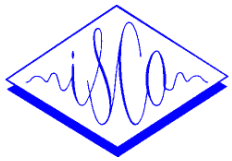

ETRW on Dialogue and Prosody
Veldhoven, The Netherlands
September 1-3, 1999

PROSODY AND PROMPT DESIGN IN A COMPUTER DIALOG SYSTEM

Gayle Ayers Elam and Sarah C. Wayland

Entropic, Inc.

ABSTRACT

This paper describes a technique for creating a small set of recorded number phrases that can be concatenated in such a way as to speak numbers ranging in value from 0.00 to 999,999,999,999.99. This was accomplished by controlling the prosody of both the number phrases themselves and the carrier phrases into which the numbers are integrated. We have successfully implemented this system for both German and English speakers.

1. INTRODUCTION

In creating a man-machine dialog system, it is important to use prompts that are both informative and intelligible. Part of being informative lies in communicating with the user about things that do not necessarily stay constant. For example, a dialog we have been developing at Entropic gives users information about bank account balances, current stock prices, and other stock information. In order to be informative, our dialog must be able to say very small numbers (e.g., 0.01), very large numbers (e.g., 999,999,999,999.99), and any size number in-between (e.g., 3,478.42). In addition, these numbers appear in a variety of different contexts (account balances, stock prices, percentages, etc.).

An obvious solution for creating intelligible spoken numbers would be to record a speaker saying every number between 0.00 and 999,999,999,999.99. However, this would require our speaker to record over 99 trillion different phrases, a task that is not only unpleasant, but untenable. Another solution would be to use a text-to-speech engine to report phrases that contain these numbers. Unfortunately, many of our customers find it difficult to listen to the speech produced by our text-to-speech engine (TrueTalk), despite the fact that its intelligibility is quite highly rated [1]. What we needed was a way to create variable number phrases that were as intelligible as human speech.

One way to do this is to have a speaker record a set of number phrases that can be concatenated to create a longer number. For example, the number 3,478.42 is composed of several number phrases – “three”, “thousand”, “four”, “hundred”, “seventy-eight”, “point”, and “forty-two”. By recording the individual phrases, and then concatenating them, we can reduce the number of phrases that must be recorded from 99 trillion to 105 (0-99, hundred, thousand, million, billion, point). This is the smallest number of recorded units that you could use to create these numbers.

We successfully implemented a German banking system using a variation of this minimizing solution. We used 454 recorded units to build monetary amounts from 0 Franken to 10,999,999.99 Franken, as described in more detail below. In short, the additional units that we recorded in this system were to account for variations in prosodic context, especially for the numbers 1 to 99. For example, consider 99 – *neunundneunzig*. There are prosodic differences in 99 when it occurs in the decimal portion of the number (*X Franken neunundneunzig* “X Francs ninety-nine”), in the number immediately before *Franken* (*neunundneunzig Franken* “ninety-nine Francs”), and in the number before *tausend* “thousand” (*neunundneunzig tausend Franken* “ninety-nine thousand Francs”).

While this approach generated reasonably intelligible numbers, the durations and pitch patterns of the recorded words were not necessarily appropriate for their positions in the large number phrases, and we felt we could do better. In addition, we needed to be able to build even larger numbers, and extending a system using this solution was unwieldly. One solution to this problem is to concatenate the smaller phrases, digitize them, derive the fundamental frequency, and then resynthesize the final number phrase with a pitch contour appropriate to the resulting (large number) phrase [2].

Our American English solution requires the speaker to record 227 phrases, and eliminates the need for the post-processing and resynthesis described above. We used 227 recorded units to build monetary amounts from $0 to $999,999,999,999.99, as described in more detail below. By determining the prosodic characteristics of number phrases in different positions in a large number, we were able to generate extremely intelligible numbers from this set of pre-recorded number phrases. We were able to integrate these numbers into a variety of carrier phrases by carefully controlling the prosody of the surrounding context.

2. METHOD

We used the following steps to create our basic recorded units and to build the system prompt sentences by concatenating the recorded units. The carrier phrase is the unchanging part of the sentence prompt. The variable items are the monetary amounts and numbers in the examples we present here.

1. Determine the exact wording of the prompts required for the system. This includes both the carrier phrases and the variable items.
2. Determine a grammatical prosody to speak the text of the prompt sentences.
3. Determine the minimum number of versions of the variable items that need to be recorded.
4. Craft the carrier phrases to concatenate well with the variable items.
5. Record the “golden voice” reading the full set of sentences that gives you complete coverage

of the carrier phrases and all required versions of the variable items.

6. Create a database of basic recording units by excising segments from the recordings.

7. Concatenate the recording units to build the system prompt sentences.

3. SYSTEM ONE: GERMAN MONETARY AMOUNTS

In this system, prompts were recorded by a native speaker of German. Segments of the recordings were excised and then concatenated to create new prompts. Using basic prosodic principles of phrasing and emphasis, we created a set of recorded items from which a much larger set of natural sounding prompts could be generated.

The first step was to determine the exact wording of the prompt sentences required for the system. This list of required system prompts gave us all the sentence contexts in which monetary amounts occurred.

The second step was to determine a grammatical prosody to speak the text of the sentences. We consulted with our native speaker of German to determine acceptable phrasing and emphasis. Whenever possible, we isolated variable items into their own prosodic phrase. This was to minimize the number of contexts that the variable items occurred in, and hence to help us with our third step: determining the minimum number of versions of the number phrases we needed to record.

We decided that we needed the recording units described in Table 1 below, and that we needed two versions of every number building block. The number phrase items were recorded in two prosodic contexts: before the word *Franken* "Francs" and at the end of an intonational phrase (with no following *Franken*). Each of the recorded items was pronounced as a separate prosodic phrase, beginning and ending with low pitch, and with high tone pitch accents on accented words. [3]

Description of Recorded Items	**Total Items**	
	Franken	**-**
1-10 Million(en) ("million")	10	10
100-900 tausend ("thousand")	9	9
1-99 tausend ("thousand")	99	99
100-900	9	9
0-99	100	100

Table 1: A total of 454 recorded items were used in the German system of building numbers.

For example, this system builds the number 3,478.42 Franken by concatenating the following four recorded items:

- drei tausend (3000)
- vier hundert (400)
- achtundsiebzig Franken (78 Franken)
- zweiundvierzig (42)

The source segments for this utterance came from the following sentences in the recording session, with the extracted segment underlined. *Ihr Sparkonto zeigt* means "Your savings account shows".

- Ihr Sparkonto zeigt drei tausend komma neun sieben Franken. (3000.97 Fr)
- Ihr Sparkonto zeigt vier hundert komma neun fuenf Franken. (400.95 Fr)
- Ihr Sparkonto zeigt achtundsiebzig Franken siebzig. (78.70 Fr)
- Ihr Sparkonto zeigt einundfuenfzig Franken zweiundvierzig. (51.42 Fr)

Note that there are two different ways of reporting the monetary amount, one version with "# komma # # Franken" and another with "# Franken #". (*komma* means "comma" or "point" and indicates the decimal point.) Conveniently, the *komma* version is pronounced with the whole number in a prosodic phrase ending before *komma* and lets us excise the number without *Franken* following. The other version is pronounced with *Franken* as the last word in the number phrase of the preceding number, with a new prosodic phrase for the decimal portion of the number. This version lets us excise the number in a prosodic phrase with *Franken*.

Bigger numbers are built up similarly, as these examples for 9,800,000 Fr and 9,899,000.99 Fr illustrate.

9,800,000 Fr:

- neun Millionen (9,000,000)
- acht hundert tausend Franken (800,000 Fr)

9,899,000.99 Fr:

- neun Millionen (9,000,000)
- acht hundert (800)
- neunundneunzig tausend Franken (99,000 Fr)
- neunundneunzig (99)

Consider the contexts for *neunundneunzig* (99) in the example above. When it occurs in the decimal portion of the number (*X Franken neunundneunzig* "X Francs 99"), it is a phrase all by itself. When it is in the number before *tausend* "thousand" (*neunundneunzig tausend Franken* "99 thousand Francs"), it is part of the phrase which also includes the words *tausend* and *Franken*.

One place where this system produces phrases with less than ideal prosody is the *acht hundert neunundneunzig tausend Franken* portion of 9,899,000.99 Fr. Ideally, all of these words would be pronounced as part of the same prosodic phrase, rather than having *acht hundert* in its own prosodic phrase. This is a shortcoming that we addressed in the English system.

As mentioned in the Method section, the fourth step in the process is to craft the carrier phrases to concatenate well with the variable items. The statement/confirmation question pair below illustrates this. The same number amount as our first example, 3,478.42 Franken, is used.

(a) Ihr Sparkonto zeigt [drei tausend] [vier hundert] [achtundsiebzig Franken] [zweiundvierzig].

(b) Moechten Sie [drei tausend] [vier hundert] [achtundsiebzig Franken] [zweiundvierzig] ueberweisen?

Both *Ihr Sparkonto zeigt* ("Your savings account shows") and *Moechten Sie* ("Would you like") were spoken as separate prosodic phrases that are prosodically marked as continuations. Both phrases end low, and the final vowels are elongated. Utterances (a) and (b) are prosodically identical at the end of the number phrase, yet (a) is a statement and (b) is a question. The entire intonational marking for the question is carried by *ueberweisen* ("to transfer") and its high rising pitch. Even though the number phrase is not recorded with continuation prosody, the question sounds quite natural, and therein lives the art of crafting carrier phrases.

Most things about the performance of this system pleased us, but there were two major things that we wanted to improve. We wanted to improve the prosodic phrasing of number units like 899,000 such that both the hundreds and tens of thousands were part of the same prosodic phrase. We also wanted to decrease our dependence on recording the 1-99 units in multiple contexts.

4. SYSTEM TWO: AMERICAN ENGLISH NUMBERS

The first version of the American English system used number phrases much like the German system. This system included no count units ("dollars", "cents", etc.), and the numbers were relatively small, with maximum values in the thousands. Typical examples of the system sentence prompts are the three below.

- Excite is up [seventeen] at [one hundred] [twenty-six].
- Today's low: [seventy].
- Today's high: [seventy-five].

The next version of the American English system included much larger numbers, into the millions and billions, and also had numbers in many more sentence contexts.

After we had done the first step and determined the exact wording of the prompt sentences required for the system, we were especially concerned about how many count unit words and phrases there were following the numbers. We could not record all the number items together in the same phrase with all the words and phrases the system required, for example "dollars", "cents", "percent", "percent change", and "shares". We also wanted to improve the quality of the big number phrases, without relying on recording units of 1-999 thousand, 1-999 million, and 1-999 billion, which was clearly untenable.

In the second step, determining a grammatical prosody to speak the text of the sentence prompts, we isolated variable items into their own prosodic phrase when possible, as in the German system. The major change in the analysis of the numbers in this system, however, was to recognize and formalize the need for recording units that were not complete prosodic phrases unto themselves. It was this decision to have basic building blocks of numbers that were internal to a prosodic phrase that allowed us to minimize the number of versions of the number phrases we needed to record. Nonetheless, we were concerned that concatenating recording units that were not complete prosodic phrases might result in less smooth sounding output, but that was a risk that we had to take. Fortunately, our preliminary attempts were encouraging.

We decided that we needed the recording units described in Table 2 below, and that we needed two or three versions of every number building block. The number phrase items were recorded in three prosodic contexts: phrase internal, at the end of an intermediate phrase, and at the end of an intonational phrase. The "final" units were recorded items pronounced as a separate prosodic phrase that was the final element in the sentence, beginning and ending with low pitch, and with high tone pitch accents on accented words. The "intermediate" units were recorded items pronounced as a separate prosodic phrase, but not sentence final. Like the "final" units, they begin and end with low pitch and have high tone pitch accents on accented words, but the end of the phrase has a bit of continuation marking to it. In the final system, only the words "thousand", "million", and "billion" were "intermediate" units. The "internal" units are not the final words in a phrase, but the phrase that they belong to meets the same description.

Description of Recorded Items	Total Items		
	Final	Inter-mediate	Internal
billion	1	1	1
million	1	1	1
thousand	1	1	1
100-900	9	0	9
0-99	100	0	100

Table 2: A total of 227 recorded items were used in the American English system of building numbers.

For example, this system builds the monetary amount $3,478.42 by concatenating the following seven recorded items.

- three (internal)
- thousand (intermediate)
- four hundred (internal)
- seventy-eight (internal)
- dollars and
- forty-two (internal)

- cents

The source segments for this utterance came from the following sentences in the recording session, with the extracted segment underlined.

- The Nasdaq is up one hundred three dollars.
- The Nasdaq is up twentyeight thousand two hundred.
- The Nasdaq is up four hundred dollars.
- The Nasdaq is up one hundred seventy-eight dollars.
- Your shares of Excite are valued at twenty-eight dollars and twenty-eight cents.
- The Nasdaq is up one hundred forty-two dollars.
- Your shares of Excite are valued at twenty-eight dollars and twenty-eight cents.

As you can see, the reanalysis of the number phrases for the big numbers (specifically changing the prosodic phrasing of the hundreds units) changed the basic building blocks. This means that the hundreds units ("one hundred") of the system prompt "Excite is up [seventeen] at [one hundred] [twenty-six]" were different in the first and the final version of the system. In the first version of the system "one hundred" was "intermediate" like the "thousand", "million", and "billion" of the expanded system, whereas in the final system "one hundred" was "internal". This change results in less emphasis and disjuncture between the hundreds and tens units.

Bigger numbers are built up similarly, as this example for $9,899,000.99 illustrates.

- nine (internal)
- million (intermediate)
- eight hundred (internal)
- ninety-nine (internal)
- thousand (internal)
- dollars and
- ninety-nine (internal)
- cents

Both occurrences of "ninety-nine" are "internal" in this example. The first one is a component part of the prosodic phrase containing "eight hundred ninety-nine thousand dollars". The second one is a component part of the prosodic phrase containing "ninety-nine cents".

This change in phrasing to allow "internal" number units addresses one of our desired improvements – to improve the quality of the big number phrases, without relying on recording units of 1-999 thousand, 1-999 million, and 1-999 billion. This is also the solution to our need to fit the numbers into all the phrase contexts that the system required (e.g. "dollars", "cents", "percent", "percent change", and "shares"). In the expanded American English system the "internal" version of the number is available to concatenate with different following counting unit phrases, such as "dollars" and "cents", and it ensures that the prosodic phrasing is handled well. In contrast, in the German system the word *Franken* is an essential part of the numbers building blocks and in essence takes care of most of the phrasing issues.

In summary, the "internal" version of "ninety-nine" is used in all of the system sentence prompts below, except for the first two which use the "final" version of "ninety-nine".

- Excite is up two at ninety-nine. (final)
- Today's high: ninety-nine. (final)
- Your shares of IBM are valued at ninety-nine cents. ($0.99)
- Your shares of IBM are valued at ninety-nine dollars. ($99.00)
- Your shares of IBM are valued at ninety-nine thousand dollars. ($99,000.00)
- Your shares of IBM are valued at eight million ninety-nine thousand dollars. ($8,099,000)
- This is a ninety-nine percent change on a volume of ninety-nine shares.
- Today your portfolio lost two thousand ninety-nine dollars, or ninety-nine percent. ($2099 or 99%)

5. DISCUSSION

By doing a careful prosodic analysis, we were able to generate a large set of highly-intelligible numbers from a limited set of pre-recorded number phrases in both German and English. In the American English system, we changed the quantity and prosodic characteristics of the pre-recorded phrases, and we found that we could create numbers that sounded even more natural when the phrase-internal or phrase-final version of the numbers was employed in the correct context. Our technique has the advantage of an economy of units, saving time in the recording process, and simplifying the collection of pre-recorded phrases.

While our numbers sound quite natural without any post-processing, future work might involve controlling the prosody of the recording phrases less stringently, and applying an appropriate PSOLA (Pitch Synchronous Overlap and Add) technique [4] to manipulate pitch and duration of the resulting number phrases for even higher-quality output. This technique could be used to smooth any fundamental frequency discontinuities between the concatenated units. This could prove especially helpful when integrating number phrases into new carrier-phrase contexts.

The proof of our technique is in the “hearing”, and we have not been able to submit our “numbers” to a formal evaluation. However, informal reactions from users of our systems indicate that the numbers created by concatenating prosodically controlled number phrases sounds quite natural.

6. REFERENCES

1. Basson, S., Yashchin, D., Kalyanswamy, A., and Silverman, K. “Comparing Synthesizers for Name and Address Provision: Field Trial Results,” *Proceedings of Eurospeech, Berlin,* 1993.

2. Pijper, J.R. de “High-quality message-to-speech generation in a practical application,” *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis, Mohawk Mountain House, New Paltz, New York,* 1994.

3. Beckman, M.E. and Elam, G.A. “Guidelines for ToBI Labelling, version 3.0.” Manuscript and accompanying speech materials, Ohio State University. [Obtain by writing to] 1998.

4. Charpentier, F. and Moulines, E. “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Proceedings of Eurospeech ’89, 2:* 13-19, 1989.