mathbf{A} . It allows for the best reconstruction of \mathbf{x} from \mathbf{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 [59, 82]. Here the columns of matrix \mathbf{A} are the eigenvectors belonging to the p largest eigenvalues of matrix $[\mathbf{S}_w^{-1}\mathbf{S}_b]$, where \mathbf{S}_w and \mathbf{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 [102]. In [102] 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) [149] overcomes this problem, and is

actually a generalization of LDA. The method usually requires the use of numerical optimization techniques to find the matrix \mathbf{A} . An exception is the method in [177], which uses the Chernoff distance to measure between-class distances and leads to a straight forward solution for \mathbf{A} . Finally, LDA and HDA can be combined with maximum likelihood linear transform (MLLT) [96], which is identical to semi-tied covariance matrices (STC) [86]. 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 (cfr. [228], section 2.1). It has been reported [244] that such a combination performs better than LDA or HDA alone. Also, HDA has been combined with minimum phoneme error (MPE) analysis [318]. 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 [213, 219].

A problem of the use of linear transforms for feature reduction is that the entire feature vector \mathbf{x} needs to be computed before the reduced vector \mathbf{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 \mathbf{A} , means that each row of \mathbf{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 \mathbf{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 [2], 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 [163]. A Maximum a Posteriori (MAP) objective function may also be used [40, 315].

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 [201, 217]. The eigenvoice approach [85, 208] 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 [140], rapid adaptation is further extended through a more accurate speaker space model, and an on-line algorithm is also proposed. In [312], 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 [139, 180, 283, 284, 312, 322].

Other forms of transformations for adaptation are also proposed in [218], 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 [3, 300].

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) [232, 234]. 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 [34, 183]. In [148], 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 [49, 195], 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 [224, 314] or used as a better seed for speaker adaptation [160]. Gender dependence is applied to whole word units, for example digits [101], or to context dependent phonetic units [224], 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 to large groups of speakers as in [18, 296]. 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 [1, 115, 172, 292].

Age dependent modeling has been less investigated, may be due to the lack of large size children speech corpora. The results presented in [48] 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 [307]. Recently, corpora for children speech recognition have begun to emerge. In [70] a small corpus of children speech was collected for use in interactive reading tutors and led to a complete children speech recognition system. In [257], 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 U.S. English.

Speaking rate notably affects the recognition performances, thus ROS dependent models were studied [194]. 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 [227], and that training procedures robust to sparse data need to be used. In that sense, comparative studies have shown that rate-adapted models per-

formed better than rate-specific models [311]. Speaking rate can be estimated on line [227], or computed from a decoding result using a generic set of acoustic models, in which case a rescaling is applied for fast or slow sentences [202]; or the various rate dependent models may be used simultaneously during decoding [39, 321].

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 [262] 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 [243] 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 [169, 222], and simulated stress token generation [26, 27]. 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 [239]. 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 [146]; and clustering training data at the utterance level provided the best performances in [256].

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 [205]; syllable boundary and stress tags in [223]; and voice characteristics in [271].

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 [31] where the acoustic densities are modulated depending on the probability of a master context HMM being in certain states. In [328], 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 [255], or hidden factors related to a clustering of the data [147, 189].

Multiple models can also be used in a parallel decoding framework [319]; then the final answer results from a "voting" process [74], or from the application of elaborated decision rules that take into account the recognized word hypotheses [14]. Multiple decoding is also useful for estimating reliable confidence measures [294].

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 to each speaker of the training data [302]. 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 performed by a maximum likelihood approach [161, 316] or from a direct estimation of the formant positions [64, 166]. More general normalization schemes have also been investigated [88], 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 [85]. A technique for "factoring-in" selected transformations back in the canonical model is also proposed in [87], 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 [11] making it much more robust in noisy environments [186]. 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 [276] and formant and auditory-based acoustic cues are used together with MFCC in [123, 252]. Correlation between pitch and acoustic features is taken into account in [144] and an LDA is applied on the full set of features (i.e. energy, MFCC, voicing and pitch) in [174]. In [52], the authors propose a 2-dimension HMM to extract the formant positions and evaluate their potential on a vowel classification task. In [90], 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 [279], 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, ...

Pitch has to be taken into account for the recognition of tonal languages. Tone can be modeled separately through specific HMMs [313] or decision trees [310], or the pitch parameter can be included in the feature vector [36], or both information streams (acoustic features and tonal features) can be handled directly by the decoder, possibly with different optimized weights [254]. 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 [171], and pitch tracking and voicing are investigated in [127]. A comparison of various modeling approaches is available in [53]. For tonal languages, pitch modeling usually concerns the whole syllable; however limiting the modeling to the vowel seems sufficient [37].

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 ro-

bust to noise [216].

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 [267] with pitch and energy as auxiliary features. Other information can also be taken into account such as articulatory information in [266] 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 [185].

Another approach for handling heterogeneous features is the TANDEM approach used with pitch, energy or rate of speech in [178]. 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 [260], or used to modify the parameters of the densities (during decoding) through multiple regressions as with pitch and speaking rate in [81].

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 [105] possibly depending on phone context. For handling hyper articulated speech where pauses may be inserted between syllables, ad hoc variants are necessary [189]. 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 [114, 253], based on phonological knowledge, data and recent studies on the syllabic structure of speech, for instance in English [99] or French [5].

In [5], 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 [99] 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 known 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 [99] 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 [156, 191]. These rely on the modeling capabilities of acoustic models that can implicitly model some pronunciation effect [60, 104, 136], provided that they are represented in the training data. In [104], 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 [6, 128]. Introducing multiple phonetic transcriptions that handle alterations produced by non-native speakers is a usual approach, and is generally associated to a combination of phone models of the native language with phone models of the target language [16, 24, 308]. However adding too many systematic pronunciation variants may be harmful [269].

Alteration rules can be defined from phonetic knowledge or estimated from some accented data [173]. Deriving rules using only native speech of both languages is proposed in [97]. [240] 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 [17], multilingual units can be used [292] or specialized models for different speaker groups can be elaborated [42]. Multilingual phone models have been investigated for many years in the hope of achieving language independent units [25, 47, 154, 249]. 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 [303] or determined from phonetic knowledge [291], 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 [79]. 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 [13, 187, 198, 255, 320]. 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 [188, 193, 198]. 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 [320], 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 [230].

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 [209], 1200 hours of training data have been used to develop acoustic models for the English broadcast news recognition task, with significant improvement over the previous 200 hours 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 [156], up to 2300 hours of speech have been used. This has been done as part of the EARS project, where training data of the order of 10000 hours 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 [298]. 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 [93].

⁴ [287] 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.

Finally, another approach is proposed in [61], 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 [268].

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.

6. Acknowledgments

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

- [1] S. Aalburg and H. Hoege. Foreign-accented speaker-independent speech recognition. In *Proc. of ICSLP*, pages 1465–1468, Jeju Island, Korea, September 2004.
- [2] Y. H. Abdel-Haleem, S. Renals, and N. D. Lawrence. Acoustic space dimensionality selection and combination using the maximum entropy principle. In *Proc. of ICASSP*, pages 637–640, Montreal, Canada, May 2004.
- [3] V. Abrash, A. Sankar, H. Franco, and M. Cohen. Acoustic adaptation using nonlinear transformations of HMM parameters. In *Proc. of ICASSP*, pages 729–732, Atlanta, GA, 1996.
- [4] K. Achan, S. Roweis, A. Hertzmann, and B. Frey. A segmental HMM for speech waveforms. Technical Report UTML Technical Report 2004-001, University of Toronto, Toronto, Canada, 2004.
- [5] M. Adda-Decker, P. Boula de Mareuil, G. Adda, and L. Lamel. Investigating syllabic structures and their variation in spontaneous french. *Speech Communication*, 46(2):119–139, June 2005.
- [6] M. Adda-Decker and L. Lamel. Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29(2):83–98, nov. 1999.
- [7] R. Andre-Obrecht. A new statistical approach for the automatic segmentation of continuous speech signals.

- IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(1):29–40, January 1988.
- [8] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer. In *Proc. of ICSLP*, pages 2037–2040, Denver, Colorado, September 2002.
 - [9] L.M. Arslan and J.H.L. Hansen. Language accent classification in american english. *Speech Communication*, 18(4):353–367, 1996.
 - [10] B. Atal. Efficient coding of LPC parameters by temporal decomposition. In *Proc. of ICASSP*, pages 81–84, Boston, USA, 1983.
 - [11] B. Atal and L. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. In *IEEE Trans. on Acoustics, Speech, and Signal Processing*, volume 24, 3, pages 201–212, June 1976.
 - [12] M. Athineos and D. Ellis. Frequency domain linear prediction for temporal features. In *Proc. of ASRU*, pages 261–266, St. Thomas, US Virgin Islands, USA, December 2003.
 - [13] E. G. Bard, C. Sotillo, M. L. Kelly, and M. P. Aylett. Taking the hit: leaving some lexical competition to be resolved post-lexically. *Lang. Cognit. Process.*, 15(5-6):731–737, 2001.
 - [14] L. Barrault, R. de Mori, R. Gemello, F. Mana, and D. Matrouf. Variability of automatic speech recognition systems using different features. In *Proc. of Interspeech*, pages 221–224, Lisboa, Portugal, 2005.
 - [15] K. Bartkova. Generating proper name pronunciation variants for automatic speech recognition. In *Proc. of ICPhS*, Barcelona, Spain, 2003.
 - [16] K. Bartkova and D. Jouvet. Language based phone model combination for asr adaptation to foreign accent. In *Proc. of ICPhS*, pages 1725–1728, San Francisco, USA, August 1999.
 - [17] K. Bartkova and D. Jouvet. Multiple models for improved speech recognition for non-native speakers. In *Proc. of SPECOM*, Saint Petersburg, Russia, September 2004.
 - [18] V. Beattie, S. Edmondson, D. Miller, Y. Patel, and G. Talvola. An integrated multidialect speech recognition system with optional speaker adaptation. In *Proc. of Eurospeech*, pages 1123–1126, Madrid, Spain, 1995.
 - [19] J. Q. Beauford. *Compensating for variation in speaking rate*. PhD thesis, Electrical Engineering, University of Pittsburgh, 1999.
 - [20] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024, February 2003.
 - [21] M.-G. Di Benedetto and J.-S. Liénard. Extrinsic normalization of vowel formant values based on cardinal vowels mapping. In *Proc. of ICSLP*, pages 579–582, Alberta, USA, 1992.
 - [22] C. Benitez, L. Burget, B. Chen, S. Dupont, H. Garudadri, H. Hermansky, P. Jain, S. Kajarekar, and S. Sivasdas. Robust asr front-end using spectral based and discriminant features: experiments on the aurora task. In *Proc. of Eurospeech*, pages 429–432, Aalborg, Denmark, september 2001.
 - [23] Mats Blomberg. Adaptation to a speaker’s voice in a speech recognition system based on synthetic phoneme references. *Speech Communication*, 10(5-6):453–461, 1991.
 - [24] P. Bonaventura, F. Gallochio, J. Mari, and G. Micca. Speech recognition methods for non-native pronunciation variants. In *Proc. ISCA Workshop on modelling pronunciation variations for automatic speech recognition*, pages 17–23, Rolduc, Netherlands, May 1998.
 - [25] P. Bonaventura, F. Gallochio, and G. Micca. Multilingual speech recognition for flexible vocabularies. In *Proc. of Eurospeech*, pages 355–358, Rhodes, Greece, 1997.
 - [26] S. E. Bou-Ghazale and J. L. H. Hansen. Duration and spectral based stress token generation for HMM speech recognition under stress. In *Proc. of ICASSP*, pages 413–416, Adelaide, Australia, 1994.
 - [27] S. E. Bou-Ghazale and J. L. H. Hansen. Improving recognition and synthesis of stressed speech via feature perturbation in a source generator framework. In *ECSCA-NATO Proc. Speech Under Stress Workshop*, pages 45–48, Lisbon, Portugal, 1995.
 - [28] H. Bourlard and D. Dupont. Sub-band based speech recognition. In *Proc. of ICASSP*, pages 1251–1254, Munich, Germany, 1997.
 - [29] B. Bozkurt and L. Couvreur. On the use of phase information for speech recognition. In *Proc. of Eusipco*, Antalya, Turkey, 2005.
 - [30] B. Bozkurt, B. Doval, C. d’Alessandro, and T. Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *IEEE Signal Processing Letters*, 12(4):344–347, 2005.
 - [31] F. Brugnara, R. De Mori, D. Giuliani, and M. Omologo. A family of parallel Hidden Markov Models. In *Proc. of ICASSP*, volume 1, pages 377–380, March 1992.
 - [32] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Z. Wei-Jin. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *IEEE Trans. on Speech and Audio Processing*, 12(4):420–435, July 2004.
 - [33] M. Carey, E. Parris, H. Lloyd-Thomas, and S. Bennett. Robust prosodic features for speaker identification. In *Proc. of ICSLP*, pages 1800–1803, Philadelphia, Pennsylvania, USA, 1996.

- [34] B. Carlson and M. Clements. Speech recognition in noise using a projection-based likelihood measure for mixture density HMMs. In *Proc. of ICASSP*, pages 237–240, San Francisco, CA, 1992.
- [35] L. Chase. *Error-Responsive Feedback Mechanisms for Speech Recognizers*. PhD thesis, Carnegie Mellon University, 1997.
- [36] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen. New methods in continuous mandarin speech recognition. In *Proc. of Eurospeech*, pages 1543–1546, 1997.
- [37] C.J. Chen, H. Li, L. Shen, and G. Fu. Recognize tone languages using pitch information on the main vowel of each syllable. In *Proc. of ICASSP*, volume 1, pages 61–64, May 2001.
- [38] Y. Chen. Cepstral domain stress compensation for robust speech recognition. In *Proc. of ICASSP*, pages 717–720, Dallas, TX, 1987.
- [39] C. Chesta, P. Laface, and F. Ravera. Connected digit recognition using short and long duration models. In *Proc. of ICASSP*, volume 2, pages 557–560, March 1999.
- [40] C. Chesta, O. Siohan, and C.-H. Lee. Maximum a posteriori linear regression for Hidden Markov Model adaptation. In *Proc. of Eurospeech*, pages 211–214, Budapest, Hungary, 1999.
- [41] G.F. Chollet, A.B.P. Astier, and M. Rossi. Evaluating the performance of speech recognizers at the acoustic-phonetic level. In *Proc. of ICASSP*, pages 758–761, Atlanta, USA, 1981.
- [42] T. Cincarek, R. Gruhn, and S. Nakamura. Speech recognition for multiple non-native accent groups with speaker-group-dependent acoustic models. In *Proc. of ICSLP*, pages 1509–1512, Jeju Island, Korea, October 2004.
- [43] R. R. Coifman and M. V. Wickerhauser. Entropy based algorithms for best basis selection. *IEEE Trans. on Information Theory*, 38(2):713–718, March 1992.
- [44] D. Colibro, L. Fissore, C. Popovici, C. Vair, and P. Laface. Learning pronunciation and formulation variants in continuous speech applications. In *Proc. of ICASSP*, pages 1001–1004, Philadelphia, PA, March 2005.
- [45] R. Cowie and R.R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication Special Issue on Speech and Emotions*, 40(1-2):5–32, 2003.
- [46] C. Cucchiari, H. Strik, and L. Boves. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication*, 30(2-3):109–119, February 2000.
- [47] P. Dalsgaard, O. Andersen, and W. Barry. Cross-language merged speech units and their descriptive phonetic correlates. In *Proc. of ICSLP*, pages 482–485, Sydney, Australia, 1998.
- [48] S. M. D’Arcy, L. P. Wong, and M. J. Russell. Recognition of read and spontaneous children’s speech using two new corpora. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.
- [49] S. Das, D. Lubensky, and C. Wu. Towards robust speech recognition in the telephony network environment - cellular and landline conditions. In *Proc. of Eurospeech*, pages 1959–1962, Budapest, Hungary, 1999.
- [50] S. Das, D. Nix, and M. Picheny. Improvements in children speech recognition performance. In *Proc. of ICASSP*, volume 1, pages 433–436, Seattle, USA, May 1998.
- [51] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 28:357–366, August 1980.
- [52] F. de Wet, K. Weber, L. Boves, B. Cranen, S. Bengio, and H. Bourlard. Evaluation of formant-like features on an automatic vowel classification task. *The Journal of the Acoustical Society of America*, 116(3):1781–1792, 2004.
- [53] T. Demeechai and K. Mäkeläinen. Recognition of syllables in a tone language. *Speech Communication*, 33(3):241–254, February 2001.
- [54] K. Demuynck, O. Garcia, and D. Van Compernelle. Synthesizing speech from speech recognition parameters. In *Proc. of ICSLP’04*, Jeju Island, Korea, October 2004.
- [55] Y. Deng, M. Mahajan, and A. Acero. Estimating speech recognition error rate without acoustic test data. In *Proc. of Eurospeech*, pages 929–932, Geneva, Switzerland, 2003.
- [56] Disfluency in spontaneous speech (diss’05), September 2005. Aix-en-Provence, France.
- [57] G. Doddington. Word alignment issues in asr scoring. In *Proc. of ASRU*, pages 630–633, US Virgin Islands, December 2003.
- [58] C. Draxler and S. Burger. Identification of regional variants of high german from digit sequences in german telephone speech. In *Proc. of Eurospeech*, pages 747–750, 1997.
- [59] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley, New York, 1973.
- [60] S. Dupont, C. Ris, L. Cuvreur, and J.-M. Boite. A study of implicit and explicit modeling of coarticulation and pronunciation variation. In *Proc. of Interspeech*, pages 1353–1356, Lisboa, Portugal, September 2005.
- [61] S. Dupont, C. Ris, O. Deroo, and S. Poitoux. Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents. In *Proc. of ASRU*, pages 29–34, San Juan, Puerto-Rico, December 2005.
- [62] E. Eide. Distinctive features for use in automatic speech recognition. In *Proc. of Eurospeech*, pages 1613–1616, Aalborg, Denmark, september 2001.

- [63] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. of ICASSP*, pages 346–348, Atlanta, GA, 1996.
- [64] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. of ICASSP*, pages 346–349, Atlanta, GA, 1996.
- [65] E. Eide, H. Gish, P. Jeanrenaud, and A. Mielke. Understanding and improving speech recognition performance through the use of diagnostic tools. In *Proc. of ICASSP*, pages 221–224, Detroit, Michigan, May 1995.
- [66] R. Eklund and A. Lindström. Xenophones: an investigation of phone set expansion in swedish and implications for speech recognition and speech synthesis. *Speech Communication*, 35(1-2):81–102, August 2001.
- [67] D. Elenius and M. Blomberg. Comparing speech recognition for adults and children. In *Proceedings of FONETIK 2004*, pages 156–159, Stockholm, Sweden, 2004.
- [68] D. Ellis, R. Singh, and S. Sivasdas. Tandem acoustic modeling in large-vocabulary recognition. In *Proc. of ICASSP*, pages 517–520, Salt Lake City, USA, May 2001.
- [69] M. Eskenazi. Detection of foreign speakers’ pronunciation errors for second language training-preliminary results. In *Proc. of ICSLP*, pages 1465–1468, Philadelphia, PA, 1996.
- [70] M. Eskenazi. Kids: a database of children’s speech. *The Journal of the Acoustical Society of America*, page 2759, December 1996.
- [71] M. Eskenazi and G. Pelton. Pinpointing pronunciation errors in children speech: examining the role of the speech recognizer. In *Proceedings of the PMLA Workshop*, Colorado, USA, September 2002.
- [72] R. Falthausen, T. Pfau, and G. Ruske. On-line speaking rate estimation using gaussian mixture models. In *Proc. of ICASSP*, pages 1355–1358, Istanbul, Turkey, 2000.
- [73] G. Fant. *Acoustic theory of speech production*. Mouton, The Hague, 1960.
- [74] J.G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *Proc. of ASRU*, pages 347–354, December 1997.
- [75] S. Fitt. The pronunciation of unfamiliar native and non-native town names. In *Proc. of Eurospeech*, pages 2227–2230, Madrid, Spain, 1995.
- [76] J. Flanagan. *Speech analysis and synthesis and perception*. Springer-Verlag, Berlin–Heidelberg–New York, 1972.
- [77] J. E. Flege, C. Schirru, and I. R. A. MacKay. Interaction between the native and second language phonetic subsystems. *Speech Communication*, 40:467–491, 2003.
- [78] E. Fosler-Lussier, I. Amdal, and H.-K. J. Kuo. A framework for predicting speech recognition errors. *Speech Communication*, 46(2):153–170, June 2005.
- [79] E. Fosler-Lussier and N. Morgan. Effects of speaking rate and word predictability on conversational pronunciations. *Speech Communication*, 29(2-4):137–158, 1999.
- [80] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen. Combination of machine scores for automatic grading of pronunciation quality. *Speech Communication*, 30(2-3):121–130, February 2000.
- [81] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama. Multiple-regression Hidden Markov Model. In *Proc. of ICASSP*, volume 1, pages 513–516, Salt Lake City, USA, May 2001.
- [82] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, New York, 1972.
- [83] P. Fung and W. K. Liu. Fast accent identification and accented speech recognition. In *Proc. of ICASSP*, pages 221–224, Phoenix, Arizona, USA, March 1999.
- [84] S. Furui, M. Beckman J.B. Hirschberg, S. Itahashi, T. Kawahara, S. Nakamura, and S. Narayanan. Introduction to the special issue on spontaneous speech processing. *IEEE Trans. on Speech and Audio Processing*, 12(4):349–350, July 2004.
- [85] M. J. F. Gales. Cluster adaptive training for speech recognition. In *Proc. of ICSLP*, pages 1783–1786, Sydney, Australia, 1998.
- [86] M. J. F. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Trans. on Speech and Audio Processing*, 7:272–281, 1999.
- [87] M. J. F. Gales. Acoustic factorization. In *Proc. of ASRU*, Madonna di Campiglio, Italy, 2001.
- [88] M. J. F. Gales. Multiple-cluster adaptive training schemes. In *Proc. of ICASSP*, pages 361–364, Salt Lake City, Utah, USA, 2001.
- [89] Y. Gao, B. Ramabhadran, J. Chen, H. Erdogan, and M. Picheny. Innovative approaches for large vocabulary name recognition. In *Proc. of ICASSP*, pages 333–336, Salt Lake City, Utah, May 2001.
- [90] P. Garner and W. Holmes. On the robust incorporation of formant features into hidden markov models for automatic speech recognition. In *Proc. of ICASSP*, pages 1–4, 1998.
- [91] P. L. Garvin and P. Ladefoged. Speaker identification and message identification in speech recognition. *Phonetica*, 9:193–199, 1963.
- [92] R. Gemello, F. Mana, D. Albesano, and R. De Mori. Multiple resolution analysis for robust automatic speech recognition. *Computer, Speech and Language*, 20:2–21, 2006.
- [93] A. Girardi, K. Shikano, and S. Nakamura. Creating speaker independent HMM models for restricted database using straight-tempo morphing. In *Proc. of ICSLP*, pages 687–690, Sydney, Australia, 1998.

- [94] D. Giuliani and M. Gerosa. Investigating recognition of children speech. In *Proc. of ICASSP*, pages 137–140, Hong Kong, April 2003.
- [95] V. Goel, S. Kumar, and W. Byrne. Segmental minimum Bayes-risk decoding for automatic speech recognition. *Trans. of IEEE Speech and Audio Processing*, 12(3):234–249, 2004.
- [96] R. A. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. In *Proc. of ICASSP*, pages 661–664, Seattle, WA, 1998.
- [97] S. Goronzy, S. Rapp, and R. Kompe. Generating non-native pronunciation variants for lexicon adaptation. *Speech Communication*, 42(1):109–123, January 2004.
- [98] M. Graciarena, H. France, J. Zheng, D. Vergyri, and A. Stolcke. Voicing feature integration in SRI’s DECI-PHER LVCSR system. In *Proc. of ICASSP*, pages 921–924, Montreal, Canada, 2004.
- [99] S. Greenberg and S. Chang. Linguistic dissection of switchboard-corpus automatic speech recognition systems. In *Proc. of ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millenium*, Paris, France, September 2000.
- [100] S. Greenberg and E. Fosler-Lussier. The uninvited guest: information’s role in guiding the production of spontaneous speech. In *Proceedings of the Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, Kloster Seeon, Germany, 2000.
- [101] S.K. Gupta, F. Soong, and R. Haimi-Cohen. High-accuracy connected digit recognition for mobile applications. In *Proc. of ICASSP*, volume 1, pages 57–60, May 1996.
- [102] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proc. of ICASSP*, pages 13–16, San Francisco, CA, 1992.
- [103] A. Hagen, B. Pellom, and R. Cole. Children speech recognition with application to interactive books and tutors. In *Proc. of ASRU*, pages 186–191, St. Thomas, U.S. Virgin Islands, November 2003.
- [104] T. Hain. Implicit modelling of pronunciation variation in automatic speech recognition. *Speech Communication*, 46(2):171–188, June 2005.
- [105] T. Hain and P. C. Woodland. Dynamic HMM selection for continuous speech recognition. In *Proc. of Eurospeech*, pages 1327–1330, Budapest, Hungary, 1999.
- [106] J. H. L. Hansen. Evaluation of acoustic correlates of speech under stress for robust speech recognition. In *IEEE Proc. 15th Northeast Bioengineering Conf, Boston, MA*, pages 31–32, Boston, Mass., 1989.
- [107] J. H. L. Hansen. Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments. In *Proc. of ICASSP*, pages 95–98, Minneapolis, Minnesota, 1993.
- [108] J. H. L. Hansen. 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*, 1995.
- [109] J. H. L. Hansen. 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, nov. 1996.
- [110] B. A. Hanson and T. Applebaum. Robust speaker-independent word recognition using instantaneous, dynamic and acceleration features: experiments with Lombard and noisy speech. In *Proc. of ICASSP*, pages 857–860, Albuquerque, New Mexico, 1990.
- [111] R. Hariharan, I. Kiss, and O. Viikki. Noise robust speech parameterization using multiresolution feature extraction. *IEEE Trans. on Speech and Audio Processing*, 9(8):856–865, 2001.
- [112] S. Haykin. *Adaptive filter theory*. Prentice-Hall Publishers, N.J., USA., 1993.
- [113] S. Haykin. *Communication systems*. John Wiley and Sons, New York, USA, 3 edition, 1994.
- [114] T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu. Pronunciation modeling using a finite-state transducer representation. *Speech Communication*, 46(2):189–203, June 2005.
- [115] X. He and Y. Zhao. Fast model selection based speaker adaptation for nonnative speech. *IEEE Trans. on Speech and Audio Processing*, 11(4):298–307, July 2003.
- [116] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde. Continuous speech recognition using joint features derived from the modified group delay function and mfcc. In *Proc. of ICSLP*, pages 905–908, Jeju, Korea, October 2004.
- [117] R. M. Hegde, H. A. Murthy, and G. V. R. Rao. Speech processing using joint features derived from the modified group delay function. In *Proc. of ICASSP*, volume I, pages 541–544, Philadelphia, PA, 2005.
- [118] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.
- [119] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing*, 2(4):578–589, October 1994.
- [120] H. Hermansky and S. Sharma. TRAPS: classifiers of temporal patterns. In *Proc. of ICSLP*, pages 1003–1006, Sydney, Australia, 1998.
- [121] L. Hetherington. New words: Effect on recognition performance and incorporation issues. In *Proc. of Eurospeech*, pages 1645–1648, Madrid, Spain, 1995.
- [122] J. Hirschberg, D. Litman, and M. Swerts. Prosodic and other cues to speech recognition failures. *Speech Communication*, 43(1-2):155–175, 2004.

- [123] J. N. Holmes, W. J. Holmes, and P. N. Garner. Using formant frequencies in speech recognition. In *Proc. of Eurospeech*, pages 2083–2086, Rhodes, Greece, 1997.
- [124] H. W. Hon and K. Wang. Combining frame and segment based models for large vocabulary continuous speech recognition. In *Proc. of ASRU*, Keystone, Colorado, 1999.
- [125] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou. Analysis of speaker variability. In *Proc. of Eurospeech*, pages 1377–1380, Aalborg, Denmark, September 2001.
- [126] X. Huang and K. Lee. On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition. In *Proc. of ICASSP*, pages 877–880, Toronto, Canada, 1991.
- [127] H. C.-H. Huank and F. Seide. Pitch tracking and tone features for Mandarin speech recognition. In *Proc. of ICASSP*, volume 3, pages 1523–1526, June 2000.
- [128] J. J. Humphries, P. C. Woodland, and D. Pearce. Using accent-specific pronunciation modelling for robust speech recognition. In *Proc. of ICSLP*, pages 2367–2370, Rhodes, Greece, 1996.
- [129] M. J. Hunt. Spectral signal processing for asr. In *Proc. of ASRU*, Keystone, Colorado, dec. 1999.
- [130] M. J. Hunt. Speech recognition, syllabification and statistical phonetics. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.
- [131] M. J. Hunt and C. Lefebvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. In *Proc. of ICASSP*, pages 262–265, Glasgow, UK, 1989.
- [132] A. Iivonen, K. Harinen, L. Keinanen, J. Kirjavainen, E. Meister, and L. Tuuri. Development of a multiparametric speaker profile for speaker recognition. In *Proc. of ICPhS*, pages 695–698, Barcelona, Spain, 2003.
- [133] E. Janse. Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. *Speech Communication*, 42(2):155–173, February 2004.
- [134] K. Jiang and X. Huang. Acoustic feature selection using speech recognizers. In *Proc. of ASRU*, Keystone, Colorado, 1999.
- [135] B.-H. Juang, L. R. Rabiner, and J. G. Wilpon. On the use of bandpass liftering in speech recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 35:947–953, 1987.
- [136] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiyang, and Z. Sen. What kind of pronunciation variation is hard for triphones to model? In *Proc. of ICASSP*, pages 577–580, Salt Lake City, Utah, May 2001.
- [137] S. Kajarekar, N. Malayath, and H. Hermansky. Analysis of sources of variability in speech. In *Proc. of Eurospeech*, pages 343–346, Budapest, Hungary, September 1999.
- [138] S. Kajarekar, N. Malayath, and H. Hermansky. Analysis of speaker and channel variability in speech. In *Proc. of ASRU*, Keystone, Colorado, December 1999.
- [139] P. Kenny, G. Boulianne, and P. Dumouchel. Eigen-voice modeling with sparse training data. *IEEE Trans. on Speech and Audio Processing*, 13(3):345–354, May 2005.
- [140] D. K. Kim and N. S. Kim. Rapid online adaptation using speaker space model evolution. *Speech Communication*, 42(3-4):467–478, April 2004.
- [141] B. Kingsbury, N. Morgan, and S. Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1-3):117–132, August 1998.
- [142] B. Kingsbury, G. Saon, L. Mangua, M. Padmanabhan, and R. Sarikaya. Robust speech recognition in noisy environments: the 2001 IBM SPINE evaluation system. In *Proc. of ICASSP*, volume I, pages 53 – 56, Orlando, FL, 2002.
- [143] K. Kirchhoff. Combining articulatory and acoustic information for speech recognition in noise and reverberant environments. In *Proc. of ICSLP*, pages 891–894, Sydney, Australia, 1998.
- [144] N. Kitaoka, D. Yamada, and S. Nakagawa. Speaker independent speech recognition using features based on glottal sound source. In *Proc. of ICSLP*, pages 2125–2128, Denver, USA, September 2002.
- [145] M. Kleinschmidt and D. Gelbart. Improving word accuracy with gabor feature extraction. In *Proc. of ICSLP*, pages 25–28, Denver, Colorado, 2002.
- [146] F. Korkmazskiy, B.-H. Juang, and F. Soong. Generalized mixture of HMMs for continuous speech recognition. In *Proc. of ICASSP*, volume 2, pages 1443–1446, April 1997.
- [147] F. Korkmazsky, M. Deviren, D. Fohr, and I. Illina. Hidden factor dynamic bayesian networks for speech recognition. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.
- [148] F. Kubala, A. Anastasakos, J. Makhoul, L. Nguyen, R. Schwartz, and E. Zavalagkos. Comparative experiments on large vocabulary speech recognition. In *Proc. of ICASSP*, pages 561–564, Adelaide, Australia, April 1994.
- [149] N. Kumar and A. G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, 26(4):283–297, 1998.
- [150] R. Kumaresan. An inverse signal approach to computing the envelope of a real valued signal. *IEEE Signal Processing Letters*, 5(10):256–259, October 1998.
- [151] R. Kumaresan and A. Rao. 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, March 1999.

- [152] K. Kumpf and R.W. King. Automatic accent classification of foreign accented australian english speech. In *Proc. of ICSLP*, pages 1740–1743, Philadelphia, PA, October 1996.
- [153] H. Kuwabara. Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate. In *Proc. of Eurospeech*, pages 1003–1006, Rhodes, Greece, 1997.
- [154] J. Köhler. Multilingual phonemes recognition exploiting acoustic-phonetic similarities of sounds. In *Proc. of ICSLP*, pages 2195–2198, Philadelphia, PA, 1996.
- [155] P. Ladefoged and D.E. Broadbent. Information conveyed by vowels. *The Journal of the Acoustical Society of America*, 29:98–104, 1957.
- [156] L. Lamel and J.-L. Gauvain. Alternate phone models for conversational speech. In *Proc. of ICASSP*, pages 1005–1008, Philadelphia, Pennsylvania, 2005.
- [157] J. Laver. *Principles of phonetics*. Cambridge University Press, Cambridge, 1994.
- [158] A. D. Lawson, D. M. Harris, and J. J. Grieco. Effect of foreign accent on speech recognition in the NATO N-4 corpus. In *Proc. of Eurospeech*, pages 1505–1508, Geneva, Switzerland, 2003.
- [159] C. Lee, C. Lin, and B. Juang. A study on speaker adaptation of the parameters of continuous density Hidden Markov Models. *IEEE Trans. Signal Processing*, 39(4):806–813, April 1991.
- [160] C.-H. Lee and J.-L. Gauvain. Speaker adaptation based on MAP estimation of HMM parameters. In *Proc. of ICASSP*, volume 2, pages 558–561, April 1993.
- [161] L. Lee and R. C. Rose. Speaker normalization using efficient frequency warping procedures. In *Proc. of ICASSP*, volume 1, pages 353–356, Atlanta, Georgia, May 1996.
- [162] S. Lee, A. Potamianos, and S. Narayanan. Acoustics of children speech: developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105:1455–1468, March 1999.
- [163] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models. *Computer, Speech and Language*, 9(2):171–185, April 1995.
- [164] R. G. Leonard. A database for speaker independent digit recognition. In *Proc. of ICASSP*, pages 328–331, San Diego, US, March 1984.
- [165] X. Lin and S. Simske. Phoneme-less hierarchical accent classification. In *Proc. of Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 1801–1804, Pacific Grove, CA, November 2004.
- [166] M. Lincoln, S.J. Cox, and S. Ringland. A fast method of speaker normalisation using formant estimation. In *Proc. of Eurospeech*, pages 2095–2098, Rhodes, Greece, 1997.
- [167] B. Lindblom. Explaining phonetic variation: a sketch of the H&H theory. In W.J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*. Kluwer Academic Publishers, 1990.
- [168] R. P. Lippmann. Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15, July 1997.
- [169] R. P. Lippmann, E.A. Martin, and D.B. Paul. Multi-style training for robust isolated-word speech recognition. In *Proc. of ICASSP*, pages 705–708, Dallas, TX, April 1987.
- [170] L. Liu, J. He, and G. Palm. Effects of phase on the perception of intervocalic stop consonants. *Speech Communication*, 22(4):403–417, 1997.
- [171] S. Liu, S. Doyle, A. Morris, and F. Ehsani. The effect of fundamental frequency on mandarin speech recognition. In *Proc. of ICSLP*, volume 6, pages 2647–2650, Sydney, Australia, Nov/Dec 1998.
- [172] W. K. Liu and P. Fung. MLLR-based accent model adaptation without accented data. In *Proc. of ICSLP*, volume 3, pages 738–741, Beijing, China, September 2000.
- [173] K. Livescu and J. Glass. Lexical modeling of non-native speech for automatic speech recognition. In *Proc. of ICASSP*, volume 3, pages 1683–1686, Istanbul, Turkey, June 2000.
- [174] A. Ljolje. Speech recognition using fundamental frequency and voicing in acoustic modeling. In *Proc. of ICSLP*, pages 2137–2140, Denver, USA, September 2002.
- [175] A. F. Litjens and A.W. Black. Knowledge of language origin improves pronunciation accuracy of proper names. In *Proc. of Eurospeech*, Aalborg, Denmark, September 2001.
- [176] E. Lombard. Le signe de l’élévation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37, 1911.
- [177] M. Loog and R. P. W. Duin. Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(6):732–739, June 2004.
- [178] M. Magimai-Doss, T. A. Stephenson, S. Ikbāl, and H. Bourlard. Modelling auxiliary features in tandem systems. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.
- [179] B. Maison. Pronunciation modeling for names of foreign origin. In *Proc. of ASRU*, pages 429–434, US Virgin Islands, December 2003.
- [180] B. Mak and R. Hsiao. Improving eigenspace-based mllr adaptation by kernel PCA. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.
- [181] J. Makhoul. Linear prediction: a tutorial review. *Proceedings of IEEE*, 63(4):561–580, April 1975.
- [182] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: Word-error minimization and other applications of confusion networks. *Computer Speech and Language*, 14(4):373–400, 2000.

- [183] D. Mansour and B.H. Juang. A family of distortion measures based upon projection operation for robust speech recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37:1659–1671, November 1989.
- [184] J. Markel, B. Oshika, and A.H. Gray. Long-term feature averaging for speaker recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 25:330–337, August 1977.
- [185] K. Markov and S. Nakamura. Hybrid HMM/BN LVCSR system integrating multiple acoustic features. In *Proc. of ICASSP*, volume 1, pages 840–843, April 2003.
- [186] A. Martin and L. Mauuary. Voicing parameter and energy-based speech/non-speech detection for speech recognition in adverse conditions. In *Proc. of Eurospeech*, pages 3069–3072, Geneva, Switzerland, September 2003.
- [187] F. Martinez, D. Tapias, and J. Alvarez. Towards speech rate independence in large vocabulary continuous speech recognition. In *Proc. of ICASSP*, pages 725–728, Seattle, Washington, May 1998.
- [188] F. Martinez, D. Tapias, J. Alvarez, and P. Leon. Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In *Proc. of Eurospeech*, pages 469–472, Rhodes, Greece, 1997.
- [189] S. Matsuda, T. Jitsuhiro, K. Markov, and S. Nakamura. Speech recognition system robust to noise and speaking styles. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.
- [190] A. Mertins and J. Rademacher. Vocal tract length invariant features for automatic speech recognition. In *Proc. of ASRU*, pages 308–312, Cancun, Mexico, December 2005.
- [191] R. Messina and D. Jouvet. Context dependent long units for speech recognition. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.
- [192] B. P. Milner. Inclusion of temporal information into features for speech recognition. In *Proc. of ICSLP*, pages 256–259, Philadelphia, PA, October 1996.
- [193] N. Mirghafori, E. Fosler, and N. Morgan. Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes. In *Proc. of Eurospeech*, pages 491–494, Madrid, Spain, September 1995.
- [194] N. Mirghafori, E. Fosler, and N. Morgan. Towards robustness to fast speech in ASR. In *Proc. of ICASSP*, pages 335–338, Atlanta, Georgia, May 1996.
- [195] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin, and K. Bartkova. Towards improving ASR robustness for PSN and GSM telephone applications. *Speech Communication*, 23(1-2):141–159, October 1997.
- [196] P. Mokhtari. *An acoustic-phonetic and articulatory study of speech-speaker dichotomy*. PhD thesis, The University of New South Wales, Canberra, Australia, 1998.
- [197] N. Morgan, B. Chen, Q. Zhu, and A. Stolcke. TRAPping conversational speech: extending TRAP/tandem approaches to conversational telephone speech recognition. In *Proc. of ICASSP*, volume 1, pages 536–539, Montreal, Canada, 2004.
- [198] N. Morgan, E. Fosler, and N. Mirghafori. Speech recognition using on-line estimation of speaking rate. In *Proc. of Eurospeech*, volume 4, pages 2079–2082, Rhodes, Greece, September 1997.
- [199] N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proc. of ICASSP*, pages 729–732, Seattle, May 1998.
- [200] I. R. Murray and J. L. Arnott. 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, 1993.
- [201] Y. S. Masaki Naito and Li Deng. Speaker clustering for speech recognition using the parameters characterizing vocal-tract dimensions. In *Proc. of ICASSP*, pages 1889–1893, Seattle, WA, 1998.
- [202] H. Nanjo and T. Kawahara. Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition. In *Proc. of ICASSP*, volume 1, pages 725–728, Orlando, FL, 2002.
- [203] H. Nanjo and T. Kawahara. Language model and speaking rate adaptation for spontaneous presentation speech recognition. *IEEE Trans. on Speech and Audio Processing*, 12(4):391–400, July 2004.
- [204] T. M. Nearey. *Phonetic feature systems for vowels*. Indiana University Linguistics Club, Bloomington, Indiana, USA, 1978.
- [205] C. Neti and S. Roukos. Phone-context specific gender-dependent acoustic-models for continuous speech recognition. In *Proc. of ASRU*, pages 192–198, Santa Barbara, CA, December 1997.
- [206] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. Automatic scoring of pronunciation quality. *Speech Communication*, 30(2-3):83–93, February 2000.
- [207] Leonardo Neumeyer, Horacio Franco, Mitchel Weintraub, and Patti Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *Proc. of ICSLP*, pages 1457–1460, Philadelphia, PA, September 1996.
- [208] P. Nguyen, R. Kuhn, J.-C. Junqua, N. Niedzielski, and C. Wellekens. Eigenvoices : a compact representation of speakers in a model space. *Annales des Télécommunications*, 55(3–4), March-April 2000.
- [209] P. Nguyen, L. Rigazio, and J.-C. Junqua. Large corpus experiments for broadcast news recognition. In *Proc. of Eurospeech*, pages 1837–1840, Geneva, Switzerland, September 2003.
- [210] F. Nolan. *The phonetic bases of speaker recognition*. Cambridge University Press, Cambridge, 1983.

- [211] National Institute of Standards and Technology. SCLITE scoring software. <ftp://jaguar.nsls.nist.gov/pub/sctk-1.2.tar.Z>, 2001.
- [212] M. Kamal Omar, K. Chen, M. Hasegawa-Johnson, and Y. Bradman. An evaluation of using mutual information for selection of acoustic features representation of phonemes for speech recognition. In *Proc. of ICSLP*, pages 2129–2132, Denver, CO, 2002.
- [213] M. Kamal Omar and M. Hasegawa-Johnson. Maximum mutual information based acoustic features representation of phonological features for speech recognition. In *Proc. of ICASSP*, volume 1, pages 81–84, Montreal, Canada, 2002.
- [214] Y. Ono, H. Wakita, and Y. Zhao. Speaker normalization using constrained spectra shifts in auditory filter domain. In *Proc. of Eurospeech*, pages 355–358, Berlin, Germany, 1993.
- [215] D. O'Shaughnessy. *Speech communication - human and machine*. Addison-Wesley, 1987.
- [216] D. O'Shaughnessy and H. Tolba. Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision. In *Proc. of ICASSP*, volume 1, pages 413–416, Phoenix, Arizona, March 1999.
- [217] M. Padmanabhan, L. Bahl, D. Nahamoo, and M. Picheny. Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems. In *Proc. of ICASSP*, pages 701–704, Atlanta, GA, May 1996.
- [218] M. Padmanabhan and S. Dharanipragada. Maximum-likelihood nonlinear transformation for acoustic adaptation. *IEEE Trans. on Speech and Audio Processing*, 12(6):572–578, November 2004.
- [219] M. Padmanabhan and S. Dharanipragada. Maximizing information content in feature extraction. *IEEE Trans. on Speech and Audio Processing*, 13(4):512–519, July 2005.
- [220] K.K. Paliwal and L. Alsteris. Usefulness of phase spectrum in human speech perception. In *Proc. of Eurospeech*, pages 2117–2120, Geneva, Switzerland, 2003.
- [221] K.K. Paliwal and B.S. Atal. Frequency-related representation of speech. In *Proc. of Eurospeech*, pages 65–68, Geneva, Switzerland, 2003.
- [222] D. B. Paul. A speaker-stress resistant HMM isolated word recognizer. In *Proc. of ICASSP*, pages 713–716, Dallas, Texas, 1987.
- [223] D.B. Paul. Extensions to phone-state decision-tree clustering: single tree and tagged clustering. In *Proc. of ICASSP*, volume 2, pages 1487–1490, Munich, Germany, April 1997.
- [224] J.J. Odell P.C. Woodland and, V. Valtchev, and S.J. Young. Large vocabulary continuous speech recognition using HTK. In *Proc. of ICASSP*, volume 2, pages 125–128, Adelaide, Australia, April 1994.
- [225] S. D. Peters, P. Stubbley, and J.-M. Valin. On the limits of speech recognition in noise. In *Proc. of ICASSP'99*, pages 365–368, Phoenix, Arizona, May 1999.
- [226] G. E. Peterson and H. L. Barney. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24:175–184, 1952.
- [227] T. Pfau and G. Ruske. Creating Hidden Markov Models for fast speech. In *Proc. of ICSLP*, Sydney, Australia, 1998.
- [228] Michael Pitz. *Investigations on Linear Transformations for Speaker Adaptation and Normalization*. PhD thesis, RWTH Aachen University, March 2005.
- [229] M. Plumpe, T. Quatieri, and D. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 7(5):569–586, September 1999.
- [230] ISCA Tutorial and Research Workshop, Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA-2002), September 2002.
- [231] L. C. W. Pols, L. J. T. Van der Kamp, and R. Plomp. Perceptual and physical space of vowel sounds. *The Journal of the Acoustical Society of America*, 46:458–467, 1969.
- [232] G. Potamianos and S. Narayanan. Robust recognition of children speech. *IEEE Trans. on Speech and Audio Processing*, 11:603–616, November 2003.
- [233] G. Potamianos, S. Narayanan, and S. Lee. Analysis of children speech: duration, pitch and formants. In *Proc. of Eurospeech*, pages 473–476, Rhodes, Greece, September 1997.
- [234] G. Potamianos, S. Narayanan, and S. Lee. Automatic speech recognition for children. In *Proc. of Eurospeech*, pages 2371–2374, Rhodes, Greece, September 1997.
- [235] R. K. Potter and J. C. Steinberg. Toward the specification of speech. *The Journal of the Acoustical Society of America*, 22:807–820, 1950.
- [236] H. Printz and P.A. Olsen. Theory and practice of acoustic confusability. *Computer Speech and Language*, 16(1):131–164, January 2002.
- [237] P. Pujol, S. Pol, C. Nadeu, A. Hagen, and H. Bourlard. Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system. *IEEE Trans. on Speech and Audio Processing*, SAP-13(1):14–22, January 2005.
- [238] L. Rabiner and B. H. Juang. *Fundamentals of speech recognition*, chapter 2, pages 20–37. Prentice Hall PTR, Englewood Cliffs, NJ, USA, 1993.
- [239] L.R. Rabiner, C.H. Lee, B.H. Juang, and J.G. Wilpon. HMM clustering for connected word recognition. In *Proc. of ICASSP*, volume 1, pages 405–408, Glasgow, Scotland, May 1989.
- [240] A. Raux. Automated lexical adaptation and speaker clustering based on pronunciation habits for non-native speech recognition. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.

- [241] ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, May 1998.
- [242] S. Saito and F. Itakura. Frequency spectrum deviation between speakers. *Speech Communication*, 2:149–152, 1983.
- [243] S. Sakauchi, Y. Yamaguchi, S. Takahashi, and S. Kobashikawa. Robust speech recognition based on HMM composition and modified Wiener filter. In *Proc. of Interspeech*, pages 2053–2056, Jeju Island, Korea, October 2004.
- [244] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen. Maximum likelihood discriminant feature spaces. In *Proc. of ICASSP*, pages 1129–1132, Jun 2000.
- [245] T. Schaaf and T. Kemp. Confidence measures for spontaneous speech recognition. In *Proc. of ICASSP*, pages 875–878, Munich, Germany, April 1997.
- [246] K.R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication Special Issue on Speech and Emotions*, 40(1-2):227–256, 2003.
- [247] S. Schimmel and L. Atlas. Coherent envelope detection for modulation filtering of speech. In *Proc. of ICASSP*, volume 1, pages 221–224, Philadelphia, USA, March 2005.
- [248] M.R. Schroeder and H.W. Strube. Flat-spectrum speech. *The Journal of the Acoustical Society of America*, 79(5):1580–1583, 1986.
- [249] T. Schultz and A. Waibel. Language independent and language adaptive large vocabulary speech recognition. In *Proc. of ICSLP*, volume 5, pages 1819–1822, Sydney, Australia, 1998.
- [250] R. Schwartz, C. Barry, Y.-L. Chow, A. Deft, M.-W. Feng, O. Kimball, F. Kubala, J. Makhoul, and J. Vandegrift. The BBN BYBLOS continuous speech recognition system. In *Proc. of Speech and Natural Language Workshop*, pages 21–23, Philadelphia, Pennsylvania, 1989.
- [251] S. Schötz. A perceptual study of speaker age. In *Working paper 49*, pages 136–139, Lund University, Dept Of Linguistic, November 2001.
- [252] S.-A. Selouani, H. Tolba, and D. O’Shaughnessy. Distinctive features, formants and cepstral coefficients to improve automatic speech recognition. In *Conference on Signal Processing, Pattern Recognition and Applications, IASTED*, pages 530–535, Crete, Greece, 2002.
- [253] S. Seneff and C. Wang. Statistical modeling of phonological rules through linguistic hierarchies. *Speech Communication*, 46(2):204–216, June 2005.
- [254] Y. Y. Shi, J. Liu, and R.S. Liu. Discriminative HMM stream model for Mandarin digit string speech recognition. In *Proc. of Int. Conf. on Signal Processing*, volume 1, pages 528–531, Beijing, China, August 2002.
- [255] T. Shinozaki and S. Furui. Hidden mode HMM using bayesian network for modeling speaking rate fluctuation. In *Proc. of ASRU*, pages 417–422, US Virgin Islands, December 2003.
- [256] T. Shinozaki and S. Furui. Spontaneous speech recognition using a massively parallel decoder. In *Proc. of ICSLP*, pages 1705–1708, Jeju Island, Korea, October 2004.
- [257] K. Shobaki, J.-P. Hosom, and R. Cole. The OGI kids speech corpus and recognizers. In *Proc. of ICSLP*, pages 564–567, Beijing, China, October 2000.
- [258] M. A. Siegler. *Measuring and compensating for the effects of speech rate in large vocabulary continuous speech recognition*. PhD thesis, Carnegie Mellon University, 1995.
- [259] M. A. Siegler and R. M. Stern. On the effect of speech rate in large vocabulary speech recognition system. In *Proc. of ICASSP*, pages 612–615, Detroit, Michigan, May 1995.
- [260] H. Singer and S. Sagayama. Pitch dependent phone modelling for HMM based speech recognition. In *Proc. of ICASSP*, volume 1, pages 273–276, San Francisco, CA, March 1992.
- [261] J. Slifka and T. R. Anderson. Speaker modification with LPC pole analysis. In *Proc. of ICASSP*, pages 644–647, Detroit, MI, May 1995.
- [262] M. G. Song, H.I. Jung, K.-J. Shim, and H. S. Kim. Speech recognition in car noise environments using multiple models according to noise masking levels. In *Proc. of ICSLP*, 1998.
- [263] C. Sotillo and E. G. Bard. Is hypo-articulation lexically constrained? In *Proc. of SPoSS*, pages 109–112, Aix-en-Provence, September 1998.
- [264] J. J. Sroka and L. D. Braida. Human and machine consonant recognition. *Speech Communication*, 45(4):401–423, March 2005.
- [265] H.J.M. Steeneken and J.G. van Velden. Objective and diagnostic assessment of (isolated) word recognizers. In *Proc. of ICASSP*, volume 1, pages 540–543, Glasgow, UK, May 1989.
- [266] T. A. Stephenson, H. Bourlard, S. Bengio, and A. C. Morris. Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables. In *Proc. of ICSLP*, volume 2, pages 951–954, Beijing, China, October 2000.
- [267] T. A. Stephenson, M. M. Doss, and H. Bourlard. Speech recognition with auxiliary information. *IEEE Trans. on Speech and Audio Processing*, SAP-12(3):189–203, 2004.
- [268] A. Stolcke, F. Grezl, M.-Y. Hwang, N. Morgan, and D. Vergyri. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In *Proc. of ICASSP*, volume 1, pages 321–324, Toulouse, France, May 2006.

- [269] H. Strik and C. Cucchiaroni. Modeling pronunciation variation for ASR: a survey of the literature. *Speech Communication*, 29(2-4):225–246, November 1999.
- [270] D.X. Sun and L. Deng. Analysis of acoustic-phonetic variations in fluent speech using Timit. In *Proc. of ICASSP*, pages 201–204, Detroit, Michigan, May 1995.
- [271] H. Suzuki, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura. Speech recognition using voice-characteristic-dependent acoustic models. In *Proc. of ICASSP*, volume 1, pages 740–743, Hong-Kong (canceled), April 2003.
- [272] T. Svendsen, K. K. Paliwal, E. Harborg, and P. O. Husoy. An improved sub-word based speech recognizer. In *Proc. of ICASSP*, pages 108–111, Glasgow, UK, April 1989.
- [273] T. Svendsen and F. Soong. On the automatic segmentation of speech signals. In *Proc. of ICASSP*, pages 77–80, Dallas, Texas, April 1987.
- [274] C. Teixeira, I. Trancoso, and A. Serralheiro. Accent identification. In *Proc. of ICSLP*, volume 3, pages 1784–1787, Philadelphia, PA, October 1996.
- [275] D. L. Thomson and R. Chengalvarayan. Use of periodicity and jitter as speech recognition feature. In *Proc. of ICASSP*, volume 1, pages 21–24, Seattle, WA, May 1998.
- [276] D. L. Thomson and R. Chengalvarayan. Use of voicing features in HMM-based speech recognition. *Speech Communication*, 37(3-4):197–211, July 2002.
- [277] S. Tibrewala and H. Hermansky. Sub-band based recognition of noisy speech. In *Proc. of ICASSP*, pages 1255–1258, Munich Germany, April 1997.
- [278] H. Tolba, S. A. Selouani, and D. O’Shaughnessy. Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm. In *Proc. of ICASSP*, pages 837 – 840, Orlando, FL, May 2002.
- [279] H. Tolba, S.A. Selouani, and D. O’Shaughnessy. 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 *Proc. of Eurospeech*, pages 3085–3088, Geneva, Switzerland, 2003.
- [280] M. J. Tomlinson, M. J. Russell, R. K. Moore, A. P. Buckland, and M. A. Fawley. Modelling asynchrony in speech using elementary single-signal decomposition. In *Proc. of ICASSP*, pages 1247–1250, Munich Germany, April 1997.
- [281] B. Townshend, J. Bernstein, O. Todric, and E. Warren. Automatic text-independent pronunciation scoring of foreign language student speech. In *Proc. of STIL-1998*, pages 179–182, Stockholm, May 1998.
- [282] H. Trauttmüller. Perception of speaker sex, age and vocal effort. In *technical report*, Institutionen för lingvistik, Stockholm universitet, 1997.
- [283] S. Tsakalidis, V. Doumptiotis, and W. Byrne. Discriminative linear transforms for feature normalization and speaker adaptation in HMM estimation. *IEEE Trans. on Speech and Audio Processing*, 13(3):367–376, May 2005.
- [284] Y. Tsao, S.-M. Lee, and L.-S. Lee. Segmental eigenvoice with delicate eigenspace for improved speaker adaptation. *IEEE Trans. on Speech and Audio Processing*, 13(3):399–411, May 2005.
- [285] A. Tuerk and S. Young. Modeling speaking rate using a between frame distance metric. In *Proc. of Eurospeech*, volume 1, pages 419–422, Budapest, Hungary, September 1999.
- [286] C. Tuerk and T. Robinson. A new frequency shift function for reducing inter-speaker variance. In *Proc. of Eurospeech*, pages 351–354, Berlin, Germany, September 1993.
- [287] G. Tur, D. Hakkani-Tür, and R. E. Schapire. Combining active and semi-supervised learning for spoken language understanding. *Speech Communication*, 45(2):171–186, February 2005.
- [288] V. Tyagi, I. McCowan, H. Bourlard, and H. Misra. Melcepstrum modulation spectrum (MCMS) features for robust ASR. In *Proc. of ASRU*, pages 381–386, St. Thomas, US Virgin Islands, December 2003.
- [289] V. Tyagi and C. Wellekens. Fepstrum representation of speech. In *Proc. of ASRU*, Cancun, Mexico, December 2005.
- [290] V. Tyagi, C. Wellekens, and H. Bourlard. On variable-scale piecewise stationary spectral analysis of speech signals for ASR. In *Proc. of Interspeech*, pages 209–212, Lisbon, Portugal, September 2005.
- [291] U. Uebler. Multilingual speech recognition in seven languages. *Speech Communication*, 35(1-2):53–69, August 2001.
- [292] U. Uebler and M. Boros. Recognition of non-native german speech with multilingual recognizers. In *Proc. of Eurospeech*, volume 2, pages 911–914, Budapest, Hungary, September 1999.
- [293] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson. Scale transform in speech analysis. *IEEE Trans. on Speech and Audio Processing*, 7(1):40–45, January 1999.
- [294] T. Utsuro, T. Harada, H. Nishizaki, and S. Nakagawa. A confidence measure based on agreement among multiple LVCSR models - correlation between pair of acoustic models and confidence. In *Proc. of ICSLP*, pages 701–704, Denver, Colorado, 2002.
- [295] D. VanCompernelle. Recognizing speech of goats, wolves, sheep and ... non-natives. *Speech Communication*, 35(1-2):71–79, aug. 2001.
- [296] D. VanCompernelle, J. Smolders, P. Jaspers, and T. Hellemans. Speaker clustering for dialectic robustness in speaker independent speech recognition. In *Proc. of Eurospeech*, pages 723–726, Genova, Italy, 1991.

- [297] S. V. Vaseghi, N. Harte, and B. Miller. Multi resolution phonetic/segmental features and models for HMM-based speech recognition. In *Proc. of ICASSP*, pages 1263–1266, Munich Germany, 1997.
- [298] A. Venkataraman, A. Stolcke, W. Wangal, D. Vergyri, V. Ramana Rao Gadde, and J. Zheng. An efficient repair procedure for quick transcriptions. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.
- [299] H. Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 25:183–192, April 1977.
- [300] R. Watrous. Speaker normalization and adaptation using second-order connectionist networks. *IEEE Trans. Neural Networks*, 4(1):21–30, January 1993.
- [301] Mitch Weintraub, Kelsey Taussig, Kate Hunicke-Smith, and Amy Snodgrass. Effect of speaking style on LVCSR performance. In *In Proc. Addendum of ICSLP*, Philadelphia, PA, USA, sep 1996.
- [302] L. Welling, H. Ney, and S. Kanthak. Speaker adaptive modeling by vocal tract normalization. *IEEE Trans. on Speech and Audio Processing*, 10(6):415–426, September 2002.
- [303] F. Weng, H. Bratt, L. Neumeyer, and A. Stomcke. A study of multilingual speech recognition. In *Proc. of Eurospeech*, volume 1, pages 359–362, Rhodes, Greece, September 1997.
- [304] T. Wesker, B. Meyer, K. Wagener, J. Anemüller, A. Mertins, and B. Kollmeier. Oldenburg logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. In *Proc. of Interspeech*, pages 1273–1276, Lisboa, Portugal, September 2005.
- [305] M. Westphal. The use of cepstral means in conversational speech recognition. In *Proc. of Eurospeech*, volume 3, pages 1143–1146, Rhodes, Greece, September 1997.
- [306] D.A.G. Williams. *Knowing What You Don't Know: Roles for Confidence Measures in Automatic Speech Recognition*. PhD thesis, University of Sheffield, 1999.
- [307] J. G. Wilpon and C. N. Jacobsen. A study of speech recognition for children and the elderly. In *Proc. of ICASSP*, volume 1, pages 349–352, Atlanta, Georgia, May 1996.
- [308] S. M. Witt and S. J. Young. Off-line acoustic modelling of non-native accents. In *Proc. of Eurospeech*, volume 3, pages 1367–1370, Budapest, Hungary, September 1999.
- [309] S. M. Witt and S. J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2-3):95–108, February 2000.
- [310] P.-F. Wong and M.-H. Siu. Decision tree based tone modeling for chinese speech recognition. In *Proc. of ICASSP*, volume 1, pages 905–908, Montreal, Canada, May 2004.
- [311] B. Wrede, G. A. Fink, and G. Sagerer. An investigation of modeling aspects for rate-dependent speech recognition. In *Proc. of Eurospeech*, Aalborg, Denmark, September 2001.
- [312] X. Wu and Y. Yan. Speaker adaptation using constrained transformation. *IEEE Trans. on Speech and Audio Processing*, 12(2):168–174, March 2004.
- [313] W.-J. Yang, J.-C. Lee, Y.-C. Chang, and H.-C. Wang. Hidden Markov Model for Mandarin lexical tone recognition. In *IEEE Trans. on Acoustics, Speech, and Signal Processing*, volume 36, pages 988–992, July 1988.
- [314] Y. Konig and N. Morgan. GDN: a gender-dependent neural network for continuous speech recognition. In *Proc. of Int. Joint Conf. on Neural Networks*, volume 2, pages 332 – 337, Baltimore, Maryland, June 1992.
- [315] G. Zavaliagos, R. Schwartz, and J. McDonough. Maximum a posteriori adaptation for large scale HMM recognizers. In *Proc. of ICASSP*, pages 725–728, Atlanta, Georgia, May 1996.
- [316] P. Zhan and A. Waibel. 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, May 1997.
- [317] P. Zhan and M. Westphal. Speaker normalization based on frequency warping. In *Proc. of ICASSP*, volume 2, pages 1039–1042, Munich, Germany, April 1997.
- [318] B. Zhang and S. Matsoukas. Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition. In *Proc. of ICASSP*, volume 1, pages 925–928, Philadelphia, PA, March 2005.
- [319] Y. Zhang, C.J.S. Desilva, A. Togneri, M. Alder, and Y. Attikiouzel. Speaker-independent isolated word recognition using multiple Hidden Markov Models. In *Proc. IEE Vision, Image and Signal Processing*, volume 141, 3, pages 197–202, June 1994.
- [320] J. Zheng, H. Franco, and A. Stolcke. Rate of speech modeling for large vocabulary conversational speech recognition. In *Proc. of ISCA tutorial and research workshop on automatic speech recognition: challenges for the new Millenium*, pages 145–149, Paris, France, September 2000.
- [321] J. Zheng, H. Franco, and A. Stolcke. Effective acoustic modeling for rate-of-speech variation in large vocabulary conversational speech recognition. In *Proc. of ICSLP*, pages 401–404, Jeju Island, Korea, October 2004.
- [322] B. Zhou and J.H.L. Hansen. Rapid discriminative acoustic model based on eigenspace mapping for fast speaker adaptation. *IEEE Trans. on Speech and Audio Processing*, 13(4):554–564, July 2005.
- [323] G. Zhou, M.E. Deisher, and S. Sharma. Causal analysis of speech recognition failure in adverse environments. In *Proc. of ICASSP*, volume 4, pages 3816–1819, Orlando, Florida, May 2002.

- [324] D. Zhu and K. K. Paliwal. Product of power spectrum and group delay function for speech recognition. In *Proc. of ICASSP*, pages 125–128, 2004.
- [325] Q. Zhu and A. Alwan. AM-demodulation of speech spectra and its application to noise robust speech recognition. In *Proc. of ICSLP*, volume 1, pages 341–344, Beijing, China, October 2000.
- [326] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke. On using MLP features in LVCSR. In *Proc. of ICSLP*, Jeju Island, Korea, October 2004.
- [327] A. Zolnay, R. Schlüter, and H. Ney. Robust speech recognition using a voiced-unvoiced feature. In *Proc. of ICSLP*, volume 2, pages 1065–1068, Denver, CO, September 2002.
- [328] A. Zolnay, R. Schlüter, and H. Ney. Acoustic feature combination for robust speech recognition. In *Proc. of ICASSP*, volume 1, pages 457–460, Philadelphia, PA, March 2005.