

Acoustic Model Adaptation in Loquendo ASR

The use of general acoustic models in the Loquendo ASR Automatic Speech Recognition engine has provided excellent performance in a vast array of applicative conditions. However, it is important to be able to adapt the acoustic models in such a way as to take into account acoustic material acquired in the field. For instance, adapting recognition to a specific individual's voice or to a set of individuals, or to the telephone environment at large.

Alongside the new version of Loquendo ASR, Loquendo recently released an innovative adaptation technique, called the **Loquendo Acoustic Model Adaptation tool**, which is based on a *patented technology* and is now another important element of the Loquendo ASR Tool-Suite. This is another step towards facilitating Loquendo partners and clients to best exploit our technology.

This paper will introduce the main concepts associated with acoustic model adaptation and will provide a description of the acoustic model adaptation process, with practical advice for an optimal use of this novel technique.

Acoustic Model Adaptation

The acoustic models released in a commercial Automatic Speech Recognition engine (ASR), such as the Loquendo ASR engine, are general models trained on large telephonic corpora, which include many speakers with a statistical distribution of age, sex, and geographic areas. They represent *an average situation* of spoken telephone speech for each language and guarantee excellent performance in the most common situations.

However, in all cases where a vocal application has certain peculiarities that differ from the standard speaker independent telephone environment, acoustic model adaptation can be used to further increase recognition performance.

For example, an application that is always used by the same speaker can be improved if the acoustic models are adapted to that speaker. In the same way, an application that is used in a particular environment (e.g. car) or using a certain kind of input device (e.g. a particular PDA or smart-phone) can be improved if the general models are adapted using vocal material recorded from user interaction.

The **most common adaptation cases** are the following:

- **Speaker**;
- **Audio Channel** and input media (e.g. specific microphone, PDA);
- **Environment** where the vocal application is used (e.g. car, office, factory);
- **Way of speaking** (e.g. regional accents, fast speaking);
- Application dependent **vocabulary** (e.g. some specific jargon, like aeronautic terms);

In all these cases, adaptation can be performed using some audio material recorded from the target application. The adapted models can greatly improve the recognition performances. The different adaptation cases have certain specific characteristics that need to be considered:

Speaker: the improvement of recognition can be great, with only a small amount of vocal material needed (a few minutes of voice recording can be sufficient). However, the recorded vocal material needs to have a broad coverage of the phonemes of the language in question, in order to guarantee a speaker adaptation that is valid for a wide range of applications.

Audio Channel: if an audio channel different from telephone is used, the adaptation of the general acoustic model to the target audio channel can produce great improvements. Please note that if you wish to keep speaker and vocabulary independence, the vocal material needs to be broad enough and should be recorded from a wide variety of speakers, pronouncing several different words.

Environment: if the vocal application is always used in a specific environment (e.g. car, factory, airplane cockpit) adapting the general acoustic model to the operating environment can be very useful. Nevertheless, adaptation cannot solve all recognition problems. If the environment is very noisy,

performances cannot equal those in quiet conditions, and adaptation can alleviate the issue but not eliminate it altogether.

Way of speaking: if your application includes some typical way of speaking, which differs from the language's average way of speaking, recognition performance can be reduced. Some of the ways of speaking that can be dealt with, to some extent, with acoustic model adaptation are: non-native accents, regional accents, childrens' or elderly persons' speech, speaking rate.

Vocabulary: a vocal application may often be based on professional jargon or a specific lexicon. For example, the jargon used by aircraft pilots, maintenance men, stockbrokers, etc. In these cases, adapting the general acoustic model to the specific words to be recognized (e.g. the name of stocks or of aircraft commands) can be useful. However, this adaptation case is somehow less important than the previous ones, as the general model usually works well with any vocabulary, and the improvement margins are limited. On the contrary, adaptation with insufficient data could even produce a degradation of the speaker independent recognition performances, as it may introduce undesired dependence on the speakers that have contributed to the adaptation material.

Acoustic Adaptation in Loquendo ASR

A module for **Acoustic Models Adaptation (AMA)** for *Loquendo ASR* has been developed and released with Loquendo ASR 6.7.0. Its features are innovative, and allow the acoustic models to be adapted to Speaker, Audio Channel, Environmental and Applicative conditions with a limited amount of vocal data. In fact, collecting massive amounts of speech recordings from specific speakers or in specific applicative conditions, in order to train the acoustic models from scratch is difficult and impractical. Instead, Loquendo AMA allows the customisation of the standard acoustic models with an amount of data that is easy to collect, boosting recognition performance. Loquendo AMA incorporates a state of the art technique for acoustic model adaptation together with an innovative patented technique, leading to improved adaptation capabilities.

Differently from many commercial ASR engines that are based on Hidden Markov Models (HMM), *Loquendo ASR integrates HMM and Neural Networks* with a hybrid approach (HMM-NN) [2][3]. Hybrid HMM-NN models [1] integrate the ability of dealing with temporal patterns, typical of HMM, with the superior pattern classification power provided by NN.

The main advantages of HMM-NN are better recognition performances, especially in the case of phonetic transcription, and superior efficiency, especially in the case of large vocabularies.

Differently from the case of standard HMM, where many adaptation algorithms are available (MAP, MLLR), there is limited available literature on proposals for the adaptation of HMM-NN. A standard technique is named Linear Input Network (LIN) and was published in the nineties [4]. A LIN is a single layer network that performs a linear mapping of the space of input parameters. This network is placed before the HMM-NN, already trained in a speaker independent manner, and is trained with the adaptation material.

LIN was the first adaptation technique to be implemented in Loquendo AMA. Following extensive experimentation, this technique was judged to be valid but not completely satisfactory. Thus, Loquendo developed a brand new technique named LHN, which was patented in June 2005 and will soon be presented, at forthcoming international conferences. This new technique has been tested on standard ARPA tests for Speaker Adaptation (WSJ0 and WSJ1 Spoke3) and has shown a large improvement with respect to LIN.

This new adaptation technique is currently the default approach in **Loquendo AMA**, however LIN can be used on request.

Adaptation in practice

The general scheme

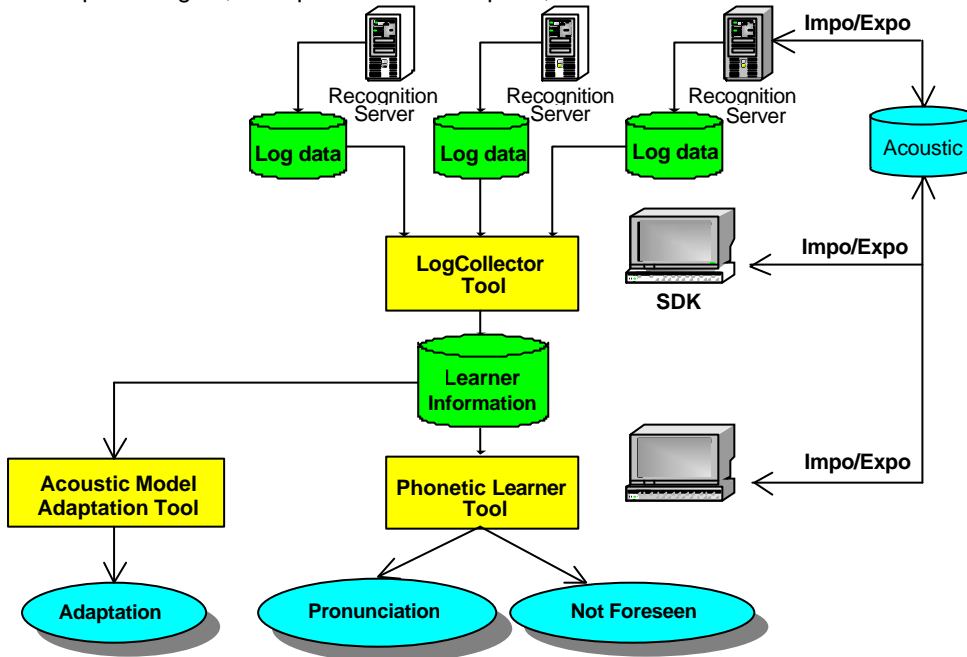
Loquendo AMA is a new tool that goes alongside the previous ones within a framework where the recognizer's performance can be improved using data collected from the on-field application.

The general scheme for Loquendo learning and adaptation from on field data is outlined in Fig. 1. The diagram also features the Phonetic Learning tool and LogCollector tool that were the object of discussion in a previous issue of the Loquendo Newsletter. The Loquendo AMA uses the LogCollector tool to gather the adaptation data and can be used in conjunction with the Phonetic Learning tool to improve the performance of the general acoustic model.

Fig. 1. The acoustic model adaptation (and phonetic learning) architecture

As depicted in the Fig.1, the global scheme is based on a three-step process: (1) data acquisition, (2) data analysis, and (3) adapted model activation.

In the data acquisition step the user must acquire new data from the recognizers running on some recognition servers. With Loquendo ASR, it is possible to do this by activating the dumping of the Log data. In this way, during the running of the application the information necessary for adaptation is saved, e.g. the speech signal, their phonetic transcription, etc.



Subsequently, data analysis must be carried out using Loquendo ASR installed on an off-line workstation. The LogCollector tool allows selecting and gathering the Log data and makes them available to Loquendo AMA. Using the Loquendo AMA, one can produce the *adapted acoustic model*.

The adaptation takes place in an unsupervised mode, where no human activity is required in order to transcribe the collected material. The collected material needed for the adaptation requires some further comment:

1. **The quantity of vocal material needed** may vary from a few minutes of voice (for speaker adaptation) to a few hours of voice (for a typical adaptation to a different environment, or to a specific application). For a complex speaker-independent application, several hours of voice material may be necessary in order to obtain some improvement through adaptation. In general, the larger the amount of material, the better the results will be. It must be noted however, that adaptation time will increase accordingly.
2. **As regards the kind of vocal material**, users are asked to bear in mind that Loquendo AMA adapts to everything that is present in the adaptation material and only to what is present. Thus, if you want to preserve speaker independence, collect material from many different speakers, while if you want to preserve open vocabulary performances, collect material with a wide phonetical variety. If you collect material containing only a limited set of words (e.g. only digits) the adapted model will perform better on that lexicon and maybe a little worse on other different lexica. However, Loquendo AMA incorporates an innovative, patented technology that avoids damaging the general performances of the adapted models even when the adaptation material is too limited and polarized in terms of lexicon.

The final adapted model activation can be carried out by exporting the adapted model from the off-line workstation and importing it into the recognition servers and configuring/re-running the application in order to use the adapted model instead of the general one. The final model shows increased recognition performance, without any additional computational charge.

The adaptation process

Loquendo AMA is capable of adapting the general acoustic model, released with the Loquendo ASR SDK, with on-field collected audio material. In this first release, the tool is a console application based software

developed for Microsoft Windows 2000/XP and Linux operating systems. In the future release, it will be provided a graphic interface as for the other tools involved in the general scheme of Fig. 1.

The adaptation process is realized in 3 steps:

- **Creation of the adaptation session:** The adaptation session must be created as the very first step in order to create the root point in the local file system where the working data and the tracing of the adaptation activity are stored. At the end of the adaptation process the session can be removed from the file system.
- **Selection of audio material:** During this phase, the tool selects the most suitable recordings for the next adaptation step from the audio material stored using the Loquendo "LogCollector tool".
- **Model adaptation:** the selected recording are used to modify the general acoustic model and to make it more suitable to the channel, environment and speakers characteristics

Following the adaptation phase, the application developer should validate the proposed adapted model using the Loquendo ASR Evaluation Tool Kit. This tool is an application-based software console developed for Microsoft Windows 2000/XP and Linux operating systems.

Adaptation Tips

Loquendo AMA allows the application developer to improve the performance of acoustic models. It is a powerful tuning tool but it must be used in a consistent way. Here are some best practice tips on using Loquendo AMA:

- a) Loquendo AMA adapts to everything that is present in the adaptation material, so avoid associating the material to unwanted aspects. E.g. if you wish to adapt to a specific microphone, but you record all the adaptation material with male speakers only, you will adapt to the specific microphone (desired) but also to male speakers (undesired!).
- b) While for speaker adaptation some minutes (5-10) of voice can be sufficient, much more data, from several speakers evenly representing the population that will use the application, is necessary if you want to adapt while maintaining speaker independence.
- c) Loquendo AMA is a time consuming off-line process. The time requested may range from few minutes to several hours depending on the amount of adaptation material and the computational power of the computer used.
- d) In its standard way of operating, Loquendo AMA is able to improve recognition performance in the case where the performance of the general models are not too low (at least 70% WA). Otherwise, Loquendo AMA may not be able to improve performance, and it is possible that they may decrease after adaptation.
- e) It is always advisable to use the Loquendo ASR evaluation toolkit to test the adapted models before rolling them out into the application. In fact, Loquendo AMA is an unsupervised process that uses the ASR to guess the correct word transcription. If the starting point given by the general models is insufficient, Loquendo AMA may, in some cases, not bring good results.
- f) In difficult cases, when the performance of the general models is not high enough (less than 70% WA), you may wish to try using the "supervised Loquendo AMA", an auxiliary tool that is not released in the standard distribution, but provided only on request. For using "supervised Loquendo AMA" some vocal material has to be collected and transcribed manually by a human operator. Please contact Loquendo , for more details.