Summary description of project context and objectives

Motivation

In order to be accepted by users, the voice of a spoken interaction system must be natural and appropriate for the content. Using the same voice for every application is not acceptable to users.

But creating a speech synthesiser for a new language or domain is too expensive, because current technology relies on labelled data and human expertise. Systems comprise rules, statistical models, and data, requiring careful tuning by experienced engineers.

So, speech synthesis is available from a small number of vendors, offering generic products, not tailored to any application domain. Systems are not portable: creating a bespoke system for a specific application is hard, because it involves substantial effort to re-engineer every component of the system. Take-up by potential end users is limited; the range of feasible applications is narrow. Synthesis is often an off-the-shelf component, providing a highly inappropriate speaking style for applications such as dialogue, speech translation, games, personal assistants, communication aids, SMS-to-speech conversion, e-learning, toys and a multitude of other applications where a specific speaking style is important.

Technical background

Speech synthesis

The problem of converting text into speech ("Text-to-speech", or TTS) is invariably broken down into two sub-problems. The text is first converted by the *front end* into a *linguistic specification*. This specification is then used to generate a speech waveform, using one of two methods – concatenation of pre-recorded speech, or generation from a previously-learned model of the speech signal. This proposal is concerned with both the *front end* and the *waveform generation* problems. For waveform generation, we will use statistical parametric models, since this method provides the flexibility required and has a well-founded mathematical framework for learning from data.

Statistical parametric speech synthesis based on HMMs and related models

Over the last decade, and particularly in the last 5 years, statistical parametric speech synthesis has risen in prominence and is now a mainstream method able to produce speech of comparable quality to the well-established method called "unit selection" in which recorded speech is segmented, re-arranged and concatenated to produce novel utterances. The statistical parametric approach has a number of overwhelming advantages over unit-selection, all arising from the fact that it is based on a statistical model. The model can be learned from speech data, in either supervised or unsupervised modes. This learning is performed to maximise an objective measure such as likelihood, generation error or modelled perceptual error. The model can be automatically adapted to small amounts of new data, enabling new voices to be created using as little as 1–100 sentences. The model generates a parametric representation of speech which is then used as input to a vocoder to generate the speech waveform.

In previous work, much of it by members of this consortium and their close collaborators, the flexibility of the statistical parametric approach has been demonstrated in a number of ways. Adaptation can be used to adapt speaker-independent models to particular individual speakers, to create and interpolate expressive and emotional voices, and even to adapt a synthesiser in one language using speech in another language. The method has been shown to be robust to noisy data and imperfect labels. We can also now control speech via the underlying movements of the articulators: this is the first step towards the speech-production-oriented layered models that will be explored in SIMPLE4ALL .

Speech signal models for speech synthesis

Most vocoders (e.g. STRAIGHT) used in statistical parametric speech synthesis are based on source-filter models. The source, for voiced speech, is typically an impulse train, optionally mixed with aperiodic noise. The quality of synthetic speech produced with such vocoders is not good enough. An improvement over the pulse train is to use an estimation of the glottal flow, which can be obtained from the speech signal using the novel technique

of glottal inverse filtering (GIF). The glottal flow characteristics are modelled by the HMMs along with the other parameters such as spectral envelope. For waveform generation, real glottal flow pulses from natural speech are manipulated to form the excitation signal in the vocoder. Recent joint work by AALTO , UH and UEDIN has yielded improved quality speech synthesis. The method also provides a route to controlling voice quality, a key requirement for natural, expressive speech.

Front-end text processing for speech synthesis

Whilst statistical parametric approaches now provide a consistent, unified model with a sound mathematical basis for the waveform generation aspect of speech synthesis, the other half of the speech synthesis problem lags far behind. A typical text-to-speech front end comprises a pipeline of many components, tokenising the input, normalising the text, predicting pronunciations, phrase breaks and pitch accents, for example. Whilst a common system *architecture* – such as that provided in toolkits like Festival or OpenMary – may be portable across languages and domains, the actual components are far from portable. Some components are hand-crafted rules, others use statistical models. It is practically impossible to optimise the entire front-end pipeline to maximise some objective function.

Even when an individual component employs a learning-from-data approach, it will invariably rely on hand-labelled data during learning. The need for hand-labelled "linguistically deep" features (e.g., intonation labels) in order to learn these models is a huge barrier, which is hard to overstate. "Linguistically deep" features require expertise to hand label, yet remain internal to the system. Such labelling is very expensive (we estimate that the most skilled labeller can only process around 10 minutes of speech per day), skilled labellers are hard to find for some languages, and there is poor inter-annotator agreement.

In statistical parametric synthesis, these expensive features are only used as context for the models – in other words, indirectly. We propose to do away with this type of feature, and the un-optimisable pipeline architecture that requires them. We will instead use shallower features, with implicit rather than explicit representation of deep linguistic structure.

The opportunity

The statistical approach to speech synthesis offers the possibility to automatically learn, optimise, adapt and tune a system. This learning can be driven by data (whose likelihood is to be maximised, for example), or other drivers such as user feedback. This project will exploit this property to produce complete learnable systems. The key challenge is to bring the front end text processing stage into a consistent statistical framework where it too can be learned from data by maximising appropriate objective functions, in conjunction with the waveform generation, leading to a complete learnable system.

Objectives

We will develop methods that enable the construction of systems from audio and text data. We will enable systems to learn after deployment. General purpose or specialised systems for any domain or language will become feasible. Our objectives are:

1. ADAPTABILITY: create highly portable and adaptable speech synthesis technology suitable for any domain or language
2. LEARNING FROM DATA AND INTERACTION: provide a complete, consistent framework in which every component of a speech synthesis system can be learned and improved
3. SPEAKING STYLE: enable the generation of natural, conversational, highly expressive synthetic speech which is appropriate to the wider context
4. DEMONSTRATION AND EVALUATION: automatic creation of a new speech synthesiser from scratch, and feedback-driven online learning, with perceptual evaluations.

Description of the work performed since the beginning of the project and the main results achieved so far

Identification of languages for study

We identified the initial set of languages to study near the start of the project, and later extended this by adding further languages. Additional languages may be added later in the project, mainly depending on external collaborators to help provide the data.

Data repository

A fully-functional repository for data has been implemented, with a user-friendly front end that captures the necessary meta-data. The repository is in routine use across the project and has been populated with several data sets.

The alignTk tool

In order to use audiobook data, or other audio with unaligned or mismatching transcriptions, it is necessary to segment the audio into 'utterances' and find the transcription corresponding to each one. In contrast to methods published by others, our method requires no initial acoustic models or language model and operates solely on the data being aligned.

VAD that can be tuned by a naive end user

Part of the processing necessary in 'found' data, including audiobooks, is to detect the regions containing speech and segment at those points, whilst avoiding segmenting at within-utterance short pauses or at stop closures. A tuneable Voice Activity Detector has been tested that allows this to be done on any data. The GMMs used to make the speech/non-speech classification are trained on the data to be segmented, using simple labelling of a small quantity of speech, that can be provided by even a naive user.

Text normalisation that is learned from data

We have demonstrated that a statistical machine translation framework is capable of learning text normalisation from parallel data. Our assumption is that naive users will find it straightforward to create such parallel data (i.e., pairs of unnormalised sentences containing numbers, abbreviations, etc alongside normalised counterparts composed only of proper words). Creating a conventional text normaliser required the construction of regular expressions, hand-crafted rules, and lookup tables, which requires much greater skill and linguistic knowledge.

Complete software framework for the new front-end

The front end, comprising the text normalisation and construction of the linguistic specification, is being implemented within a new software framework. This was our preferred choice, rather than re-purposing an existing framework such as that from Festival. This is because any existing framework would presuppose a pipeline architecture, which is not suitable for our ultimate goal of end-to-end optimisation.

Vector Space Models over letters and words

The principal technique developed so far for constructing a linguistic specification from input text, without reliance on POS taggers, dictionaries, phrase break predictors, and so on, is the use of Vector Space Models to find – in an unsupervised manner – representation that can replace things like POS tagsets or phonetic classes.

Prominence tagging

Experiments in the annotation of prominence from the acoustic signal have given promising results and the method was included in an entry to the Blizzard Challenge 2012 as part of the GlotHMM system.

The GlottHMM vocoder

The vocoder is now mature enough for use across the project and has been tested on various data sets through collaborations between consortium partners. Evaluations have shown it to be superior to the widely-used STRAIGHT

vocoder under most conditions, and always at least as good. It can control voice quality and so far experiments have explored this capability using normal, breathy and Lombard speech styles. The Lombard synthetic speech was found to be more intelligible in noise than normal synthetic speech and in fact as intelligible as natural Lombard speech.

Acoustic and prosodic properties of speech in various speaking styles

Deliverable D5.1 presents a large-scale exploration of these properties on two databases which each contain speech in several styles. Voice source features, derived using the methods from the GlotHMM vocoder were examined in detail for several voice types and for a continuum of degrees of speaking effort.

Speaking style identification experiments

We tested whether it is possible to distinguish between speaking styles using acoustic features. This is a necessary step before proceeding to attempt first to model those features for speech synthesis and then later to predict the corresponding synthesiser control parameters from text in order to generate an appropriate speaking style. Good results were obtained.

Speaker diarization on speech with diverse styles

If we are to build speech synthesisers on diverse data with various speaking styles, then the data preparation pipeline must be robust to such data. Speaker diarization is part of this pipeline when processing data containing multiple speakers, and we have demonstrated that this is robust to wide variations in speaking style and that good Diarization Error Rates can be obtained.

Crowdsourcing word prominence annotations of speech

Whilst we are aiming to escape the constraints of supervised learning from expensive hand-labelled data, we are still considering the use of supervised learning in cases where the labelled data can be obtained cheaply, which generally means from naive listeners. In this experiment, we tested this idea for prominence labelling, with promising results.

Project website

A website with three distinct areas has been created: internal site for recording meetings, external collaborators site for sharing information with them, public-facing site for general dissemination of our work.

External collaborators group

We have formed a group of highly-relevant but sufficiently diverse external collaborators, with a preference for non-academic groups where possible but also allowing for relevant academic groups (e.g., where they can assist us in testing our methods on a wider variety of languages).

First annual report to the external collaborators group

At the end of the first period, we created an attractive and concise report that has ben distributed to all external collaborators.

The address of the project public website is `www.simple4all.org`