ISCA Technical and Research Workshop: *Adaptation Methods in Automatic Speech Recognition* Sophia Antipolis, France, 29-30 August 2001

Speech recognition scores grew dramatically in the last 8 years for applications in quiet environments with dedicated speakers reading texts. The presence of noise, as well as the spontaneous speaking mode, still constitute major difficulties in using vocal interfaces for commercial applications. Also, the variability of the transmission channel (telephone channel, wireless channel or different microphones) is a major obstacle to robust recognition. A solution would be to retrain speech unit models (phonemes, triphones, syllables) for each new operating condition; however, this is completely unrealistic since users need immediate access to the voice interaction. Today, adaptation is the most promising solution. Adaptation can be defined as a method to modify the general model parameters trained previously under favorable conditions and on a large database by using the few data collected on the spot when the user wants to place a vocal command or request.

For that reason, Professor Chris J. Wellekens, Institut Eurécom, Sophia Antipolis, and Dr. Jean Claude Junqua, President of Panasonic Speech Technology Laboratories, Santa Barbara, California, organized a 2-day workshop on August 29-30, 2001, devoted to adaptation. The workshop had the support of ISCA (International Speech Communication Association) at Sophia Antipolis. Seventy-five researchers from around the world attended the workshop, that was also a satellite event of Eurospeech 2001 in Aalborg (Denmark).

Each half day was devoted to a wide domain in adaptation. An expert introduced the topic by presenting the state of the art. Then participants presented their contribution in this wide domain via poster sessions, concluding with a panel discussion prepared by moderators that gave participants the opportunity to have fruitful scientific exchanges. The different domains were:

- Speaker Adaptation. Invited lecturer: Phil Woodland; Panel moderator: Hervé Bourlard
- Acoustic Model Adaptation. Invited lecturer: Shigeki Sagayama; Panel moderator: Mazin Rahim
- Pronunciation Adaptation. Invited lecturer: Helmer Strik (presentation by Lou Boves); Panel moderator: William Byrne
- Language Model Adaptation. Invited lecturer: Jerôme Bellegarda; Panel moderator: Renato de Mori.

In his review paper on speaker adaptation, Phil Woodland pointed out the different adaptation modes. If the actual text is known a priori, then the parameters of the model can be modified in a supervised way, but unsupervised techniques may rely on confidence measures on recognized text.

Three different families of techniques are observed today:

- the Maximum a Posteriori techniques (MAP)
- the Maximum Likelihood Linear Regression (MLRR)
- the techniques using speaker clustering or speaker-space methods (eigenvoices).

The MAP criterion provides more robust estimates with less data but the adaptation focuses on the data seen during the adaptation phase and can even degrade unobserved models.

This can be partially circumvented by using the correlation between the parameters and updating even unobserved models. In another MAP technique, called structural MAD, Gaussians of the models are organized in a tree structure and a mean vector offset and diagonal covariance scaling are updated recursively descending from the root of the tree to the leaves.

The MLLR techniques are quite popular. The parameters of an affine transformation of mean vectors are estimated using the ML criterion. MLLR can also be applied on covariance matrices. In constrained MLLR, the use of similar parameters for mean vector transformation as for covariance transformation can be seen as a transformation of the data. Instead of using speaker independent seed models, it is possible to use speaker dependent models available from the general training data. This technique is known as speaker adaptive training.

The last family uses clustering. Data are divided in speaker groups (the most trivial is by gender) and models are trained for each group. Then the choice of a group for a current speaker is made by a hard decision. A better approach is the cluster adaptive training (CAT) where the mean vectors for the current speaker is estimated as a combination of the mean vectors of the groups. A similar idea is used in the eigenvoice techniques, where a high dimensional space describes the speakers observed in the training data. Then, using principal component analysis, the size of this space is reduced and speakers are projected in this reduced space.

Extensions of these techniques are discussed, as is vocal tract length normalization (VTLN), which has drawn a new interest in the last years.

Shigeki Sagayama developed analytic methods for acoustic model adaptation to speaker change and also to noise and channel variations. He describes the Jacobian adaptation, which is a way to use sensitivity networks related to models. The method can be applied to noisy cepstra, to time derivatives of cepstral features, to mean vectors, and to covariance matrices.

A generalization of the Jacobian adaptation is Vector Field Smoothing (VFS).

To conclude the paper, Shigeki Sagayama compared these techniques to MAP, MLLR, structural approaches, clustering techniques (eigenvoices) and feature compensation (vocal tract length normalization and cepstrum mean normalization).

Helmer Strik's paper was presented by Lou Boves. It was devoted to the very important problem of pronunciation adaptation. Speakers do not pronounce the words exactly as they are phonetically transcribed (canonical transcription) in a dictionary. Using the canonical transcriptions in supervised training leads to erroneous speech unit models. Also, building word models from canonical transcriptions will not lead to robust recognizers. Using human expertise for labeling actual pronunciation is quite expensive (in time and in money) and not consistent. Application of rules of phonetics may help, but is far from covering all observable variabilities (i.e., knowledge based methods). Efforts were devoted to the generation of pronunciation variants from the training data themselves (data driven). Two steps are necessary. First, the pronunciation variants should be found, and second, a technique should discovered on how to use them in speech recognition.

The most popular technique used for generating information about pronunciation variability is forced alignment (dynamic programming), the results of which can be used to derive rewrite rules, train an artificial neural network or decision trees, and calculate confusion matrices. The variants are usually generated using one of these techniques.

At a first glance, increasing the number of transcriptions of a word must increase the chance to recognize this word. But a large number of transcriptions also increases the confusability between the words. As a consequence, the variants per word should be carefully selected and their number limited. The criteria used are the frequency of occurrences, an ML criterion, confidence measures and the degree of confusability.

The last invited review paper, presented by Jerôme Bellegarda, was on the topic of Statistical Language Model Adaptation. Syntactic and semantic constraints play a pivotal role in speech recognition.

Also, a same message can be formulated along different sentences and these various sentences can be seen as stochastic realizations of the message.

Natural language constraints can be approximated by context free grammars, but their implementation in an ASR is difficult. As a consequence, stochastic finite state automata (regular grammars) are regularly used. However, by adding the Markovian assumption, the use of statistical n-grams is frequently preferred. The n-grams are trained. Much better results are obtained with training material selected in a given task than with more training material but from different tasks.

Adaptation of a statistical language model consists in the tuning of the n-gram probabilities of a background task to a specific one. Different approaches are discussed.

In model interpolation, a linear combination between background probabilities and probabilities poorly trained on the current task is used as a new statistical model and the interpolation weights are estimated (e.g. by using the EM algorithm or a MAP adaptation). In a second approach, called constraint specification, the adaptation corpus is used to extract features that the statistical language model is constrained to satisfy. The criterion used is the minimum discrimination information (MDI). A third family of techniques is the exploitation of the general topic of the discourse. Here the information extracted from the adaptation language model is used to improve the background model based on semantic classification. More generally the semantic knowledge instead of the topic information can be used. All approaches using syntactic knowledge assume that the background and the adaptation tasks share a common grammatical infrastructure. The background model can then be used for initial syntactic modeling while the adaptation corpus is used for the re-estimation of the associated parameters.

Forty contributions on these domains were presented as posters and are published in the Proceedings. The panel sessions were also quite active and very well prepared by the moderators.

On the first evening, a sunset trip along the Esterel Golden Coast (red rocks) took the participants to the Car Museum of Mougins, where they were welcomed by traditional Provencal dancers and had dinner.

The workshop was sponsored by the Conseil Général des Alpes Maritimes, Swisscom and Texas Instruments. Proceedings, as well as a CD-ROM containing also the material used by the authors for their presentations, can be bought at the ISCA secretariat (info@isca-speech.org).

Chris J.Wellekens
Dept. Multimedia Communications
Institut Eurécom
Sophia Antipolis, France
e-mail: wellekens@eurecom.fr