Speech Output Generation in GoalGetter

Esther Klabbers*†

Abstract

In this paper a method for speech output generation in data-to-speech systems is proposed, called phrase concatenation, which tries to find a balance between naturalness and flexibility of the speech output. The GoalGetter system, which generates spoken monologues on football matches, serves as an example. The phrase concatenation technique involves concatenating prerecorded words and phrases, which is new in that different prosodic versions of otherwise identical phrases are recorded.

Introduction

The main issue addressed in this paper is the problem of generating high quality speech in data-to-speech systems, i.e., systems which present data in the form of spoken monologues, sometimes also called concept-to-speech systems. Data-to-speech generation is a relatively new area of research. Traditionally, research on spoken-language generation was mainly undertaken within the separate fields of natural-language generation and text-to-speech synthesis. State-of-the-art language generation is capable of generating flexible utterances and texts, but often the intonational properties are not taken into account. Text-to-speech synthesis often fails to generate adequate prosody due to the lack of information available in texts. In contrast to text-to-speech systems, explicit discourse models can be reliably constructed in data-to-speech systems, so that a more natural prosody can be achieved.

The method of speech output generation is explained in the context of a simple data-to-speech system called GoalGetter, which generates spoken monologues on football matches. GoalGetter generally works as follows: it takes as input a Teletext page that contains summary information on a particular football match. The Teletext page lists the two teams that played against each other, the score, which players scored when, etc. From this concise information, the language generation module (LGM) generates a coherent text using syntactic templates. The output text, enriched with prosodic markers, is passed on to the speech generation module (SGM), which makes it audible through one of two output modes, i.e., diphone synthesis or phrase concatenation.

---

*IPO, Center for Research on User-System Interaction

†This research is carried out within the framework of the Priority Programme Language and Speech Technology (TST). The TST-Programme is sponsored by NWO (Netherlands Organization for Scientific Research).

Before explaining the phrase concatenation technique, it is necessary to get a general idea of the working of the LGM. It is responsible for the content and form of the utterances and the prosodic properties, and as such sets the pre-conditions the SGM has to satisfy.

1 Language Generation in GoalGetter

The technique used for natural language generation in GoalGetter was originally developed at IPO for an English-spoken database query system called Dial-Your-Disc (DYD). This system generates spoken monologues about compact discs with musical compositions written by Mozart (van Deemter, Landsbergen, Leermakers, and Odijk 1994). The architecture of the LGM is depicted in Figure 1.

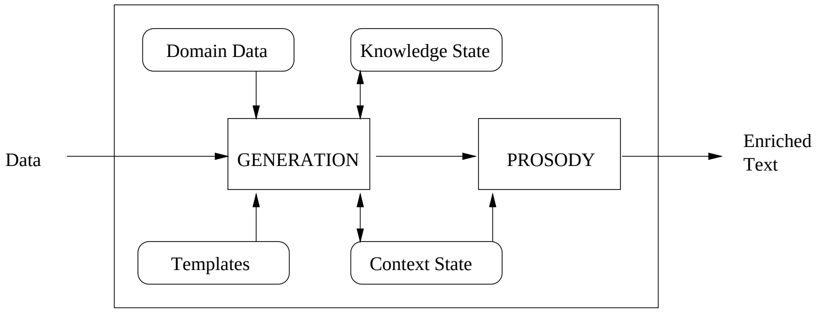
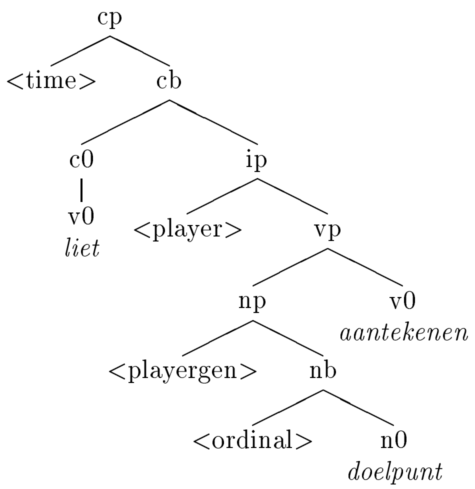


Figure 1: The architecture of the Language Generation Module (LGM)

team 1: PSV goals 1: 1 team 2: Ajax goals 2: 3 goal 2: Kluivert (5) goal 2: Kluivert (18) goal 2: Blind (83/pen) goal 1: Nilis (90) referee: Van Dijk spectators: 25.000 yellow 1: Valckx	"De "wedstrijd tussen "PSV en "Ajax / eindigde in "@een // - "@drie /// "Vijfentwintig duizend "toeschouwers / bezochten het "Philipsstadion /// "Ajax nam na "vijf "minuten de "leiding / door een "treffer van "Kluivert /// "Dertien minuten "later / liet de aanvaller zijn "tweede doelpunt aantekenen /// De % "verdediger "Blind / verzilverde in de "drieentachtigste minuut een "strafschop voor Ajax /// Vlak voor het "eindsignaal / bepaalde "Nilis van "PSV de "eindstand / op "@een // - "@drie /// % "Scheidsrechter van "Dijk / "leidde het duel /// "Valckx van "PSV kreeg een "gele "kaart ///

Figure 2: Example input and output of the LGM

The input for the *Generation* module in the LGM is formed by a textual representation of a teletext page on a particular football match (see Figure 2). It also uses a database that contains fixed background data about e.g., the names of the stadiums and the field positions for each player (defender, goalkeeper). To generate sentences, the Generation module uses a set of so-called syntactic templates. These are basically syntactic parse trees with fixed parts, *carriers*, and variable parts, *slots*, in which other syntactic templates can be inserted. An example template is depicted in Figure 3. The templates have conditions attached to them about when they can be used. For instance, a template expressing the number of spectators of a match can only be used after the match was introduced, e.g. by naming both teams. In order to be able to check which information is already known, a *Knowledge State* is maintained. Furthermore, to ensure the well-formedness of referring expressions used to fill the template slots, we need information about which discourse objects have been mentioned, and how and when they have been referred to. This is recorded in the *Context State*. Each piece of information in the data structure can be expressed by at least one template. To allow for more variation in the output text, more templates can be implemented to express the same information in different ways, which are selected randomly.



CONDITIONS:
TOPIC goalscoring
time ← express:[currentevent.time, currentmatch, c]
player ← express:[currentevent.player, currentmatch, c] / nom
playergen ← express:[currentevent.player,currentmatch,c] / gen

Figure 3: Syntactic template for the sentence *Dertien minuten later liet Kluivert zijn tweede doelpunt aantekenen*

In the last stage of text generation, the *Prosody* module computes the accents and prosodic boundaries taking the properties of the Context State into account. The accentuation algorithm is based on a version of Focus-Accent Theory (van Deemter (1994); Dirksen (1992)), where binary branching metrical trees are used to represent the relative prominence of nodes with respect to pitch accent. After accentuation, phrase boundaries are assigned. The output of the LGM is an enriched text i.e., a coherent text with prosodic markers (see Figure 2), which is passed on to the SGM. The prosodic markers will be discussed in Section 3.1. For a more extensive explanation of the LGM see (Klabbers, Odijk, de Pijper, and Theune 1996).

2 Speech output generation methods

In commercial data-to-speech systems, it is important that the voice output interface be of high quality. There are several methods to provide a system with speech output, each with their advantages and disadvantages. Three methods are distinguished here, viz. the use of prerecorded speech, speech synthesis and speech concatenation.

2.1 The use of prerecorded speech

A maximum degree of naturalness can be achieved by playing back digitally stored natural speech. In the past, several information announcement systems have been created to provide such services as weather, motoring and tourist information, recipes, and bed-time stories. The speech output was created by simply making recordings of the whole information base and playing a loop or disc continuously throughout the day (Waterworth 1984). This approach has two main disadvantages. Firstly, memory and storage limitations will become a problem once the vocabulary of the system becomes too large. Secondly, the approach is highly inflexible in that entire messages have to be re-recorded to update the vocabulary.

For GoalGetter, the vocabulary consists of a limited set of carrier sentences and a more extensive set of variable words that can be inserted in the slots (*slot fillers*). Even though the vocabulary is within limits (approx. 2000 words), the total number of combinations is almost innumerable. Adding a new football player to the vocabulary would necessitate the recording of a large set of new sentences in which this player can occur. Therefore, for GoalGetter, using prerecorded speech is not a feasible method.

2.2 Speech synthesis

An alternative that yields a maximum degree of flexibility is the use of synthetic speech. This method requires much less memory than stored-waveform techniques. One way of producing synthetic speech is by *allophone* or *formant synthesis* which attempts to approximate the acoustic output of a speaker. In the DYD system the DECTALK formant synthesizer was used (Allen, Hunnicutt, and Klatt (1987) discusses its predecessor MITalk). It models the vocal tract transfer function by simulating formant frequencies, bandwidths and amplitudes. The process is controlled by 20 - 40 parameters which are updated every 5 - 10 ms. For this approach, extensive knowledge is needed on how the acoustic properties of the speech signal evolve over time. The parameters are highly correlated with production and propagation of sound in the oral tract. Various sorts of voices can be generated, as well as different speaking styles, speaking rates, etc. One of the drawbacks of this approach is that the automatic technique of specifying parameters is still unsatisfactory. The majority of parameters has to be optimized manually.

Current speech synthesizers usually produce speech by means of *diphone synthesis*. A diphone database consists of small segments excised from human speech, that cover the transitions between any two sounds of a given language. The manual preparation of the appropriate speech segments can be time-consuming, but once the inventory is constructed, there is only moderate computational power needed. Diphone concatenation is less flexible than formant synthesis, since only one voice can be synthesized. When a different voice is needed, a new diphone database has to be constructed.

Intelligibility of synthetic speech can be quite high. Diphone synthesis usually has a higher intelligibility rate than formant synthesis. However, recent evaluations show that when both types of synthetic speech are sent through a telephone channel, intelligibility decreases significantly. In GSM conditions, intelligibility drops even further (Rietveld, Kerkhoff, Emons, Meijer, Sanderman, and Sluijter 1997). Furthermore, naturalness still leaves a great deal to be desired. This leads to the conclusion that speech synthesis is not yet suitable for use in commercial applications.

Diphone synthesis has been implemented as one of the output modes in GoalGetter in order to test the prosody rules in the LGM. Because the LGM generates an orthographic representation with a unique phonetic representation[1], it is possible to do errorless grapheme-to-phoneme conversion by lexical lookup instead of rules. The phonetics-to-speech system SPENGI (SPeech synthesis ENGIne), developed at IPO, provides GoalGetter with PSOLA-based diphones (Pitch Synchronous Overlap and Add, Charpentier and Moulines (1989)). However, the prosodic and durational realization rules in SPENGI have not been optimized for the GoalGetter domain. In the rest of this paper, we focus on another output mode, namely that of speech concatenation.

2.3 Speech concatenation

The key to generating high quality speech output is to find a balance in the tradeoff between naturalness and flexibility. In that respect, concatenating prerecorded units like words and phrases appears to be a good alternative. With this approach, a large number of utterances can be pronounced on the basis of a limited number of prerecorded words and phrases, saving memory space and increasing flexibility. This technique is practical only if the application domain is limited and remains rather stable. Speech concatenation is used in most voice response services, but often the method is so straightforward, that it is not even mentioned in publica-

[1]It could also generate a phonetic representation directly.

tions. The necessary words and phrases are simply recorded and the concatenated sentences are played back when required. This approach has two major problems:

1. Very careful control of the recordings is needed. Usually, this is not accounted for, so that differences in loudness, rhythm and pitch patterns occur, leading to disfluencies in the speech. Phrases seem to overlap in time, creating the impression that several speakers are talking at the same time, at different locations in the room. These prosodic imperfections are often disguised by inserting pauses, which are clearly audible and make the speech sound less natural. As far as the differences in loudness are concerned, these can be remedied by manipulating the overall energy of the material after recording without loss in quality. Differences in rhythm and pitch patterns are more difficult to correct. PSOLA manipulation only works for some voices without deterioration of the speech quality.

2. The words that serve as slot fillers are recorded in one prosodically neutral version only. This makes it practically impossible to exploit the two most important features of intonation:

   (a) Highlighting *informational* structure by means of accentuation, i.e. by accenting important and new information, while deaccenting old or given information.

   (b) Highlighting *linguistic* structure by means of prosodic phrasing, i.e. by melodically marking certain syntactic boundaries and by using pauses at the appropriate places.

One simple application that takes the prosodic properties into account is a telephone number announcement system described in Waterworth (1983). In order to increase the naturalness of the long number strings, they are split into smaller chunks. Digits are recorded in three versions with different intonation contours. There is a *neutral form*, a *terminator*, with a falling pitch contour, and a *continuant*, with a generally rising pitch. Experiments showed that people preferred this method over the simple concatenation method.

Another application called Appeal, which is a computer-assisted language learning program, uses a more sophisticated form of word concatenation to deal with prosodic variations (de Pijper 1997). The words have been recorded embedded in carrier sentences to do justice to the fact that words are shorter and often more reduced when spoken in context. The duration and pitch of the words are adapted to the context using the PSOLA technique. This ensures a natural prosody, but the coding scheme may deteriorate the quality of the output speech to some extent.

3 Speech output generation in GoalGetter

Our approach to concatenating words and phrases requires no manipulation or coding of the recordings, so the quality of the speech is not affected at that point. A good speech output quality is obtained by recording several prosodic variants of otherwise identical phrases and words. In this way, a large number of utterances can be pronounced on the basis of a limited number of prerecorded phrases, saving memory space and increasing flexibility. This technique can be used whenever there is a carrier-and-slot situation, i.e., there is a limited number of types of utterances (carriers, templates) to be pronounced, with variable information to be inserted in fixed positions (slots) in those utterances. GoalGetter obviously fits this situation well. The carriers are the syntactic templates, and these have slots for variable information, such as match results, football team names, names of individual players, and so on.

To determine which words and phrases have to be recorded and how many different prosodic realizations are needed, a thorough analysis of the material to be generated is a necessary phase in the development of a phrase database.

3.1 Prosodic markers

As mentioned before the intonation of a sentence should serve to highlight informational and linguistic structure. In order to generate the proper pitch contour for a given sentence, one needs to integrate intonational, accentual and surface-syntactic information. The LGM has this information readily available and passes it on to the SGM in the form of prosodic markers. There are two basic types of markers: accent markers and phrase boundary markers. In GoalGetter, there are also special, application-specific, markers.

- *Accent markers:* A word can be either accented or unaccented. In the enriched text, accents are indicated with a double quote (") before the accented word. Deaccentuation rules are based on the given-new distinction (van Deemter 1994). As mentioned before, proper accentuation highlights informational structure. Deaccentuation is necessary in GoalGetter because accentuating given information leads to unnatural results and can even result in unintended interpretations. Recently, a third type of accent, viz. *contrast accent*, has been implemented in the LGM. However, the prosodic realizations associated with this type of accent have not yet been included in the SGM. Therefore, we leave this accent type out of consideration in this paper. The interested reader is referred to Theune (1996) (this volume) for a discussion on the prediction of contrastive accent in data-to-speech generation systems.

- *Phrase boundary markers:* Prosodic boundaries are indicated by slashes in the enriched text. The number of slashes (1, 2 or 3) denotes the strength of the boundary. The sentence final boundary (///) is the strongest one. Words which are clause-final or which precede a punctuation mark other than a comma are followed by a major phrase boundary (//). A minor boundary (/) precedes a comma and constituents to the left of an I', C' or maximal projection. This is a slightly modified version of a structural condition proposed by Dirksen and Quené (1993).

  In longer texts, containing more complicated constructions, one might want to distinguish more levels. Sanderman (1996) uses five levels for generating texts with more natural phrasing.

- *Special markers:* The symbols % and @ are used to trigger particular application-specific prosodic realizations not immediately related to accentuation and boundary marking. They are only used in the phrase concatenation mode. In order to use them in the diphone mode we need robust rules that specify how these special prosodic versions are realized, which are unavailable at the moment. The @-sign is used to mark the numbers reflecting the score. This is because in Dutch the score of a match is pronounced in a special way: the two accented numbers are realized with a so-called flat hat (a steep rise on the first accented word and a steep fall on the second one, with high pitch in between), which in Dutch is normally used only if there is no intervening boundary. The fact that the first accented word is lengthened and that a small pause seems appropriate, on the other hand, suggests that a boundary should be there.

  The %-sign is used to mark nouns that are followed by a noun phrase functioning as an adjunct, as in *de %verdediger de Boer kreeg een gele kaart* 'the defender de Boer got a yellow card'. The noun *de verdediger* can also occur in isolation where it has a longer duration and often receives an accent. In the case where it is marked with a %-sign, a different prosodic variant is chosen which is shorter and does not have an accent. This phenomenon seems to be general in Dutch and as such ought to be incorporated in the prosody rules.

3.2 Prosodic realization

Once the content and prosodic properties of the text is known, a phrase database can be developed, which provides the words and phrases that have to be concatenated. For the slot fillers, we chose to use six different prosodic realizations, one for each context described in terms of accentuation and phrasing attributes. Stylizations of these prosodic realizations are depicted in Figure 4. The special markers are not indicated in Figure 4, because they apply to a small group of words only.

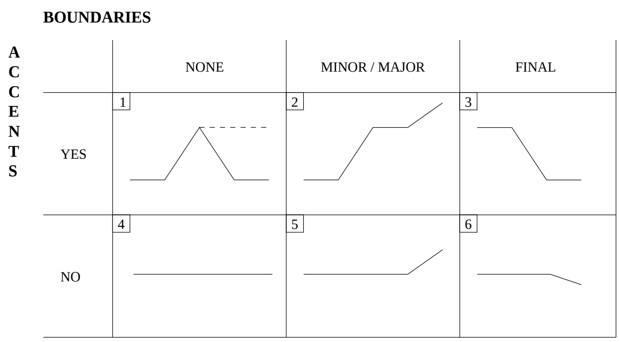


Figure 4: Stylized examples of the pitch contours needed

The six different prosodic realizations, described in terms of the IPO Grammar of Intonation ('t Hart et al. 1990), are:

1. A slot filler that is accented and does not occur before a phrase boundary is produced with the pitch movement that is most frequently used, the so-called *hat pattern*, which consists of an accent-lending rise and fall on the same syllable. This contour often corresponds to the prosodically neutral version that is used in straightforward concatenation techniques. Sometimes, the penultimate and the final accent in a sentence are combined, and instead of two hat patterns, one *flat hat* is realized. In Figure 4 this contour is obtained by combining the rise of (1) with the fall of (3). GoalGetter uses this construction mainly in time expressions that occur at the end of the sentence.

2. An accented slot filler which occurs before a minor or a major phrase boundary is most often produced with a rise to mark the accent and an additional continuation rise to signal that there is a non-final phrase boundary. A short pause is added after the word.

3. An accented slot filler which occurs in final position receives a final fall. A longer pause follows the word. This contour co-occurs with a rise in a preceding word.

4. Unaccented slot fillers are pronounced in a neutral fashion without any pitch movement associated to them.

5. Unaccented slot fillers occurring before a minor or a major phrase boundary only receive a small continuation rise. This type of words does not occur very often in the GoalGetter domain, since the LGM usually puts a minor or major phrase boundary immediately after an accented constituent.

6. Unaccented slot fillers in a final position are produced with final lowering.

When recording the material for the phrase database, the slots in the carrier sentences are filled with dummy words, so that the fixed phrases to be stored in the database can be excised easily. The slot fillers such as team and player names are embedded in dummy sentences that provide the right prosodic context. The sentences are constructed in such a way as to make the speaker produce the standard prosodic realization naturally. The intonation in the fixed phrases is not very critical, so the speaker may use his own intuitions to determine how to pronounce them.

3.3 Generating speech

In order to make a text audible, the proper words and phrases have to be concatenated by an algorithm which performs a mapping between the enriched text (with accentuation and phrasing markers), and the phrases that have to be selected. The different prosodic variants are selected on the basis of the prosodic markers. The algorithm recursively looks for the largest phrases to concatenate into sentences.

At concatenation time, the slot fillers are surrounded by short pauses of 50 ms, which are hardly perceivable, but which give the speech a less hasty character. Because the slot fillers usually contain the important information, they are supposed to stand out slightly from the rest of the sentence, which is an additional reason why introducing small pauses is not disturbing.

3.4 Selection of speaker and speaking style

The choice of an appropriate speaker is essential for the success of the application. Cox and Cooper (1981) conducted a survey to find out what properties in a human's voice make it suitable for use in a telephone information system. The results showed two important factors influencing the preferences of the listeners, i.e. agreeableness and assertiveness (which is also associated to the notion of self-confidence). In their experiments, female speakers were marked up for assertiveness whereas male speakers were marked down for that quality. Because of this property, there seemed to be a slight preference to use a female speaker in telephone announcement systems.

Speaking style also constributes to the output quality of the speech. Two important factors associated to speaking style are speaking rate and pitch range. When selecting a speaker, these factors have to be taken into account. A speaker should not speak too fast, since that gives the concatenated speech a restless, nervous quality. Especially small words like function words will sound as if they have been cut off abruptly. A speaker's pitch range should not be too excessive, as disfluencies in the speech are more likely to occur.

4 Conclusion

This paper describes a method for speech generation in the GoalGetter system. It has been demonstrated that with a sophisticated phrase concatenation technique, we can obtain speech output with a very good quality. As mentioned before, this technique is only suitable when there is a stable and fairly limited application domain. Once the language generation module generates too flexible output and the slot fillers change continuously, the phrase concatenation technique will prove to be too inflexible. Therefore, we are continuing our efforts to improve the diphone synthesis technique.

References

Allen, J., M. Hunnicutt, and D. Klatt (1987). *From Text to Speech: the MITalk System.* Cambridge: Cambridge University Press.

Charpentier, F. and E. Moulines (1989). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Proceedings EUROSPEECH'89, Paris, France*, Volume 2, pp. 13–19.

Cox, A. and M. Cooper (1981). Selecting a voice for a specified task: the example of telephone announcements. *Language and Speech 24*, 233–243.

de Pijper, J. (1997). High quality message-to-speech generation in a practical application. In J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (Eds.), *Progress in Speech Synthesis*, pp. 575–586. New York: Springer-Verlag.

Dirksen, A. (1992). Accenting and deaccenting: A declarative approach. In *Proceedings of COLING 1992, Nantes, France*, pp. 865–869.

Dirksen, A. and H. Quené (1993). Prosodic analysis: The next generation. In van Heuven and Pols (Eds.), *Analysis and Synthesis of Speech: Strategic Research Towards High-Quality Text-to-Speech Generation*, pp. 131–144. Berlin - New York: Mouton de Gruyter.

Klabbers, E., J. Odijk, J. de Pijper, and M. Theune (1996). GoalGetter: From Teletext to speech. In *IPO Annual Progress Report*, Volume 31, pp. 66–75.

Rietveld, T., J. Kerkhoff, M. Emons, E. Meijer, A. Sanderman, and A. Sluijter (1997). Evaluation of speech synthesis systems for Dutch in telecommunication applications in GSM and PSTN networks. To appear in Proceedings of EUROSPEECH'97, Rhodes, Greece.

Sanderman, A. (1996). *Prosodic Phrasing: production, perception, acceptability and comprehension.* Ph. D. thesis, Eindhoven University, Eindhoven.

't Hart, J., R. Collier, and A. Cohen (1990). *A Perceptual Study of Intonation: an Experimental Phonetic Approach to Speech Melody.* Cambridge: Cambridge University Press.

Theune, M. (1996). Goalgetter: Predicting contrastive accent in data-to-speech generation. In J. Landsbergen, J. Odijk, K. van Deemter, and G. Veldhuijzen van Zanten (Eds.), *Proceedings CLIN VII*, Eindhoven.

van Deemter, K. (1994). What's new? A semantic perspective on sentence accent. *Journal of Semantics 11*, 1–31.

van Deemter, K., J. Landsbergen, R. Leermakers, and J. Odijk (1994). Generation of spoken monologues by means of templates. In *Proceedings of TWLT 8*, Twente, pp. 87–96. Twente University.

Waterworth, J. (1983). Effect of intonation form and pause durations of automatic telephone number announcements on subjective preference and memory performance. *Applied Ergonomics 14*(1), 39–42.

Waterworth, J. (1984). Interaction with machines by voice: a telecommunications perspective. *Behaviour and Information Technology 3*(2), 163–177.