Résumé du projet de recherche (Langue 1)

Background In order to improve the perceived quality of speech communications systems, so-called wideband standards have been developed in recent years. One example is the adaptive multi-rate wideband (AMR WB) speech codec, a part of the 3rd Generation Partnership Project (3GPP). The current trend involves super-wideband (SWB) speech signals with an acoustic bandwidth in excess of 14kHz (Enhanced Voice Services EVS codec developed by 3GPP, Skype Opus codec). Both narrowband and wideband infrastructures are likely to co-exist for some time. There is thus a need to ensure the backward compatibility of narrowband technologies with current and future wideband infrastructure. In consequence, artificial means of extending narrowband speech to wideband speech have been investigated over the last 20 years, e.g. [1, 2, 3, 4]. This technology is known as artificial bandwidth extension (BWX). Most recent approaches to BWX are based on some form of joint-density modelling in which the missing, wideband spectral components are estimated from the available narrowband components, e.g. [1, 3]. These statistical approaches focus on extending the narrowband spectral envelope while avoiding discontinuity artefacts typical of the earlier vector quantisation (VQ) approaches. The work in [3] was among the first to explore a statistical approach based on the modelling of narrow and wideband speech signals using Gaussian mixture models (GMMs), a joint-density estimation procedure and a traditional source-filter speech model. This work was inspired by developments in voice transformation and conversion, e.g. [5, 6]. Most of the recent work has focused on the same source-filter approach with variations in the statistical approach used to extend the narrowband spectral envelope, e.g. via a hidden Markov model (HMM) approach [4]. The work proposed here aims to extend the statistical approaches of the past and to apply the very latest developments in voice conversion to BWX. The first approach should be an extension of the solution in [4] to SWB. The second approach should be a less complex approach based on, for instance, a simple replica of the spectrum as a low-cost approach with potential for rapid integration into Intel products. The third approach for consideration was originally proposed in a study of automatic speaker recognition vulnerabilities to spoofing through voice conversion. The algorithm essentially converts one person’s voice towards that of another, target speaker using two speaker models: one of the source speaker and another of the target speaker. This PhD programme will adapt this technique to BWX and SWB by using models of different bandwidths rather than speakers.
Voice conversion Current Bandwidth Extension from NB signal into WB developed within Intel is derived from work exposed in [4]. Based on combination of extension of Linear Predictive Coefficients (LPC) and of residual signal through spectrum mirroring, such methods has shown good compromise between audio quality and complexity. This project will study the use of similar principle for the bandwidth extension from the WB single into a SWB signal, with maximum re-use of already existing solution. Small adaptation to specificity of WB signal should permit to get reduction of complexity, as it is a priori easier to extrapolate SWB LPC coefficients when WB LPC are available, in comparison to the extrapolation from NB to WB. Bandwidth extension can be achieved by far more simple solutions, like the mirroring or duplication of the full spectrum of the signal. If such solution have shown their poor quality for the extension of NB signal into WB, their quality could be sufficient the WB to SWB conversion. Indeed getting more information already available in the spectrum of voice up to 8 kHz may avoid the need of smart interpolation, based for instance on LPC analysis. Accordingly, as simple reference algorithms or potentially future integration of low-cost bandwidth extension in Intel platform, such simple approach will be investigated. Voice conversion (VC) is a sub-domain of voice transformation [7]. The specific goal of VC is to convert one speaker’s voice x towards that of another y according to a conversion function y=F(x,?), where ? is the set of conversion parameters. Early VQ approaches produced the same discontinuity artefacts observed with VQ implementations of BWX. The joint-density GMM approach alleviates these problems but over-smoothing can produce audible artefacts which degrade the quality of converted speech. A multitude of different approaches are reported in the literature; VC is an active research area with a growing community. This project will study the application of a recent VC approach to BWX investigated in the context of automatic speaker recognition (ASR). While the work will investigate different approaches, on account of its ability to synthesize high quality converted speech signals with no discernible processing artefacts, the initial focus will involve Gaussian dependent filtering (GDF). This algorithm is detailed in the appendix. The core idea is to convert between different bandwidths instead of converting between different speakers, i.e. to replace the target speaker model with a same-speaker, wideband model. Initially, aside from the learning of narrowband and wideband speaker models, the only modification required in order to apply GDF voice conversion to BWX is the extension of the excitation signal. Initially this will be achieved through a similar approach to that in [4] and afterwards through a conversion approach similar to GDF. It will also be possible to simplify the GDF approach by removing the constraints relevant to ASR. In summary, today’s state-of-the-art approached to BWX originated from related work in voice conversion. This project will advance the past work by exploiting the latest developments in voice conversion and related work in ASR.