AUTOMATIC SPEECH RECOGNITION AND INTRINSIC SPEECH VARIATION

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

Multitel, Eurecom, Universität Oldenburg, France Telecom, Acapela Group,
Loquendo, Politecnito Torino, Université d'Avignon, McGill University

ABSTRACT

This paper briefly reviews state of the art related to the topic of speech variability sources in automatic speech recognition systems. It focuses on some variations within the speech signal that make the ASR task difficult. The variations detailed in the paper are intrinsic to the speech and affect the different levels of the ASR processing chain. For different sources of speech variation, the paper summarizes the current knowledge and highlights specific feature extraction or modeling weaknesses and current trends.

1. INTRODUCTION

Major progress is being recorded regularly on both the technology and exploitation of Automatic Speech Recognition (ASR) and spoken language systems. However, there are several 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 speech intrinsic variations naturally present in speech. For instance, specific aspects of the speech signal like foreign accents, precludes the use by specific populations. Also, some applications, like directory assistance, 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 itself. These create a wide range of variations: speaker, gender, speech rate, vocal effort, regional accents, speaking style, non stationarity...

2. VARIATION IN SPEECH

2.1. Speaker characteristics

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

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

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

The preponderance of lower frequencies for carrying the linguistic information has been assessed by both perceptual and acoustical analysis and justify the success of the non-linear frequency scales such as Mel, Bark, Erb. Other approaches aim at building acoustic features invariant to the frequency warping [8, 9]. A direct application of the tube resonator model of the vocal tract lead to the different vocal tract length normalization (VTLN) techniques: speaker-dependent formant mapping [10], transformation of the LPC pole modeling [11], frequency warping, either linear [12] or non-linear [13], all consists in modifying the position of the formants in order to get closer to an "average" canonical speaker. Incidentally, channel compensation techniques such as the cepstral mean subtraction or the RASTA filtering of spectral trajectories, also compensate for the speaker-dependent component of the long-term spectrum [14, 15].

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

2.2. Foreign and regional accents

As introduced earlier, accent is one of the major components of inter-speaker variability, as demonstrated in [2]. And indeed, compared to native speech recognition, performances degrades when recognizing accented speech and even more for non-native speech recognition [18]. In fact accented speech is associated to a shift within the feature space [19]. For native accents the shift is applied by large groups of speakers, is more or less important, more or less global, but overall

1-4244-0469-X/06/$20.00 ©2006 IEEE

acoustic confusability is not changed significantly. On the opposite, for foreign accents, the shift is very variable, is influenced by the native language, and depends also on the level of proficiency of the speaker.

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

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

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

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

2.3. Speaking rate and style

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

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

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

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

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

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

2.4. Age

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

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

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

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

dictionary, and the use of language models which are customized for children speech have all been tried [47].

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

2.5. Emotions

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

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

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

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

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

2.6. Conclusion[2]

In general, this review further motivates research on the acoustic, phonetic and pronuncition limitations of ASR, and also pinpoints specific representation weaknesses. It is for instance aknowledged that pronunciation discrepancies is a major factor of reduced performance (in the case of accented and spontaneous speech)

Breakthroughs and improvements in general feature extraction and modeling techniques that provide more resistance to speech and production diversity are hence desirable. General techniques such as compensation, adaptation, multiple models, additional acoustic cues and more accurate/detailed models are popular to adress the problems posed by speech variation. Also, driven by the commoditization 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. [60].

Such complementary and multidisciplinary perspectives are currently receiving more interest and, coupled with up-to-date knowledge from phonetics as well as auditory and cognitive sciences, are probably keys to significantly push the technology forward.

[2] The topics adressed here are only briefly touched upon. A more comprehensive overview will however be provided at the scientific workshop organized by the DIVINES project *httm://www.divines-project.org* in May 2006.

3. REFERENCES

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

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

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

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

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

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

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

[8] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 40–45, Jan. 1999.

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

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

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

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

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

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

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

[16] C. Lee, C. Lin, and B. Juang., "A study on speaker adaptation of the parameters of continuous density hidden markov models," *IEEE Trans. Signal Processing.*, vol. 39, no. 4, pp. 806–813, April 1991.

[17] C. Leggetter and P. Woodland., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer, Speech and Language*, vol. 9, no. 2, pp. 171–185, April 1995.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[37] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, "Effects of disfluencies, predictability, and utterance position on word form variation in english conversation," *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1001–1024, Feb. 2003.

[38] Jean-Claude Junqua, "The lombard reflex and its role on human listeners and automatic speech recognisers," *JASA*, vol. 93, no. 1, pp. 510–524, Jan. 1993.

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

[40] M. Adda-Decker, P. Boula de Mareuil, G. Adda, and L. Lamel, "Investigating syllabic structures and their variation in spontaneous french," *Speech Communication*, vol. 46, no. 2, pp. 119–139, June 2005.

[41] 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. Speech Audio Process.*, vol. 12, no. 4, pp. 349–350, July 2004.

[42] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajic, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, T. Ward, and Z. Wei-Jin, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 420–435, July 2004.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[60] P. Nguyen, L. Rigazio, and J.-C. Junqua, "Large corpus experiments for broadcast news recognition," in *Proc. of Eurospeech*, 2003.