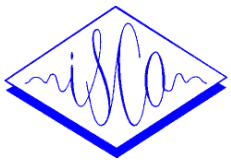

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

Using the ORATOR® Synthesizer for a Public Reverse-Directory Service: Design, Lessons, and Recommendations

Murray F. Spiegel

Bellcore (Bell Communications Research)
445 South Street, Room 2E-252, Morristown, New Jersey, USA 07960
Telephone: 1-201-829-4518; Fax: 1-201-829-5963; Email: spiegel@bellcore.com

ABSTRACT

Many telecommunication applications for text-to-speech synthesis (TTS) involve speaking names and/or addresses. Two services often considered are a caller-name identification service (often called "Audible Caller Name Delivery" or "Who's Calling" in the U.S.). and a reverse directory service. This paper will describe our research, development, and recommendations in providing speech synthesis for a heavily used reverse directory service. Issues involved include highly accurate name pronunciation, high synthesis intelligibility, and required directory preprocessing.

Keywords: reverse directory service, name pronunciation, speech synthesis

1. INTRODUCTION

The technology of text-to-speech synthesis (TTS) has long been "maligned" as a mature technology no longer in need of additional research, yet TTS has been slow to be applied in services for the general public. Although it is assumed that many automated information services are desired, such as (in the telecommunications domain) reverse directory and caller-name identification ("Who's Calling"), few have been offered. One reason is that regulatory concerns about privacy have led to caution on the part of telecommunication companies. An equally high barrier is that synthesis technology is *not* mature enough to withstand being mindlessly dropped into a service – close attention to a myriad of details is still necessary to create a successful public service using TTS.

The first challenge represented by these services is that proper pronunciation of names is particularly difficult, since there are so many (the U.S. has over 1.5 million different surnames). In fact, the listings in telephone directories contain not only the rich mixture of names of English and non-English origin, reflecting the general population, but also a large proportion of common words and neologisms for business names. In our synthesis research we devoted considerable attention to developing rules for name pronunciation that accurately reflect how names are pronounced in the U.S. The paper will describe the strategies used to obtain the highest name pronunciation scores of any synthesizer.

A second challenge is that most telecommunication services are intended for use by a public largely inexperienced with listening to synthetic speech. Synthetic speech is not yet as understandable as natural speech. Furthermore, the message content in these services (names, addresses) contains very little semantic redundancy, which in ordinary text helps alleviate the perceptual errors that occur when listening to synthetic speech. Thus, these services require high segmental intelligibility for success. Our demisyllable-based synthesis provides the high segmental intelligibility necessary for telecommunication applications.

A third challenge is that telephone directories have been organized for purposes that are entirely different from speech synthesis-based services: a) Telephone directories contain many non-standard abbreviations that interfere with message comprehension unless correctly translated; b) The text in directories is often presented in reversed or "scrambled" word order, with the most significant word placed first for indexing purposes. The public-service offering that uses our synthesizer employs special-purpose preprocessors: telephone directory parsers.

2. TECHNOLOGY REQUIREMENTS

The current state of TTS falls somewhat short of mimicking the naturalness and capabilities of human speech. Improvements in TTS naturalness, intelligibility, vocal characteristics, and pronunciation accuracy require more research on speech production, perception, pronunciation, etc. The following sections briefly overview our current efforts to achieve high pronunciation accuracy and segmental intelligibility.

2.1 Name Pronunciation Accuracy

Although many applications for synthesis involve reading text in a specific language (e.g. English), telecommunication applications rely on databases that are largely composed of names. Sensible pronunciations for the names in these databases are critical to many telecommunication services, yet for many reasons high accuracy for name pronunciation is

difficult to achieve:[1] there are millions of names; new names are continually created (especially those of products and business); no dictionary specifies pronunciations for most names; names found in several countries originate from a great many languages; name pronunciation rules are different than for words of a language; and gathering data on name pronunciation (and verifying the accuracy of rules or dictionaries) is not easy. However, through years of careful research, the best systems now approach or may actually exceed average human performance.[2]

There are two general approaches for producing name pronunciations: through rules (either hand-developed or via automatic analogizers) or through dictionaries. Rules offer compactness and, if well written, generalizability for infrequent names. Dictionaries offer ease of construction and, if well verified, high accuracy for names covered by the dictionary. The pronunciation rules used in the ORATOR® synthesizer were developed by referencing very large databases of telephone company listings, representative of names of people, places, and businesses in the U.S. Having access to telephone listings allowed us to validate the synthesizer's pronunciations.

Synthesizers addressing the task of name pronunciation either require the user to select name- or word-pronunciation rules and dictionaries or rely on lexicographical cues (e.g., capitalization) to determine which are needed. However, such cues or selection processes are nearly useless for telephone directories, which are a rich mixture of names and English words. In fact, business listings may be the most widely used database in which names and words are so thoroughly mixed. If residence listings are pronounced using name-pronunciation rules and if business listings are pronounced with English-word pronunciation rules, there will be high error rates for business listings.[1] Thus, for the ORATOR system we developed a *unified* set of rules and dictionaries for both proper names and English words. In business listings, many ambiguities are lexically resolved based on word-pair statistics from large text databases. This relieves the user from the responsibility of deciding which rule set to use, and produces appropriate pronunciations for many applications.

The approach for our pronunciation rules, described more fully elsewhere,[3] involves several stages: a) a text-normalization and acronym-analysis stage, where abbreviations and numbers are expanded, and where it is determined whether acronyms are to be pronounced; b) a dictionary of exceptionally pronounced words and names, including common alternate pronunciations for names; c) an ethnic classification of input words that guides pronunciations (so that a foreign-language name is not pronounced according to the rules of that language, nor is it completely Anglicized); d) an extensive morphological analysis involving nearly 2,000 morphs occurring in names; and finally, e) a unified set of pronunciation rules for names and words.

In developing the rules, it was especially beneficial to have lists of surnames, first names, city and country names, as well as business names. The latter are especially instructive, as small businesses contain inventive neologisms important for dictionaries, rules, and morphological analysis. Our experience suggests that in order to achieve highest name pronunciation accuracy when developing rules or dictionaries, it is important not to rely on any one individual's intuitions nor to pronounce names strictly in accordance to the pronunciation rules of the original language. Instead, surveying representative people and verifying via telephone lists is very important.

One final comment on the topic of name pronunciation for those considering developing pronunciation dictionaries for names (e.g. ONAMASTICA and related projects): it will be wise to include several alternate pronunciations when possible. In synthesis applications users could cycle through alternates for a pronunciation they prefer, rather than being subjected to a more laborious customization process. Alternate pronunciations will also help services that use automatic speech recognition: an automated directory agent needs to match names in the incoming speech that will be pronounced in any one of several ways.

2.2 Synthesis Intelligibility

The measured differences between the clarity or intelligibility of synthetic speech and human speech can be small on simple tests such as the MRT or DRT, but on tests more representative of real applications, synthetic speech does not closely approach human speech.[4] The deleterious effects of synthetic speech distortions can be largely overcome by highly-motivated listeners, with long-term experience with synthesized speech, when listening to texts with adequate semantic redundancy.

Unfortunately, the casual user of telephone services is not (yet) accustomed to synthetic speech, nor do they want to be. Furthermore, the information presented in many of these services (e.g., names) contain very little semantic redundancy. That is, a correctly pronounced, but poorly synthesized, name can be misheard as another name, since nearly any phonologically allowed sound sequence may be a name. Thus, high intelligibility is essential for services involving names. Our approach, concatenative synthesis using a demisyllable inventory, provides state-of-the-art intelligibility. We are currently working to significantly improve the intelligibility of our

---

ORATOR is a registered trademark of Bellcore

1. Using a look-up scheme that removed plurals and other inflectional affixes, roughly 66% of the items (word types) in business listings are not found in an extensive dictionary (i.e., they only occur as names). In residence listings 99% of the items were not found with dictionary look-up, as expected. By contrast, only 1% of the items in a typical issue of the New York Times are not found.

synthesizer through a new inventory encoding scheme.[5]

2.3 Naturalness

The last factor associated with synthetic speech considered here is its naturalness. Naturalness is the most immediately apparent feature of synthetic speech - listeners can usually assess naturalness long before they can estimate the clarity or pronunciation accuracy of synthetic speech. Nevertheless, naturalness is synthetic speech's most elusive quality: attempts to produce a sensible metric for evaluating naturalness have been only partially successful, and research aimed at improving naturalness has made frustratingly slow progress.

A continuing hurdle to the wide use of synthetic speech is lack of naturalness. Many applications for speech technology will not be elective to the ultimate users, but will be thrust upon them. As a result, there will be resistance to some TTS services until the speech is more pleasant to listen to. More applications of synthetic speech will open up as synthetic speech continues to attain higher naturalness.

3. APPLICATION REQUIREMENTS

Services based on the information contained in telephone company databases offer enormous potential. Unfortunately, these databases were designed for services that have quite different needs than those of TTS. Without proper attention to methods for overcoming these differences, services based on these databases will fail as readily as if woefully inadequate speech quality was used. Following is a sampling of the problems that can be encountered. More detail and proposed solutions are contained in [6].

Some databases do not distinguish between types of listings (residence, business, government), which can present several problems. Proper pronunciation of homographs, some of which can be either names or words, is helped with knowledge of whether a listing is a residence or not. For instance, in the U.S. one finds the name Fried vs. the English word "fried," and the Arabic name Said vs. the English word "said." Many cases can be disambiguated via context, but not all. A second problem is that word orderings are often different for residence and business listings. While the residence of *Jane Smith* is probably stored as "Smith Jane", her business may be listed as "Jane Smith Distributors". A third problem is that abbreviations and acronyms can often not be identified as such. Many may look like potentially legitimate names, yet if pronounced as if they were a name might be in error. For example, in the U.S. Bek, Jos, Maj, and Corp occasionally occur as family names, yet BEK is also an acronym (to be pronounced "B E K"), and the rest are frequently used abbreviations for the words Joseph, Major, and Corporation.

Some databases contain no indication of whether a name is a first or last name, or lack such indication in some contexts (e.g. business listings). Cues such as titles and initials, if used in a consistent fashion, may help indicate the correct ordering some of the time. A likelihood estimate could determine the most likely name ordering, but is likely to require memory- and/or compute-intensive processes[2] and contain considerable error.[3] Even when permitted by computer resources, this approach probably should be considered only where re-ordering through other means obtains a high error rate.

A related problem is that customer information may not indicate which portion is the address and which is part of a name. Thus, "Buildings and Grounds" may be part of a name whereas "Main and Pine" is part of the address.[7] This creates problems if field-dependent processing must be applied, or if customers request repetition or spelling of the information in a given field.

Databases may store information in an order designed for rapid retrieval for an existing service, such as placing the most important words first. Often the text ordering is *completely* inappropriate for synthesis. Listings like "Embassy House The Apartments," "Adults Sexually Abused As Children The Center For," and "Bingham Carleton Dr & Carol, Rev," if sent unaltered to a synthesizer, will simply not be understood.

When a database lacks identification of upper vs. lower case, the distinction between acronyms and other words is lost. As mentioned above, many family names share the same letters as acronyms.

Databases often contain additional information that is used for other purposes. Text such as this is found: "Call after 5pm," "TouchTone phone only," "See also [company_name]," or "Delbarton telephone number." If this text is not removed, customers will be annoyed, or worse, thoroughly confused by the extraneous information in the audio message. Due to similarity with information desired in other contexts, some extraneous information cannot be easily identified for automatic removal.

Some databases contain field length restrictions that severely limit the text available for listing information. Customers seeing a visual display "Jonathan Rosenb" or "Jonthn Rosnbrg" will recognize information is missing and might even be able to

2. In the U.S., there are 1.5 million family names and tens of thousands of first names; even approximating the statistics for infrequent names (which constitute a large proportion of all names) is difficult.

3. Although surnames are often used for first names (of the most common 1000 surnames, 640 occur as first names), a very high proportion of first names are also surnames (998 of the most frequent 1000 first names occur as surnames). For example, in the U.S. the name *Ray Charles* is highly likely in either order, since both *Ray* and *Charles* are very common surnames and given names. Names such as these (Warren Christopher, Robert Morris, Dean Martin) are not unusual.

correctly reconstitute the listing. However, those who must listen to synthesis of truncated text will be confused. In addition, clerks providing information in too-small fixed-length fields can produce idiosyncratic, uncommon abbreviations, e.g.: "Best Secrtrl Crp." Again, many of these can be resolved when displayed visually, but not when synthesized.

Even in the absence of field length restrictions, abbreviations are used for the sake of easy data entry. Unless data is entered by highly trained personnel or is screened automatically, listings will contain inconsistently applied abbreviations ("Lincolnshire" abbreviated as *Linclnshire, Linclnshr, Linclnshre,* or *Lincnshr*) and gratuitous abbreviations, leading to highly ambiguous translations ("Comm" standing for *Communications, Commercial, Committee, Common, Commission, Commonwealth, Commerce,* and many other words). These will be misunderstood without proper translation.

4. USER REQUIREMENTS

Different users have different needs; service developers must know the particular needs of their own customers. The amount of familiarity users have with information and its semantic redundancy is a key component to how readily users will comprehend synthesized speech. If users intend to transcribe information, then speaking rates, pauses, and even phrasing may need to be different than if users are merely verifying information. Repetition and intelligent spelling options should be available. One should consider adding phrase boundaries, lengthening pauses, or slowing the speech rate during repetitions. Because automatic preprocessing to correct large, frequently updated databases may never be 100% accurate, information should be available through a back-up mode, e.g. visually or via human operators.

Before customers use a service, they must understand its operation. This can be achieved via a well-tested, intuitive user interface; an analog with existing services; or conventional prompts. When prompts are used, a decision must be made whether to accept synthetic speech for the prompts or to use recorded human speech. No single choice can be appropriate for all conditions. When the prompts are simple or somewhat predictable, synthetic speech may be better for both the prompts and the requested information. Not only may the transition between dissimilar voice qualities prove jarring, but providing a small amount of familiarity with the synthetic voice in advance may improve comprehension. However, if the prompts are complicated, human speech may be more appropriate. Other factors also govern optimal voice selection, such as how frequently a typical customer accesses the service and the complexity of interaction.

5. A PUBLIC REVERSE DIRECTORY SERVICE

A U.S. telephone company (Ameritech) has recently implemented a public-access, reverse directory service using the ORATOR system for synthesis. Approximately 2 - 3 million listings per city are accessed in the service. Non-standard abbreviations are expanded largely via lookup in field-dependent tables. A 1400-instruction preprocessor for database-specific word reordering and phrasing, which attends to most of the problems mentioned in this paper, was developed over the course of one year. Both the abbreviation tables and the directory preprocessor have very high accuracies (above 99%). Yet for these large databases many listings still contain errors because the distribution of errors has a very long tail. Every rule or table entry for fixing the remaining errors corrects only a handful of database entries.

REFERENCES

[1] Spiegel, MF "Pronouncing names automatically," Proc. of AVIOS, p. 107-132, 1985.

[2] Golding, AR and Rosenbloom, PS "A comparison of Anapron with seven other name-pronunciation systems," Journal of AVIOS, 1994. In preparation.

[3] Spiegel, MF and Macchi, MJ "Development of the ORATOR synthesizer for network applications: Name pronunciation accuracy, morphological analysis, customization for business listings, and acronym pronunciation," Proc. of AVIOS, Bethesda, MD, 1990.

[4] Spiegel, MF, Altom, MJ, Macchi, MJ, and Wallace, KL "Comprehensive assessment of the telephone intelligibility of synthesized and natural speech," Speech Communications, *9*, p. 279-292, 1990.

[5] Macchi, MJ, Altom, MJ, Kahn, D, Singhal, S, and Spiegel, MF "Intelligibility as a function of coding method for template-based speech synthesis," Eurospeech '93, 1993.

[6] Spiegel, MF "Coping with telephone directories that were never intended for synthesis applications," Proc. of ESCA workshop on Applications of Speech Technology, Bavaria, Germany, 1993.

[7] Kalyanswamy, A and Silverman, K "Say What? Problems in preprocessing names and addresses for text-to-speech conversion," Proc. of AVIOS, 1991.