Corpus-Based Speech Synthesis: Methods and Challenges

Bernd Möbius

Institute of Natural Language Processing
University of Stuttgart

Abstract

Corpus-based approaches to speech synthesis have been advocated to overcome the limitations of concatenative synthesis from a fixed acoustic unit inventory. The frequency of unit concatenations in, e.g., diphone synthesis has been argued to contribute to the perceived lack of naturalness of synthetic speech. The key idea of corpus-based synthesis, or unit selection, is to use an entire speech corpus as the acoustic inventory and to select at run-time from this corpus the longest available strings of phonetic segments that match a sequence of target speech sounds in the utterance to be synthesized, thereby minimizing the number of concatenations and reducing the need for signal processing.

This paper reviews the assumptions underlying this synthesis strategy and the different approaches to unit selection, as well as the major challenges encountered by corpus-based methods. One of the biggest problems to date is the relative weighting of acoustic distance measures. We further argue against the quest for ever larger speech databases and instead advocate the development of criteria that will help design a speech database with optimal coverage of the target domain—which is often the whole language. We also show that word- or syllable-based approaches are only feasible in strictly closed application domains.

1 Introduction

It has been argued that the large number of concatenation points in a synthesized utterance produces a perceptual impression of unnaturalness (e.g., (Donovan and Woodland, 1999)), even if the spectral discontinuities at the concatenation points are reduced by a careful inventory design based on phonetic criteria. In diphone synthesis, there is a concatenation point in each segment.

A paradigm shift occurred when researchers began to design corpus-based synthesis strategies that consider acoustic units of variable length. A non-uniform unit concatenation method was first proposed by Sagisaka of ATR (Sagisaka, 1988; Takeda, Abe, and Sagisaka, 1990), and a parallel line of research that eventually became known as *unit selection* evolved at the same institution (Black and Campbell, 1995; Hunt and Black, 1996). The complexity of acoustic inventory design shifted from an offline to a runtime selection of units.

The key idea of unit selection is to use an entire speech corpus as the acoustic inventory and to select from this corpus the longest available strings of phonetic segments that match a sequence of target speech sounds in the utterance to be synthesized, thereby minimizing the number of concatenations and reducing the need for signal processing. In an ideal world, the target utterance would be found in its entirety in the speech database and simply played back by the system without any concatenations and without any signal processing applied, effectively rendering natural speech.

Given the complexity and combinatorics of language and speech, this ideal case is extremely unlikely to actually occur in unrestricted application domains. However, given a speech database of several hours worth of recordings, chances are that a target utterance may be produced by a small number of units each of which is considerably longer than a classical diphone or demisyllable. In general, unit selection

speech databases tend to be much larger than diphone databases.

The most extreme view of the unit selection approach has been implemented in the CHATR TTS system (Black and Taylor, 1994), which follows the strategy of not performing any modifications by signal processing whatsoever. The underlying assumption is that the listener will tolerate occasional spectral or prosodic mismatches in an utterance if the quality of the output speech in general approaches that of natural speech.

In the early versions of CHATR this strategy appeared to fail even for languages with comparably simple phonotactics such as Japanese; in informal demonstrations the system would sometimes render locally unintelligible speech. The reason was that the relative weights assigned to segmental and prosodic features were unbalanced, and the selection algorithm would occasionally sacrifice the segmental identity if a competing unit string was found that happened to contain prosodic features that matched closely those required for the target utterance. Obviously, a match between the targeted and the selected segmental strings is an indispensible condition for speech synthesis systems.

In the following, we first present the commonly applied unit selection algorithms, with and without the use of decision trees (sections 2 and 4, respectively). We then address the problems of finding appropriate distance measures and of training the cost functions, and some solutions to these problems (section 3). We also describe the attempt to use phonological descriptors as selection criteria (section 5) and to develop word or syllable concatenation systems for restricted application domains (section 7). The important topic of the design of an optimal speech database for unit selection is discussed in section 6.

2 Selection algorithm

The selection algorithm attempts to minimize two types of cost, one for *unit distortion* and one for *continuity distortion*. Unit distortion (*target cost* in (Hunt and Black, 1996)) is a measure of the distance of the candidate unit from the desired target. Continuity distortion (*concatenation cost* in (Hunt and Black, 1996)) is a measure of the distance between two adjacent units at the concatenation point. This type of cost reflects how well a candidate unit will join with the previously selected unit.

The units in the speech dabase are annotated with multidimensional feature vectors that comprise both segmental and prosodic properties of speech; the annotation is produced by some offline manual or automatic procedure. The feature vector of the target is computed from text at runtime. Therefore, to compute the unit distortion cost only those features can be taken into account that are computable from text. The continuity distortion, on the other hand, can exploit all the features in the feature vector because here only unit candidates are compared that had been available for offline annotation.

Each unit in the database is represented by a state in a state transition network, where the state occupancy costs are given by the measure of unit distortion and the state transition costs are given by the measure of continuity distortion. This design is somewhat reminiscent of HMM-based speech recognition systems. The key difference is in the use of cost functions in the unit selection framework as opposed to the probabilistic models used in speech recognition.

The unit selection algorithm proposed by Hunt and Black (1996) selects from the database an optimal sequence of units by finding the path through the state transition network that minimizes the combined target and concatenation costs.

The unit of choice in this approach is usually a phone-sized unit, or even a sub-

phone unit such as a *half-phone* (Beutnagel et al., 1999; Conkie, 1999) or *demiphone* (Balestri et al., 1999), defined as the portion of the speech signal that is delimited by a phone boundary and a diphone boundary. At first glance this type of unit does not appear to reduce the number of concatenation points in an utterance, compared to a diphone inventory. Another potential disadvantage is that the boundaries between phone-sized units are often in regions that are characterized by rapid spectral and waveform changes, viz. at the transitions between speech sounds. On the other hand, the selection process encourages the exploitation of longer units that contain sequences of phones because units that naturally occur together in the database will have no continuity distortion cost.

3 Distance measures and weights

One of the lessons learned from the early CHATR experiments was to appreciate the difficulties that the weighting of numerous acoustic features presents. This problem has been addressed, if not yet solved, in a number of papers. Two different approaches to weight training have been implemented.

The first, called *weight space search* (Black and Campbell, 1995; Campbell and Black, 1997), samples the total space of weights by means of an analysis-by-synthesis scheme: an utterance is synthesized from the preliminary best set of units in the database, and its distance from the natural waveform is measured. This process iterates over different weight settings and utterances until the process converges on a globally best set of weight values.

The second approach (Hunt and Black, 1996) determines the weights for the two cost functions separately and trains the target costs using *multiple linear regression*. The advantage of the regression method over the iterative weight space search, which is still used for training the concatenation costs, is that separate weights for different

phone classes can be generated, and it is also computationally much less expensive. This method is still used in more recent versions of CHATR (Campbell, Higuchi, and Black, 1998).

New weight training procedures that enhance the efficiency of the exhaustive weight search training and also refine the regression weight training method were recently proposed (Meron and Hirose, 1999). Weight space search was made more efficient by splitting the analysis-by-synthesis process into two separate processes, viz. selection and scoring. The savings in calculations can be put to use either to find better weight combinations by increasing the size of the search space, or to make the weights more robust by considering a larger number of sentences.

Another improvement was made to the regression training, which can now be applied to train target and concatenation costs simultaneously. This is desirable because the two types of costs are not independent of each other. Finally, the new method also considers the costs of prosodic modifications at synthesis time.

Because determining the concatenation costs can be computationally very expensive, one might think of precomputing offline and caching all possible concatenation costs. For all practical purposes this approach is doomed to fail because of the sheer number of such unit combinations. The AT&T system, for example, uses an 84,000 demiphones inventory, yielding 1.8 billion possible unit pairs, plus another 36 million possible mid-phone transitions, a prohibitive number.

However, experiments showed that a subset of 1.2 million unit pair concatenation costs provides a coverage of 99% and that a cache constructed offline from this subset produces unit sequences that are almost entirely (98.2%) identical to units selected using the full search space (Beutnagel, Mohri, and Riley, 1999). The synthesized output speech is reported to be virtually indistinguishable from the optimal selection.

Establishing the relationship between computed ("objective") distances and per-

ceptual differences is a difficult task, and the body of research on this topic is quite small and mainly focussed on speech coding (Quackenbush, Barnwell, and Clements, 1988). In early unit selection experiments (Black and Campbell, 1995) the mean Euclidean cepstral distance between the feature vectors of the target unit and those of the candidate units in the database was used as a score for the set of weights. However, the cepstral distance measure appeared to give higher priority to unit distortion, often at the expense of continuity distortion, whereas human listeners prefered smoother transitions at the concatenation points.

Some insight into the usability of objective distance measures as predictors of perceptual differences in unit selection is provided by Wouters and Macon (Wouters and Macon, 1998; Macon, Cronk, and Wouters, 1998). They attempt to find measures that best predict phonetic variations in the realizations of phonemes. These measures are intended to reflect specific phonetic changes instead of overall quality of distorted (coded) speech and to quantify the distance between two candidate units. Some of the most well-known measures such as mel-based cepstral distance and the Itakura-Saito distance were found to be quite useful, yielding a moderate correlation ($r = 0.66$) with perceptual distances. The authors feel, however, that this strength of correlation is still not sufficient for objective distance measures to be reliable predictors of perceptual differences.

If we assume that the perfect sequence of units to generate a target sequence is not available in a speech database, we need to be able to predict qualitatively or, better, quantitatively the amount of mismatch and distortion that is produced by any of the units that do exist in the database.

Holzapfel and Campbell (1998) use fuzzy logic as the mathematical framework to compute the suitability of a candidate unit for synthesis in a given context. They propose a suitability function for each feature, and the implementation requires that each suitability be in the range of acceptable distances. The effects of big mismatches

are emphasized, and even one unacceptable distance in one unit will disallow the whole sequence of candidate units. The relative importance of a particular criterion can be expressed by the shape of its suitability function. The shapes can be preset according to *a priori* knowledge, or initialized heuristically, or optimized by experiment, with the latter solution requiring subjective perception tests to be run on the synthesized speech.

The implementation of this procedure reduced the amount of signal processing in the Papageno speech synthesis system developed at Siemens (Holzapfel and Campbell, 1998) but failed to significantly improve CHATR, which works without signal processing to begin with.

4 Context clustering

Parallel to the ATR-style unit selection approach, and indeed starting a few years earlier, an alternative method was developed that is based on decision tree clustering (Nakajima and Hamada, 1988; Nakajima, 1994; Itoh, Nakajima, and Hirokawa, 1994; Wang et al., 1993). The key idea is to cluster into equivalence classes all realizations of phonemes that are found in a single-speaker database. Equivalence classes are defined by segmental phonetic context. Clustering is performed by decision trees that are constructed automatically such that they maximize the acoustic similarity within each equivalence class. Each leaf in the tree is represented by a segment ("allophone") and its features, as extracted from the database.

One advantage of this method is that it automatically determines the relative importance of different contextual and coarticulatory effects. Through interpolation even context specifications that were not seen during training can be met. Problematic is the use of units that correspond to speech sounds because they are bounded on both ends by rapid spectral changes. At the same time, at least in Nakajima's exper-

iments for Japanese (Nakajima and Hamada, 1988; Nakajima, 1994), no smoothing or interpolation is performed around the concatenation because the authors argue that essential coarticulatory effects are captured within the units as a consequence of the applied context-oriented clustering algorithm. A modified version of the clustering method has been implemented in the English speech synthesizer developed at Cambridge University (Donovan and Woodland, 1999) and in the IBM speech synthesizer (Donovan and Eide, 1998).

As stated previously, the concept of non-uniform unit concatenation was first proposed by Sagisaka (Sagisaka, 1988; Takeda, Abe, and Sagisaka, 1990). In this work we also encounter for the first time the distinction between unit distortion and continuity distortion, under the notions of spectral pattern difference between the target and the segments that are available in the database, and segment discontinuity, respectively. However, spectral pattern difference was defined heuristically and the continuity criterion was not used during the selection process.

An extension of the non-uniform unit concatenation and the context-oriented clustering approaches was then proposed by the ATR research group (Iwahashi, Kaiki, and Sagisaka, 1992). In their implementation, context clustering is performed by calculating the Euclidean distance between the centroids of segments in different triphonic contexts, considering only the preceding and following contexts. Notice that only the vowel spectrum is evaluated.

The prototypicality of a selected vowel is measured by its Euclidean distance from the cluster centroid. This measure corresponds to the concept of measuring unit distortion. Continuity distortion is straightforwardly calculated as the spectral distance between two adjacent units around the concatenation point. Finally, the ideal cut and concatenation point is determined by taking into account the rate of spectral change as a predictor of speech quality degradation caused by concatenation.

The cost functions are minimized by dynamic programming, and the optimal

sequence of units is found that minimizes the global cost over the sentence to be synthesized. While this methodology still involves considerable signal processing—speech representation is in the cepstral domain and smoothing is performed for unit boundary adjustment—it became the forerunner to CHATR.

A more general and powerful solution to the problem of minimizing inter-segmental distortion was presented by (Iwahashi and Sagisaka, 1995; Sagisaka and Iwahashi, 1995). A unit set is selected from a large database that simultaneously minimizes spectral discrepancies between units as well as the distance of each sound from its cluster centroid. This is a hard optimization problem, and the authors use iterative improvement methods (a deterministic method) and simulated annealing (Kirkpatrick, Gelatt, and Vecchi, 1983; van Laarhoven and Aarts, 1987) (a probabilistic method) to overcome the combinatorial difficulties and arrive at a sub-optimal solution.

The restrictions of this method are, firstly but least importantly, that it is computationally almost prohibitively expensive; secondly, inter-segmental distortion was measured at the temporal mid point of the vowel, semi-vowel, or nasal, in Iwahashi and Sagisaka's diphone-based selection experiments; and thirdly, the optimization algorithm assumes that only one candidate unit sequence exists for each target phone sequence. However, progress was also made, because the proposed method simultaneously minimizes unit and continuity distortions, if only approximately, and because prosodic properties of speech, such as pitch pattern, phone duration and amplitude, were added to the segmental spectral features as selection criteria.

A merger of the context clustering and unit selection approaches was proposed by Black and Taylor (1997) and implemented as an experimental waveform synthesis component in the Festival speech synthesis system (Black, Taylor, and Caley, 1999). Here an offline automatic clustering of segments according to their phonetic and prosodic contexts is performed. The population of candidate units is partitioned

into clusters, such that each cluster contains only units that are similar to each other based on some distance measure.

The key advantage of the merged procedure is that the unit database is organized offline such that the search effort at runtime is significantly reduced. Instead of evaluating all available unit candidates, only the most appropriate cluster of potential candidates is selected by means of a decision tree and then searched for the best unit. The gain in compute time and effort may then be put to use by applying more elaborate optimization algorithms, for instance in the domain of signal processing and prosodic modification.

Macon and colleagues at OGI (Macon, Cronk, and Wouters, 1998; Cronk and Macon, 1998) have proposed several improvements to the clustering method. They enhance the decision tree's capability to generalize to input vectors unseen in the training data. This is achieved by using cross-validation during tree growing to optimize the decision of when to stop partitioning the data. Generalization power will suffer if trees are allowed to grow too far and become biased towards the units seen in training. The authors also observe that their method of tree-structured clustering yields fuller clusters with lower objective distances.

In the experiments reported in (Black and Taylor, 1997) no signal modification was performed. The authors emphasize, however, that it might be advantageous to allow prosodic modifications if major discontinuities can thereby be avoided; this had previously been suggested by Hauptmann (1993). The cost of eventually necessary signal modification could be included in the scoring during unit selection.

Prosodic features of speech are explicitly incorporated as selection criteria in the recent version of the Eloquens speech synthesis system developed at CSELT (Balestri et al., 1999); they also play a crucial role for the design of the speech database. It is argued that imposing model-based artificial prosody is still required to achieve the desired level of prosodic flexibility. In the current implementation sentence modes

other than declaratives are generated by rule-based prosodic modifications.

In this system the unit selection algorithm extracts the longest suitable sequence of demiphones, taking into account categorical prosodic labels as well as their acoustic correlates in terms of F_0 and duration values. The segmental context is evaluated by a bell-shaped window function centered on the demiphone in question. A unit similarity measure is also applied, as is a concatenation factor that encourages the selection of co-occurring demiphones. Experiments suggest that driving the unit selection by categorical prosodic features yields better results than matching numerical prosodic values.

5 Phonetic and phonological trees

By way of a phonological specification of an utterance, explicit reference to acoustic properties is avoided altogether (Allen, 1992). The key idea in this kind of approach is that most of the variability in the speech signal is predictable and that units selected from the appropriate context are likely to have the right specifications.

For instance, vowel reduction in English comprises not only the substitution of a full vowel with a schwa and some local durational adjustments; much more subtle and complex spectral and timing modifications are involved in the process, and one may get them for free if the right and right-sized unit is selected from the appropriate context. In the prosodic domain, given a perfect match of target and candidate contexts, pitch and duration will closely resemble the desired contours and values, and no error-prone model-mediated specification is required. Another advantage of this method is that the linguistic text analysis components of a TTS system tend to generate phonological representations more reliably than phonetic specifications.

In BT's Laureate speech synthesis system (Breen and Jackson, 1998) unit se-

lection is based exclusively on phonologically motivated criteria, disregarding the actual acoustic feature vectors in the target and the candidate units. Two speech sounds are defined to be matching each other if their phonological annotations are identical. Phonetic attributes and the phonological context of a speech sound are represented by a *phoneme context tree*, a pre-processed structure on which the run-time unit search is performed.

A context tree window, centered on the phone in question, defines the length of the unit, in terms of number of phones, that is considered for unit selection. While the window size can be of arbitrary length, computational efficiency limits the unit length to be triphonic in the Laureate system. During a database search all triphonic units matching the target specification are entered into a workspace. To determine the similarity between the target and the candidates a non-binary distance metric is applied that relies on a set of abstract attributes. The weight, or relative importance, of individual discriminating features is assigned through a combination of knowledge bases, including linguistic theory, signal processing, and clustered acoustic data. The attribute set comprises a subset of the classical distinctive features (Chomsky and Halle, 1968), articulatory features, features related to syllable structure, and other suprasegmental features. Segment identity is assigned higher priority than context matching.

Taylor and Black (1999) present a similar approach, *phonological structure matching*, where phonological information, such as canonical pronunciation, positional factors and accentuation, is used for unit selection instead of narrow phonetic transcriptions and absolute duration and F_0 values. In contrast to the BT implementation (Breen and Jackson, 1998), the basic unit may be a single phone, a sub-syllabic constituent (e.g., syllable onset), a syllable, a word, or a phrase. For every target the entire database, represented as phonological trees, is searched, starting with the highest node level and resorting to daughter nodes whenever no candidate

is found. As a result of the search, candidate units will appear at various positions and levels in the tree, and they will correspond to units of arbitrary length in the database.

One potential drawback of Taylor and Black's (1999) approach is that word boundaries appear to represent hard boundaries in the phonological tree. The authors argue that coarticulation has been found to be stronger within constituents than across constituent boundaries at all levels in the tree. However, the claim that the word boundary is a strict coarticulation barrier is at best controversial—consider examples such as *sun glass* [sʌŋglæs] and *this year* [ðɪʃiɹ]—and diphone inventories of state-of-the-art synthesizers never fail to include cross-word units (Portele, 1996; Möbius, 1999). Notice that the idea of reduced coarticulation at word boundaries also underlies the recent word concatenation approaches (Lewis and Tatham, 1999; Stöber et al., 1999) (see section 7).

When selecting the best candidate, prosody is classified as secondary information in Taylor and Black's method. The reason is presumably that the phone identity needs to be protected from being overruled by, e.g., perfectly matching prosodic features. Signal processing based on residual-excited LPC is performed to modify duration and pitch if even the best candidate shows too much of a prosodic discrepancy from the target. Thus, natural prosody is preserved wherever possible and model-based prosody is imposed whenever required. A similar strategy has been suggested by Conkie (1999) who concludes that unit selection does not render signal processing obsolete and that both techniques can be usefully applied in the same framework.

This design again reflects the current lack of appropriate relative weighting of large sets of features. It is also an illustration of the combinatorics of language and speech, due to which many units will not be found in a speech database.

6 Speech database

Defining the optimal speech database for unit selection has become one of the most important research issues in speech synthesis. A well-designed speech corpus has a huge impact on the quality of the synthesized speech, no matter what the basic unit is defined to be, a phone, a demiphone, a diphone, or even a triphone.

In the earlier unit selection experiments databases were of moderate size, albeit usually larger than for diphone inventories, typically less than one hour's worth of speech. For instance, the IBM speech synthesizer (Donovan and Eide, 1998) was trained on 45 minutes of speech, the related one at Cambridge University on 60 minutes (Donovan and Woodland, 1999). The monolingual speech databases of CHATR included sets of phonetically balanced sentences and isolated words as well as radio news sentences (Black and Campbell, 1995). Shortly later it was estimated that 40 minutes of speech would be generally adequate, and as little as 20 minutes for Japanese (Campbell, Higuchi, and Black, 1998). Even more recently, database design amounted to the instruction for the speakers to "bring a novel or short-story of their own choice" (Campbell, 1999a, page 43).

This strategy appears to be in overt conflict with the belief that to be able to benefit from long acoustic units, a meticulous design of the text materials to be recorded is required. The database should be designed or constructed such as to include all relevant acoustic realizations of phonemes, a point made already by Iwahashi and Sagisaka (1995).

In building the two-hour speech database of the AT&T speech synthesis system (Beutnagel et al., 1999; Conkie, 1999), the main focus was on achieving robust unit selection that enables the synthesis quality to be consistently high. To this end, diphone coverage was carefully controlled so that sufficient examples of pairs of speech sounds were collected. Moreover, different types of textual materials were considered

that were intended to be close to the target applications, including newspaper text and interactive "prompt-style" sentences, to cover a variety of prosodic contexts and speaking styles. The second important decision with an impact on synthesis quality was to manually correct the segmentation and annotation of the speech database after a first-pass automatic processing. As previously reported by Black and Campbell (1995), accurately labeled smaller databases tend to yield better synthetic quality than large automatically segmented databases.

In the CSELT system (Balestri et al., 1999) large text corpora that are representative of the intended domain were statistically analyzed to define the segmental and prosodic coverage. From these corpora a much smaller subset of text that has the same coverage was extracted with the help of a greedy algorithm (van Santen and Buchsbaum, 1997). The resulting text materials were intended to be well-formed sentences of regular structure and reasonable length, and they should enable the speaker to read them easily and with the expected prosodic patterns. Redundant portions of the text can be pruned to further reduce the amount of recordings.

Reducing the size by removing redundant sentences and units was also suggested by Campbell (1999b). The underlying idea is that smaller databases can be segmented and annotated more reliably. The problem is, however, that in order to be able to decide which parts of the corpus can be pruned without a significant loss of coverage, choice of units, and thus output speech quality, the original (larger) body of speech already needs to have been annotated—a vicious circle.

It has also been argued plainly that databases need to become even larger (e.g., (Campbell, 1999a)). More formally, the AT&T group has observed that each time the number of units in the database was increased, the quality of the output speech also improved significantly (Conkie, 1999).

How big, then, does the database have to be to achieve optimal coverage? There are hardly any systematic studies of coverage in the area of speech synthesis, with

the exception of (van Santen, 1997), and the results from this study are quite discouraging.

For example, van Santen constructed a contextual feature vector for diphone units that included key prosodic factors such as word accent status and position in the utterance. He then computed the *coverage index* of training sets, which is defined as the probability that all diphone-vector combinations occurring in a randomly selected test sentence are also represented in the training set. It turned out that a training set of 25,000 combinations had a coverage index of 0.03, which means that the probability is 0.03 that the training set covers all combinations occurring in the test sentence. To reach a coverage index of 0.75 a training set of more than 150,000 combinations is required. Given that the factors used for the feature vector were coarse and few, unit selection approaches based on diphone units would require absurdly large speech databases to achieve reasonable coverage.

Moreover, as van Santen points out, the results look even worse if one computes coverage indexes of training sets that are selected from text genres which differ from the genre of the test sentences. Together with equally discouraging results for a few other scenarios, such as for using obstruent-terminated units, these findings shed an unfavorable light on corpus-based speech synthesis approaches that attempt to cover an unrestricted domain—typically, the whole language—by simply re-sequencing recorded speech.

If it is practically impossible to construct an optimal speech database, what are the requirements of a corpus if approximate coverage is the goal? The answer, again, is tentative, and pessimistic.

Many aspects of language and speech can be characterized as belonging to the LNRE (*Large Number of Rare Events*) class of distributions. LNRE classes have the property of extremely uneven frequency distributions: while some members of the class have a high frequency of occurrence, i.e. they are types with a high token

count, the vast majority of the class members are extremely rare, and many of them are in fact types that occur only once in a given corpus. LNRE distributions are also observed, for instance, in feature vectors used in segmental duration modeling (van Santen, 1995; Möbius and van Santen, 1996) and in the estimation of word frequencies (Baayen, 2000).

Evidently, LNRE distributions also play a crucial role in data-driven synthesis. For example, Beutnagel and Conkie (1999) report that more than 300 diphones out of a complete set of approximately 2000 diphones occur only once in a two-hour database recorded for unit selection. These rare diphones were actually included in the database only by way of being embedded in carefully constructed sentences; from the start, they were not expected to occur naturally in the recorded speech at all. The authors also observe that the unit selection algorithm prefers these rare diphones for target sentences instead of concatenating them from the smaller demiphone units, which means that they also generate superior synthesis quality compared to the demiphone solution.

For the construction of the database for a new Japanese synthesis system (Tanaka et al., 1999) 50,000 multi-form units were collected that cover approximately 75% of Japanese text. Multi-form units are designed to cover all Japanese syllables and all possible vowel sequences, realized in a variety of prosodic contexts. In conjunction with another set of 10,000 diphone units this database accounts for 6.3 hours of speech. Given the relatively simple syllable structure of Japanese, the emphasis should be on *only* 75% coverage. Figure 2 in (Tanaka et al., 1999) illustrates that increasing the unit inventory to 80,000 does not result in a significantly higher coverage, and the growth curve appears to converge to about 80%. The authors state that for unrestricted text the actually required number of units approaches infinity, and that the majority of the units are rarely used—a characteristic of LNRE distributions. The question of how to get to near 100% coverage remains unanswered, in

fact even unasked.

The LNRE characteristics of speech are often unrecognized and the pertinent problems underestimated. For example, it is a common attitude to accept poor modeling of less frequently seen or unseen contexts because "they are less frequently used in synthesis" (Donovan and Woodland, 1999, page 228). The perverse nature of LNRE distributions is the following: the number of rare events is so large that the probability of encountering at least one of these events in a particular sample, such as in a sentence to be synthesized, approaches certainty.

In the Laureate system (Breen and Jackson, 1998) an attempt is made to optimize the database based on linguistic criteria. The result is a speech database that contains at least one instance of each diphone in the language. This baseline inventory is augmented by embedding the diphones not in carrier phrases but in phonetically rich text passages. This self-restrained optimization attempt is a consequence of the fact that annotation and quality control is considered to be too unreliable for larger databases. The authors argue that it is also difficult to ensure a consistent speaking style in a large set of recordings and that speech segments from very different styles will result in a patchwork of concatenated speech. Speaking style itself is currently not considered to be a useful selection criterion.

Established techniques from speech recognition have been applied not only in the work by Hunt and Black (1996) but also by Holzapfel and Campbell (1998), in the latter work to enhance generalization to unseen cases in runtime unit selection. They train a set of triphone HMM's on the speech database to assess the similarity of segmental contexts. All contexts of each phone are first pooled; the pools are then iteratively split according to phonetically motivated criteria, with a maximum likelihood criterion ensuring optimal improvement of the models with every split of a cluster. By classifying the contexts according to the criteria learned by the clustering tree, triphone contexts that do not occur in the database and were unseen during

training can be reconstructed and mapped appropriately, a standard procedure in speech recognition (Jelinek and Mercer, 1980; Young, 1992). A similar approach was implemented in Microsoft's TTS system (Huang et al., 1996; Hon et al., 1998).

7 Word and syllable concatenation

> "Attempts to record and play back words have not been successful, largely due to the large and changing number of words and the need to make contextual adjustments." (Allen, 1992, page 768)

For restricted domains a version of the unit selection method might be feasible that exploits units larger than demiphones, phones, or diphones. In the most recent version of the synthesis component developed in the Verbmobil project (Wahlster, 1997), a word concatenation approach has been implemented (Stöber et al., 1999). The Verbmobil domain comprises a fixed lexicon of about 10,000 words from the travel planning domain.

Each word in the domain was recorded in a variety of prosodic and positional contexts. A statistical analysis of a text corpus from the pertinent domain, i.e. real travel planning dialogs, was performed to achieve the best coverage of possible sentence structures. Additionally, names of months and week days, numbers and a few other high-frequency words in the domain were recorded in all relevant prosodic contexts. Position in the utterance, sentence mode, segmental reduction, and prominence were passed as relevant information to the unit selection algorithm. The only signal processing step applied was a simple amplitude smoothing on all adjacent words that do not co-occur in the database.

However, the Verbmobil domain is not restricted in all respects. Its lexicon has a loophole that allows proper names to sneak into the domain. To synthesize

these names, and novel words in general, the system resorts to diphone synthesis. This strategy is not altogether satisfactory because the quality difference between phrases generated by word concatenation and the high-entropy novel words synthesized from diphones is too striking. To extend the word concatenation approach to unrestricted domains, a procedure is suggested that would enable the system to synthesize words from syllables and syllables from phones. Furthermore, imposing prosodic manipulations of the synthetic signal is again an option to consider (Stöber et al., 1999).

A system based on word and syllable concatenation has also been presented by Lewis and Tatham (1999), for the limited domain of weather forecasting. The system, MeteoSPRUCE, has an inventory of 2000 recorded monosyllabic and polysyllabic words. There are numerous problems with this approach. For instance, monosyllables are embedded in a fixed-context carrier phrase during recordings, making them almost automatically inappropriate for recombination. Also, some of the recombination rules appear to be of an ad-hoc nature, such as "cut three periods from the start or end of syllables whose onsets or codas are periodic." The authors admit that such rules will probably have to be modified for other voices or recording rates. These problems notwithstanding, they are confident that their synthesis strategy can be extended to much larger databases and to unrestricted text-to-speech scenarios.

8 Looking ahead

Recent unit selection based speech synthesis systems can be characterized by their uneven performance. When good unit sequences are available in the speech database, the speech output quality approaches that of natural speech. But very often stretches of almost perfect synthetic speech are interrupted by some very poor unit, usually a

consequence of distortions at the concatenation point. Evidently, the main problem is to achieve a consistently high speech quality.

In the light of the mostly successful concatenations it is tempting to brush aside the depressing results of van Santen's (1997) study, which seem to insinuate that unit selection can never work because of the complexity and combinatorics of language and speech. But then, do the occasional, perceptually very disturbing glitches not in fact confirm van Santen's results? It is the rare vectors and combinations that are poorly modeled, and one or the other of these rare events indeed shows up when utterances are synthesized, just as predicted by the LNRE distribution models.

There are at least two avenues for further progress in unit selection. One promising line of research is to increase the coverage of speech databases by a careful design, i.e., by defining the linguistic and phonetic criteria that the database should meet. This should be complemented by further systematic studies of the correlations between objective distance measures and perceptual ones. The second area of research is the design of databases for restricted application domains, where the distributions of linguistic factors are known. In this type of systems speaking style can become a useful selection criterion, and it might even be feasible to include idiosyncratic or context-dependent pronunciation variants in the speech database.

References

Allen, Jonathan. 1992. Overview of text-to-speech systems. In Sadaoki Furui and M. Mohan Sondhi, editors, *Advances in Speech Signal Processing*. Marcel Dekker, New York, pages 741–790.

Baayen, Harald. 2000. *Word Frequency Distributions*. Kluwer, Dordrecht.

Balestri, Marcello, Alberto Pacchiotti, Silvia Quazza, Pier Luigi Salza, and Stefano

Sandri. 1999. Choose the best to modify the least: a new generation concatenative synthesis system. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 5, pages 2291–2294.

Beutnagel, Mark and Alistair Conkie. 1999. Interaction of units in a unit selection database. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 3, pages 1063–1066.

Beutnagel, Mark, Alistair Conkie, Juergen Schroeter, Yannis Stylianou, and Ann Syrdal. 1999. The AT&T Next-Gen TTS system. In *Collected Papers of the 137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association: Forum Acusticum (Berlin, Germany).* Paper 2ASCA_4.

Beutnagel, Mark, Mehryar Mohri, and Michael Riley. 1999. Rapid unit selection from a large speech corpus for concatenative speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 2, pages 607–610.

Black, Alan W. and W. Nick Campbell. 1995. Optimising selection of units from speech databases for concatenative synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Madrid, Spain)*, volume 1, pages 581–584.

Black, Alan W. and Paul Taylor. 1994. CHATR: a generic speech synthesis system. In *Proceedings of the International Conference on Computational Linguistics (Kyoto, Japan)*, volume 2, pages 983–986.

Black, Alan W. and Paul Taylor. 1997. Automatically clustering similar units for

unit selection in speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece)*, volume 2, pages 601–604.

Black, Alan W., Paul Taylor, and Richard Caley, 1999. *The Festival speech synthesis system—system documentation.* CSTR Edinburgh. Edition 1.4, for Festival version 1.4.0. [http://www.cstr.ed.ac.uk/projects/ festival/manual-festival_toc.html].

Breen, Andrew P. and P. Jackson. 1998. Non-uniform unit selection and the similarity metric within BT's Laureate TTS system. In *Proceedings of the Third ESCA Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 373–376.

Campbell, W. Nick. 1999a. A call for generic-use large-scale single-speaker speech corpora and an example of their application in concatenative speech synthesis. *Technical Publications, ATR Interpreting Telecommunications Research Laboratories*, pages 42–47.

Campbell, W. Nick. 1999b. Reducing the size of a speech corpus for concatenation waveform synthesis. *Technical Publications, ATR Interpreting Telecommunications Research Laboratories*, pages 90–91.

Campbell, W. Nick and Alan W. Black. 1997. Prosody and the selection of source units for concatenative synthesis. In Jan van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, editors, *Progress in Speech Synthesis.* Springer, New York, pages 279–292.

Campbell, W. Nick, Norio Higuchi, and Alan Black. 1998. Chatr: a natural speech re-sequencing synthesis system. Draft, April 8, 1998. [http://www.itl.atr.co.jp/chatr/papers.html].

Chomsky, Noam and Morris Halle. 1968. *The Sound Pattern of English.* Harper & Row, New York.

Conkie, Alistair. 1999. Robust unit selection system for speech synthesis. In *Collected Papers of the 137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association: Forum Acusticum (Berlin, Germany).* Paper 1PSCB_10.

Cronk, Andrew and Michael Macon. 1998. Optimized stopping criteria for tree-based unit selection in concatenative synthesis. In *Proceedings of the International Conference on Spoken Language Processing (Sydney, Australia)*, volume 5, pages 1951–1954.

Donovan, Robert E. and E. M. Eide. 1998. The IBM trainable speech synthesis system. In *Proceedings of the International Conference on Spoken Language Processing (Sydney, Australia)*, volume 5, pages 1703–1706.

Donovan, Robert E. and P. C. Woodland. 1999. A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and Language*, 13:223–241.

Hauptmann, A. G. 1993. A first experiment in concatenation synthesis from a large corpus. In *Proceedings of the European Conference on Speech Communication and Technology (Berlin, Germany)*, pages 1701–1704.

Holzapfel, Martin and Nick Campbell. 1998. A nonlinear unit selection strategy for concatenative speech synthesis based on syllable level features. In *Proceedings of the International Conference on Spoken Language Processing (Sydney, Australia)*, volume 6, pages 2755–2758.

Hon, Hsiao-Wuen, Alex Acero, Xuedong Huang, Jingsong Liu, and Mike Plumpe. 1998. Automatic generation of synthesis units for trainable text-to-speech sys-

tems. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (Seattle, WA)*, volume 1, pages 293–296.

Huang, Xuedong, Alex Acero, Jim Adcock, Hsiao-Wuen Hon, John Goldsmith, Jingsong Liu, and Mike Plumpe. 1996. Whistler: A trainable text-to-speech system. In *Proceedings of the International Conference on Spoken Language Processing (Philadelphia, PA)*, volume 4, pages 2387–2390.

Hunt, Andrew J. and Alan W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (München, Germany)*, volume 1, pages 373–376.

Itoh, K., Shin-ya Nakajima, and T. Hirokawa. 1994. A new waveform speech synthesis approach based on the COC speech spectrum. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (Adelaide, Australia)*, volume 1, pages 577–580.

Iwahashi, Naoto, Nobuyoshi Kaiki, and Yoshinori Sagisaka. 1992. Concatenative speech synthesis by minimum distortion criteria. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (San Francisco, CA)*, volume 2, pages 65–68.

Iwahashi, Naoto and Yoshinori Sagisaka. 1995. Speech segment network approach for an optimal synthesis unit set. *Computer Speech and Language*, 9:335–352.

Jelinek, Frederick and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice.*

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220:671–680.

Lewis, Eric and Mark Tatham. 1999. Word and syllable concatenation in text-to-speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 2, pages 615–618.

Macon, Michael W., Andrew E. Cronk, and Johan Wouters. 1998. Generalization and discrimination in tree-structured unit selection. In *Proceedings of the Third ESCA Workshop on Speech Synthesis (Jenolan Caves, Australia)*, pages 195–200.

Meron, Yoram and Keikichi Hirose. 1999. Efficient weight training for selection based synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 5, pages 2319–2322.

Möbius, Bernd. 1999. The Bell Labs German text-to-speech system. *Computer Speech and Language*, 13:319–358.

Möbius, Bernd and Jan van Santen. 1996. Modeling segmental duration in German text-to-speech synthesis. In *Proceedings of the International Conference on Spoken Language Processing (Philadelphia, PA)*, volume 4, pages 2395–2398.

Nakajima, Shin-ya. 1994. Automatic synthesis unit generation for English speech synthesis based on multi-layered context oriented clustering. *Speech Communication*, 14:313–324.

Nakajima, Shin-ya and Hiroshi Hamada. 1988. Automatic generation of synthesis units based on context oriented clustering. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (New York, NY)*, pages 659–662.

Portele, Thomas. 1996. *Ein phonetisch-akustisch motiviertes Inventar zur Sprachsynthese deutscher Äußerungen.* Niemeyer, Tübingen.

Quackenbush, Schuyler R., T. P. Barnwell, and Mark A. Clements. 1988. *Objective Measures of Speech Quality.* Prentice Hall, Englewood Cliffs, NJ.

Sagisaka, Yoshinori. 1988. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (New York, NY)*, pages 679–682.

Sagisaka, Yoshinori and Naoto Iwahashi. 1995. Objective optimization in algorithms for text-to-speech synthesis. In W. Bastiaan Kleijn and Kuldip K. Paliwal, editors, *Speech Coding and Synthesis.* Elsevier, Amsterdam, pages 685–706.

Stöber, Karlheinz, Thomas Portele, Petra Wagner, and Wolfgang Hess. 1999. Synthesis by word concatenation. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 2, pages 619–622.

Takeda, Kazuya, Katsuo Abe, and Yoshinori Sagisaka. 1990. On unit selection algorithms and their evaluation in non-uniform speech synthesis. In Gérard Bailly and Christian Benoît, editors, *Proceedings of the ESCA Workshop on Speech Synthesis (Autrans, France)*, pages 35–38.

Tanaka, Kimihito, Hideyuki Mizuno, Masanobu Abe, and Shin-ya Nakajima. 1999. A Japanese text-to-speech system based on multi-form units with consideration of frequency distribution in Japanese. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 2, pages 839–842.

Taylor, Paul and Alan W. Black. 1999. Speech synthesis by phonological structure matching. In *Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary)*, volume 2, pages 623–626.

van Laarhoven, P. J. M. and E. H. L. Aarts. 1987. *Simulated Annealing: Theory and Applications*. Reidel, Dordrecht.

van Santen, Jan P. H. 1995. Computation of timing in text-to-speech synthesis. In W. Bastiaan Kleijn and Kuldip K. Paliwal, editors, *Speech Coding and Synthesis*. Elsevier, Amsterdam, pages 663–684.

van Santen, Jan P. H. 1997. Combinatorial issues in text-to-speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece)*, volume 5, pages 2511–2514.

van Santen, Jan P. H. and Adam L. Buchsbaum. 1997. Methods for optimal text selection. In *Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece)*, volume 2, pages 553–556.

Wahlster, Wolfgang. 1997. VERBMOBIL: Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache. Technical report, DFKI, Saarbrücken. Verbmobil-Report 198.

Wang, W. J., W. Nick Campbell, Naoto Iwahashi, and Yoshinori Sagisaka. 1993. Tree-based unit selection for English speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (Minneapolis, MN)*, volume 2, pages 191–194.

Wouters, Johan and Michael W. Macon. 1998. A perceptual evaluation of distance measures for concatenative speech synthesis. In *Proceedings of the International*

Conference on Spoken Language Processing (Sydney, Australia), volume 6, pages 2747–2750.

Young, Steve. 1992. The general use of tying in phoneme-based HMM speech recognisers. In *Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing (San Francisco, CA)*, volume 1, pages 569–572.