Estimation and Uncertainty Processing Techniques for Signal Transmission and Recognition

Antonio M. Peinado, José L. Pérez-Córdoba, Victoria E. Sánchez,
Ángel M. Gómez, José L. Carmona, José A. González López Juan A. Morales Cordovilla

Departamento de Teoría de la Señal, Telemática y Comunicaciones
Universidad de Granada, Granada (Spain)
amp@ugr.es

Abstract

This paper presents the research project TEC2010-18009/TCM proposed for its funding by the Ministerio de Ciencia e Innovacin (MICINN). The main goal of this research project is the development of two groups of techniques for the processing of noisy, damaged or lost information: estimation and uncertainty processing. We will consider two different applications with a clear parallelism: speech recognition in adverse environments and error concealment for robust multimedia transmission (speech and video).

1. Introduction

During the last years, our research team has been developing its work on remote speech recognition (RSR). An RSR system employs client/server architecture for transmitting speech signals or parameters from a thin client (typically mobile devices such as smart-phones, VoIP phones or PDAs) to a powerful remote server where recognition is performed. Through a series of R+D projects, we have considered several issues related to RSR. Thus, we have dealt with robustness against adverse acoustic environments and against degraded transmission channels (mobile networks, IP, WAN,). In order to tackle with this last problem, our research team has employed, among others, two approaches especially attractive due to their statistical nature. First, we have considered several estimation techniques (mainly Bayesian) and, specially, the minimum square error (MMSE) criterion. Also, it must be taken into account that the transmitted speech features will not be, even after estimation at the decoder, reliable. The uncertainty processing techniques try to obtain some kind of reliability measures which can be incorporated to the speech recognizer in order to improve the system performance. In this proposal, the research team wants to exploit this previous experience and the results obtained in RSR about estimation/uncertainty techniques with the aim of deepening in them with a more general perspective, extending them to new problems and applications. In order to do that, we must undertake two general issues:

1. Estimation techniques. The most important point for their application is to obtain an statistical model (suitable for every application and signal) for the information (signal/parameters) generation and transmission process which can be integrated in a Bayesian estimation framework. Also, it is important to do a suitable use of the available data (distorted by noise or errors).

2. Uncertainty processing techniques. Basically, we consider here the noise-robust speech recognition problem, since state-of-the-art speech recognizers usually adopt a statistical approach, which allows a natural management of uncertainty. Two issues must be considered here: which reliability measures are to be used and how the speech recognizer is modified to incorporate these measures.

Estimation and uncertainty are related issues since, as mentioned, the estimation process involves uncertainty. Thus, although the reliability measures may be obtained from different criteria, it is also possible extracting them from the probability distributions employed for estimation.

2. Goals, Concepts and State of the Art

Taking into account the discussion in the previous Section, the goal of this proposal can be summarized as follows (see Fig. 1).

• The research team seeks a continuation of its traditional research, automatic speech recognition, as main application. This way, this project continues previous projects developed by the team on RSR, although focusing now on the acoustic noise problem. This is considered the main issue in order to obtain a ubiquitous and pervasive human-machine interaction. In particular, we intend to develop both estimation and uncertainty processing techniques under a joint conception and with the aim of allowing the collaboration between them. As we will see later in this section, missing data techniques provide a suitable framework for this goal.

• We also intend to extend estimation techniques to other applications and, in particular, to multimedia signal transmission. Thus, our team seeks the translation of its previous experience to a new field whith important applications such as VoIP telephony, audio/image/video streaming or DVB-H television. As mentioned, our group has already work in this field and, in particular, on robust speech transmission for RSR. This way, we will consider first the general problem of speech transmission but, also, we intend to exploit synergies and explore new

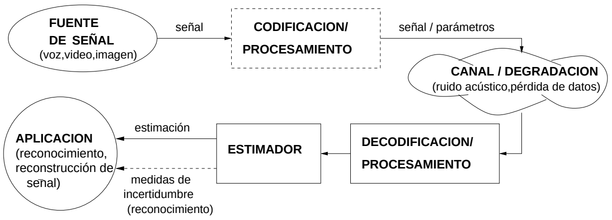

Figure 1: *Estimation and uncertainty processing for different applications.*

This work has been supported the project TEC2007-66600.

applications such as images and, mainly, video.

• Finally, another objective will be the development of techniques which may complement estimation/uncertainty. For example, in transmission applications, source and channel coding techniques not only can be combined with estimation, but also they may be required by the estimator in order to provide useful a posteriori information. Regarding speech recognition, we must also consider complementary techniques (preprocessing, noise detection and modeling) which can support and collaborate with estimation/uncertainty.

The following subsections are devoted to the development of the concepts mentioned above, as well as a review for every considered application.

2.1. Robust Speech Recognition

During the last years, automatic speech recognition (ASR) has shown its utility in different applications, although it is specially promising for human-computer interaction in portable and mobile devices (mobile phones, PDAs, e-books). This fact is reflected in the strategic research agendas of several technology platforms such as eMobility or eMOV. Mobility requires ASR system which must work in noisy environments. Then, robustness against acoustic noise is a crucial issue which is researched by many gropus devoted to ASR. This proposal intends to research this topic employing estimation and uncertainty techniques. In the field of robust speech recognition, the estimation techniques try to recover the speech features when acoustic noise contaminates the speech signal. MMSE estimation has been very successful when combined with stereo data for training, that is, when the same speech signals are available with and without noise, since this allows a suitable modeling of noise. One of the first approaches where SNR-dependent cepstral normalization (SDCN) and codeword-dependent cepstral normalization (CDCN). Since then, other related techniques as multivariate Gaussian cepstral normalization (RATZ), stereo-data linear compensation of environments (SLICE), multienvironment model-based linear normalization (MEMLIN), uncertainty decoding or stereo-data stochastic mapping (SSM) have been developed. All these techniques share the idea of assuming a parametric data pdf (normally Gaussian mixture models, GMM). In a recent work, the research team proposes a first approach to robust ASR based on MMSE and VQ modeling (instead of GMMs) and cepstral normalization. Another purpose of this project is the application of uncertainty processing. As mentioned, the statistical nature of speech recognizers, usually based on HMMs, facilitates the integration of this kind of processing. Uncertainty processing techniques are based on the fact that not all the speech features are equally reliable due to random nature of noise. Therefore, if we can measure their level of reliability, we can also modify the recognition engine to deal with this fact. This way, it is possible to propose an heuristic exponential feature weighting of the HMM observation probabilities according to the input feture reliability. This approach has found also application in RSR systems for channel error mitigation. An alternative is that of considering that every input speech feature is a random variable instead of a deterministic value. Thus, the probability distributions at the recognizer can be modified according to this uncertainty point of view. We must also mention a group of estimation/uncertainty techniques usually gathered under the name of missing data techniques (MD). The stating point of these techniques consists of identifying, usually in the spectrogram, which regions are reliable (speech prevails over noise) and not reliable (noise prevails over speech). As a result, we obtain masks which can be binary (0=not-reliable, 1=reliable) or continuous, where every mask value provides the reliability level of that part of the spectrogram. Main procedures to compute these masks are those based on instantaneous SNR, those based on speech harmonic properties, or those based on intensity level difference in stereo recording. More recently, speech fragment decoding techniques, based on auditory scene analysis have been proposed. Once reliable and not reliable regions are obtained, there are two possibilities for ASR with MD: imputation and marginalization. Imputation is the same as estimation when reliability masks are employed to reconstruct non reliable regions from reliable ones and a priori speech models. These models can be the same as those employed for recognition or simplified models employed only doe feature extraction. Once the clean speech features are estimated, speech recognition is performed as usual. On the other hand, the marginalization techniques are a type of uncertainty processing techniques which manage uncertain data modifying the recognition engine. This way, reliable parts are employed without modifying the recognizer, while the non reliable parts are marginalized by the acoustic models up to the received energy level (under the assumption of speech plus noise energy additivity). Imputation and marginalization have several advantages and drawbacks. Thus, marginalization techniques provide an optimal classification under the assumption of missing data. These techniques can be only applied to speech models trained with spectral features (Mel-filterbank,Gammatone). However, the performance obtained with these features is lower than that obtained with cepstrum. This fact justified the success of imputation versus marginalization, since imputation still allows the use of cepstral features. In spite of this clear advantage of the imputation techniques, they have the problem of a higher sensitivity to mask estimation errors. The study and development of new imputation/marginalization hybrid schemes is one of the objectives of this project. These hybrid techniques would have the advantage of allowing cepstral features along with uncertainty processing at the recognizer.

2.2. Robust Multimedia Transmission Systems

In multimedia transmission systems, robust techniques against channel errors (bit change, packet loss) are usually classified into three types: interactive error control, techniques based on channel coding and error concealment techniques. In this project, we pay particular attention to the latter type, error concealment (EC), which tries to alleviate the possible changes of the transmitted information caused by the transmission channel. The EC techniques employ different approaches, although all of them share the principle of exploiting the multimedia signal redundancies in order to reconstruct damaged or lost data. Repetition, interpolation/extrapolation and estimation are examples of the most important EC techniques. Estimation, similar to interpolation, can employ data correctly received before and after the error interval, although estimation uses explicitly a statistic model of the source and channel in order to reconstruct, which provides a clear advantage. In this proposal, we focus our research on EC by estimation. Also, we consider channel coding techniques that can combine with EC techniques, such as FEC (forward error correction) codes and interleaving, in which the research team has experience. In the following, we review the state of art in EC techniques applied to different multimedia signals, although we also include those channel coding techniques that could cooperate with the first ones.

2.2.1. Speech/Audio Error Concealment

Traditionally, EC techniques applied to coded speech have been oriented to avoid annoying sounds in the synthesized signal. Thus, most of these techniques are based on the substitution of the lost frames and progressive muting. This trend has been also implemented in popular speech coding standards (EFR, AMR, IS-641, G.723.1, G.729). During last years, VoIP telephony has given rise to new techniques designed for increasing the transmission robustness. One of the main problems to face has been the predictive behavior of the existing speech codecs. This implies that the errors can be propagated forward during a period considerably longer than the duration of the errors themselves. Thus, new coding schemes have been proposed in order to limit the error propagation by removing, totally or partially, the prediction. In special, we must remark the success of the iLBC codec that removes the possibility of error propagation by means of an intra-frame coding, although with a considerable increase of bit rate. In other cases, the propagation is limited by means of periodic intra-frames that act as firewalls, emulating the MPEG standard of video. Nevertheless, the most extended codecs are predictive and they are based on the CELP (coded excitation linerar prediction) paradigm. Thus, during the last years, some EC techniques have been proposed for this type of codecs. A first approach is based on the use of frames received longer than the delay imposed by the anti-jittering butter. Although these frames are not usable directly, they can be used in order to resynchronize the excitation and, therefore, remove the error propagation. In some works the backward excitation energy is bounded since it is the main reason of the error propagation. This loss of energy is compensated by means of the innovation codebook (algebraic codebook) or a glottal-pulse codebook is provided in order to encode the first subframe after a loss for voiced sounds. Also, some works employ the Bayesian estimation as EC technique. As aforementioned, the Bayesian philosophy is based on combining the evidence contained in the signal with the a priori knowledge of the probability distribution of the source process. This is achieved by means of the conditional probability of the possible transmitted information given the received one. In our case, the a priori knowledge is will be given by the present redundancies in the signal. Several EC techniques based on Bayesian methodology can be found in the literature. The MMSE estimator (based on obtaining the expected value of the damaged or lost data given the available information) has been extensively used in speech transmission. In particular, the MMSE formulation using hidden Markov models (HMM) as source model has provided excellent results in speech transmission and parameter transmission for remote speech recognition. Other possible Bayesian estimator is the one based on the maximum a posteriori (MAP) that, unlike MMSE, uses the mode of the a posteriori distribution as estimate. Depending on the type of probability distribution, the MAP estimator can be advantageous. In audio signals coded using MDCT is important to stress the use of the maximum likelihood (ML) for the reconstruction of the lost spectral information. One of the most important aspects in our proposal is the combination of FEC codec with EC techniques. This topic has been slightly discussed in audio/speech, possibly, due to the hard delay restrictions imposed by the telephony applications (FEC codes introduce variable delays), although these restrictions are less exigent in streaming applications. Thus, the MMSE estimation of the LSP coefficients is combined with media-specific FEC codes (previous LSPs encoded using a secondary codebook) and introducing a maximum delay of 2 frames. Our group presents previous experience in the use of this type of FEC codes and their combination with EC techniques based on MMSE, as well as the use of suitable interleavers for a Bayesian EC.

2.2.2. Image/Video Error Concealment

First, we review the error concealment in images, since EC techniques for video are often based on those ones corresponding for images. In the field of image transmission, it is common the use of techniques based on interpolation. Thus, in coding schemes based on transform, several techniques carry out an interpolation of the adjacent transformed coefficients to those lost/corrupted ones, or work in an alternative domain such as the wavelet transform. In the spatial domain, it is proposed a pixel interpolation using weights derived from the frontier pixels of the adjacent macroblocks. In the field of Bayesian estimation, the MMSE estimation of lost data (with DPCM coding) has been proposed using the quantized indices received before and after the lost ones. An improvement to the previous techniques is carried out in by means of the use of soft-bits (an error probability is assigned to each bit) that can be complemented with the use of MRF. In a combination of the technique presented with FEC codes is proposed in order to carry out an iterative decoding scheme. The problem of error concealment in video sequences is a generalization of the case for images. Thus, these EC techniques use intra-frame information (image problem) and temporal correlations (inter-frame information). These EC techniques can be classified into spatial methods, temporal methods and mixed ones. Regarding spatial methods, a weighted interpolation for H.26l is proposed, also a scheme based on directional decision and intra-frame prediction is employed for H.264/AVC, and a combination of interpolation with analysis/synthesis of textures is proposed in order to obtain a perceptual optimum result. Also, Bayesian approaches based on MAP estimation and MRF, similar to those ones presented for images. Regarding temporal methods, a simple and effective way of reconstruct the corrupted/lost areas of a image is their substitution by the corresponding areas of the previous frame. However, this approach does not present good results in the case of fast movement or sudden scene changes. Other possibility is the use of boundary matching algorithms for recovering lost/damaged movement vectors. The mixed methods present the advantage of employing all the available correlations in order to provide the lost information. The combination of channel coding and EC techniques is also frequent in the field of Video error concealment. Thus, some works apply a flexible macroblok reordering in order to make easier the concealment task. Intra-coding techniques try to mitigate the error propagation of predictive codecs by means of removing the prediction in some macroblocks. The question here is to determine when intra-encoded macroblocks must be inserted taking into account the EC technique included in the decoder. Other alternative is the use of additional information to help the EC method. Thus, in [Frossard01] the introduction of additional data is proposed in order to resynchronize a video sequence when the loss of MPEG-2 packets reduces the video quality (after applying EC) under a determined threshold. In [Kim01], the EC technique is combined with a coding scheme based on two movement vectors (applied to two different macroblocks) in order to prevent the loss of these references. In [Zhu09] a method of using the redundant frames (media-specific FEC) is proposed for H.264/AVC as EC. Finally, data-hiding techniques offer a way to hide some parameters in the bit-stream in order to help the

EC techniques and, thereby, to achieve a better quality.

3. Project Objectives

The reasons why we consider relevant this proposal and the starting hypothesis that support the project objectives can be summarized in next list:

• Bayesian estimation techniques have provided excellent results in speech transmission and recognition applications. However, we think that that the potential of these techniques has not been fully exploited in the field of robust speech recognition in noisy environments. Thus, current techniques have been more focused on modeling the noise effect over the speech features, but not on providing a whole model of source and channel (degradation by noise). We also consider that the possibilities of the Bayesian techniques must be increased with complementary procedures as noise detection and characterization.

• It can be considered as universal principle that any measuring process is inherently uncertain. Thus, during the last years, several methods accepting this principle have arisen. Robust speech recognition is especially suitable in this framework, given the random nature of the noise, as well as the statistical fundaments of current speech recognizers (usually based on hidden Markov modeling). Therefore, we consider that uncertainty processing techniques deserve more analysis and development as it is proposed in the present project proposal.

• Through previous projects, our research team has acquired a wide experience on robust transmission for remote speech recognition (RSR) systems by means of estimation techniques. We think the translation of this experience to a more general field, as robust multimedia transmission, may be very fruitful. Robust speech transmission is an immediate extension of this experience. Also, we consider that it is particularly interesting to extend also our experience to other fields such as image and, especially, video transmission. We also think that statistical mitigation techniques must be combined with certain channel coding techniques specially adapted to the statistical ones.

3.1. Background And Previous Results

In Section 2, we already summarized the main work which supports the research proposed in this Project. Anyway, we extract below the most relevant contributions with respect to the main points mentioned in Section 3:

Estimation for robust speech recognition. There are a number of recent contributions [1] which show that a suitable selection of the statistical models employed for estimation can provide important performance improvements. The introduction of an additional for the source process has been only considered in recent publication, although has been widely studied by our research team for RSR systems [2, 3, 4].

Uncertainty processing for robust speech recognition. Softdata techniques provide a smart framework where the estimates to be processed by the recognizer are considered random variables instead of deterministic values. However, this previous work only considers Gaussian distributions, what is, in general, false, and leaves space for future improvements. Also, exponential feature weighting, even being a heuristic technique, can provide a better performance than the soft-data approach [4]. Therefore, we think that exponential weighting deserves a more in-depth analysis, applying a more formal view, in order to obtain suitable reliability measures. Finally, missing data (MD) techniques provide a joint framework for estimation and uncertainty processing based on the search of reliable regions over the speech spectrograms. Recent work points out that this point of view has a perceptual fundament which makes the MD approach especially interesting for research over the next years.

Robust multimedia signal transmission. In the case of robust speech transmission, our research team [2, 5, 4], as well as other groups, has already successfully applied estimation techniques for error concealment in RSR systems. Thus, we think that the extension and generalization of these techniques to new speech application (VoIP, videoconference, streaming) is quite timely. In this sense, the combination of estimation with media-specific FEC or interleaving looks very convenient [3]. Finally, although estimation techniques have already been applied to image/video, they have not been applied extensively. A possible reason for it is the computational complexity involved by 2D and 3D statistical models [2]. This issue has already been researched by our research team. Again, the combination of concealment techniques with cannel coding is particularly promising.

3.2. Specific Objectives

Our objectives can be summarized as follows:

1. Robust speech recognition against acoustic noise by means of Bayesian estimation techniques (especially MMSE), focusing on new statistical models for the whole process (speech source, and noisy and clean feature spaces) and on the computational efficiency of the resulting estimators. We include here new techniques for complementing and supporting the Bayesian ones (preprocessing, noise detection and modeling).

2. Development of uncertainty processing techniques for robust speech recognition which may be combined with estimation. We consider here new ways for uncertainty measuring and their incorporation to the recognition engine. We also consider new variants of missing data techniques, seeking the collaboration between imputation and marginalization, as well as the definition of new reliability masks.

3. Development of robust multimedia signal transmission techniques for speech and images/video. Again, we consider very especially estimation techniques for damaged/lost data reconstruction, focusing on the statistical modeling of source and channel. As in objective 1, we include complementary techniques (media-specific FEC, interleaving) which may collaborate with estimation.

3.3. Additional Information

For more detailed information about the project and a complete list of bibliographical references, please visit project web page: http://ceres.ugr.es/tsc/tetitrs/

4. References

[1] J.A. González, A. Peinado, A. Gómez, J.L. Carmona, J.A. Morales: ”Efficient VQ-based MMSE for robust speech recognition”, in Proc. ICASSP, 2010.

[2] A.M. Peinado, V. Sánchez, J. Pérez-Córdoba, A. Rubio: ”Efficient MMSE-based channel error mitigation techniques. Application to distributed speech recognition over wireless channels”. IEEE Trans. Wireless Commun., vol. 4 no. 1, 2005.

[3] A.M. Gómez; A.M. Peinado; V. Sánchez; A.J. Rubio; ”Combining Media-Specific FEC and Error Concealment for Robust Distributed Speech Recognition Over Loss-Prone Packet Channels”, IEEE Trans. Multimedia, Vol. 8, No. 6, 2006.

[4] J.L. Carmona; A.M. Peinado; J.L. Pérez-Córdoba; A.M. Gómez; ”MMSE-Based Packet Loss Concealment for CELP-Coded Speech Recognition”, IEEE Trans. Audio Speech Lang. Process., Vol. 18, No. 6, 2010.

[5] A. M. Gómez, A.M. Peinado, V. Sánchez, A.J. Rubio: ”Recognition of Coded Speech Transmitted Over Wireless Channels”. IEEE Trans. on Wireless Communications, vol. 5, pp. 2555-2562, September 2006.