A Polyglot Domain Optimised Text-To-Speech System for Railway Station Announcements

Csaba Zainkó, Mátyás Bartalis, Géza Németh, Gábor Olaszy

Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics, Hungary
{zainko,bartalis,nemeth,olaszy}@tmit.bme.hu

10.21437/Interspeech.2015-311

Abstract

Announcements at railway stations are a major information source for passengers. In order to ensure high intelligibility, the traditional solution is to use recorded prompts with "slot filling" of variable data. If a data type (e.g. train name) changes new recordings have to be made. Even with careful design the quality of the system will gradually deteriorate due to change of the voice of the voice talent, speech rate, etc.

Advances in corpus-based technology have allowed the introduction of text-to-speech solutions into this application domain. In this paper our solution for a flexible, single voice based polyglot system is described. It is currently implemented for Hungarian and English with plans underway for German. Hungary, being at the geographic center of Europe is at the crossroads of rail connections to more than 15 countries. The Hungarian system announces the Hungarian variant of station names while the English system shall read them in the official language of the country (e.g. Venice is 'Velence' in Hungarian and 'Venezia' in Italian).

The system has been in operation at the largest passenger railway station of Hungary since June 2014 and has been installed for more than 60 other stations and stops.

Index Terms: polyglot speech synthesis, railway passenger information system, text-to-speech

1. Introduction

Railway information systems have been a subject for speech technology research for a long time. Although it is a limited domain area, it offers a wide variety of challenges. The first large scale, multilingual telephone-based speech dialogue research project in this area was the LE-3 project ARISE (Automatic Railway Information Systems for Europe) between 1996 and 1998 [1]. These systems provided travel information over the telephone with varying detail for the rail networks of France, the Netherlands and Italy. TTS systems of the time could not provide sufficient quality for the task. The traditional approach was prompt concatenation with insertion of variable data recorded in a single, indifferent pronunciation [2]. This solution caused perceivable discontinuities and also required continuous upgrading of the studio recordings. This solution was applied in Hungary until recently also for station announcements. This approach was later enhanced by including several prosodic variations of variable data [3][4].

With the improvement of corpus-based and unit selection approaches it became reasonable to apply these techniques for station announcements in several countries. Some commercial companies focus business units for this purpose (e.g. Acapela [5] and Ivona [6]). Researchers have tried to apply solutions for the task in various languages using both corpus-based and statistical parametric techniques. A unit selection approach is reported for Czech in [7] and [8]. A HMM-based approach was applied for Chinese [9].

There is not too much emphasis in the literature on mixed language related problems of railway announcements. It may be due to the fact that most reports come from large countries where international traffic is not significant. Polyglot synthesis was introduced in Switzerland [10]. This was first implemented as a combination of three separate diphone-based systems derived from a speaker who spoke German, French and Italian at a native level. Recently HMM-based speaker adaptation was tested by phoneme mapping in Japan [11] and India [12]. In our system we needed the natural quality of corpus-based solutions. Our aim was to provide better speech quality with a corpus-based domain-optimized approach than that of the traditional prompt concatenation solution.

In 2007 there was a feasibility study in Hungary on the possible implementation of a corpus-based domain optimized solution for railway announcements [13]. The test location is the railway station of a small town with about 30 trains a day. The announcements are in Hungarian only. The script of the recordings was based on the traditional local announcements of the station extended with generic messages. Altogether 1200 sentences were recorded. The quality and flexibility was sufficient. This system has been in operation since 2008. It was the basis for the system reported in this paper.

2. System requirements

The traditional, former announcement systems of the Hungarian Railways are based on phrase concatenation. The main problems with these solutions are the following:

- Each station has its own database and concatenation rules. It takes several weeks to generate the new dataset in case of timetable changes.
- If a variable data (e.g. train name) changes or a new message type is required (e.g. replacement buses because of track reconstructions), a new recording is required. Voice talents are usually difficult to reach so the generation of the new data may last for a long time.
- Some of the systems are rather old (up to 20 years) and the voice of the voice talent has changed significantly so there are disfluencies in the announcements.
- Depending on regions and languages there are several voice talents (e.g. different voices for Hungarian, English and German).

To overcome the above problems Hungarian Railways intended to introduce TTS technology with the following basic requirements which were targeted by our solution:

- Similar or better speech quality for domain specific announcements with the Hungarian TTS system than with the traditional solution. Hungarian accent is accepted for foreign languages.
- Easy and quick (real-time), text-input based generation of announcements.
- Intelligible speech output even for out-of-domain text input.
- A single voice for all messages and languages.
- English (and later German) announcements for Intercity and international trains.
- Local speech technology and linguistic support.

3. The TTS System

The announcement system is composed of three major subsystems (see Figure 1): (i) he *controller* subsystem through which the operator can issue commands for announcement generation based on the timetable and announcement information database input by a railway officer, (ii) the public address system (denoted by *PA system* in Figure 1) containing the audio cabling, amplifiers and loudspeakers, (iii) our *TTS system*. See TTS details in the sections below.

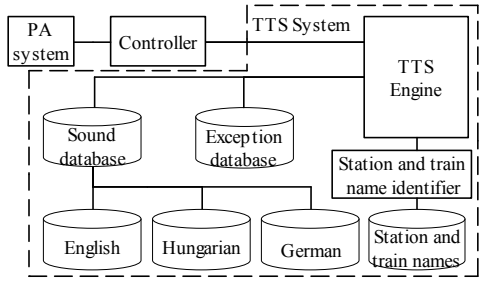

Figure 1: *Block diagram of a large station installation*

3.1. Script design

Script design is a crucial issue in corpus-based speech synthesis. Both domain and unit coverage of the language have to contain all the variants that could be required for optimal reproduction of speech. Our initial text corpus was the list from the feasibility study extended by additions derived by a greedy algorithm from the traditional announcement list of some large railway stations. The general database has eight types of smaller blocks: station names, station names with suffix(es), train names, messages with changing content, messages with fixed content, time elements, platform elements, others. The first Hungarian version of the script was composed of 2410 sentences. To cover most of the country 900 new sentences have been added. The English script contains further 577 sentences for domain coverage and 1133 sentences from ARCTIC [14] for general coverage. Schemes were defined for different message types (arrival, departure, travelling via, etc.). The wording of the messages had to take into account the limited knowledge of English of several foreigners travelling across Hungary (e.g. instead of "*The train calling at Szob...*" we use "*The train stopping at Szob…*")

3.1.1. Station names

For the pronunciation of station names the following specifications were given: in the Hungarian subsystem all foreign station names had be pronounced according to their official pronunciation (e.g. *Villach*) except if they had a Hungarian name too (e.g. the Slovak *Bratislava* in Hungarian pronounced as *Pozsony*). In the English and German announcements the Hungarian station names had to be pronounced in Hungarian, all others according to their official (not English) pronunciation (e.g. *The train arrives from Wien*).

The system can handle and pronounce 2031 Hungarian station names and 732 international ones. For Hungarian, a solution was developed for suffixed station names in order to avoid the recording of all suffixed versions (~14000). The following strategy was implemented: station names (STN) were stored in a separate block (2031 in text and waveform). For station names with suffixes (STNX) a reduced number of variants was constructed taking into consideration the suffix forms and the ending syllables of station names. Some examples for suffixes that were used for the grouping are given below.

Where (*n, on, en, ön, án, én)* (*Budapest**en***)
From where (*ból, ből, ról, ről, áról, éről)* (*Budapest**ről***)
To where (*ba, be, ra, re, ára, ére)* (*Budapest**re***)

Another grouping concerned the final syllable types found in Hungarian station names. From these two groups a minimal set of station names having a suffix were created (1420 vs. 14000) that cover all possible suffixed forms of station names.

During studio recordings station names (normal and suffixed) were read in carrier sentences. In one sentence 6 station names were listed. The beginning and the end of the sentence was fixed, the station names were placed inside the sentence. During the synthesis process the program detects station names in the text. If it is a suffixed one, the algorithm concatenates the appropriate suffix from the suffixed block to the original station name intelligently.

3.1.2. Train names

Altogether 143 train names are handled in the announcement system. These names are stored in separate blocks both in written form and in their spoken carrier sentences. A train name may occur in the message to be synthesized only at the beginning (*PTE Intercity train arrives from Pécs at platform 10.*) or inside (*We inform our passengers that the PTE Intercity train is delayed.*). The continuous change in marketing policy results in the frequent change of train names. In one year 11 new names were added to the timetable and 7 were withdrawn.

3.2. Speech databases

3.2.1. Voice talent selection

Selecting the voice talent was a complex process. A female voice was the target in order to have less echo in large halls. The selected voice had to meet three conflicting criteria, i.e. speech technology, subjective aspects and management ones.

As for speech technology, it was important that the voice matches to our software tools (good articulation, clear voice, normal speaking rate) for the most accurate automatic annotation and segmentation results possible in order to reduce the need for manual corrections.

As for the subjective aspect the voice should be accepted by the public. It had to be pleasant and highly intelligible. On the other hand, the selected voice talent must be relatively young (in order to be able to support further modifications and developments for several years), be experienced in studio recordings and have multilingual competence. The selected person was a well-trained female radio announcer. She is native in Hungarian and Romanian and has some accent for English and German. She read the Slovak, Polish, Czech, Russian etc. station names with good pronunciation.

3.2.2. Studio recordings

We apply a master sentence to force the speaker to keep the same sound timbre, F_0 and speaking style during long recordings (3-4 hours) and at each new occasion of a studio recording. The master sentence is quite long (11 words) containing station names as well. Listing of station names occurs frequently, so prosody planning took it into account. The listed station name must have close to neutral prosody. During the synthesis process when concatenating many station names after one another only a slight change in speech quality is allowed. Close to neutral means no accent, approximately the same F_0 and the same articulation speed. To reach this the master sentence was played over headphones after every 25 read sentences (or if needed). The announcer repeated it aloud while listening to the recorded master version until the sound timbre was acceptable for our speech expert. This approach needs a professional announcer, who can adjust her voice at will and who can keep the adjusted voice and the style (speed, sentence ending, etc.) for longer periods when speaking.

3.2.3. Database processing

Speech processing is semi-automatic. All read sentences are aligned to the text. Sound boundaries are determined by an automatic speech recognizer in forced alignment mode [15] and checked by a human expert. There are several English sentences where the language is mixed (international station names) that is why a new acoustic and language model was applied in the speech recognizer. The Hungarian database contains 8 hours of speech for domain coverage. The English database is about 2 hours long.

3.3. Corpus-based TTS

This new system was based on our solution introduced in [16]. It was extended with methods for handling mixed language sentences.

3.3.1. Special prosodic features

The text processing component consists of two basic parts: general and special domain-related rules. The processing of abbreviations, foreign names is the same as in our general TTS systems [16]-[18]. The handling of domain specific sentences shall deal with some special words with unique prosody. Station names are critical units of announcements, so they have to be highly intelligible. A sentence often contains several of them in a list. There are two contradicting requirements. On the one hand the reading speed of the list should be fast to decrease the length of the sentence. On the other hand there should be pauses to increase intelligibility. This problem was solved by introducing a short pause. Station names are identified in the text and ordered in a list with short pauses. It is also a dictionary-based solution. Some station names are composed of several parts. These names are pronounced without a pause (e.g. 'München Hauptbahnhof'). These names are tagged as connected words. Similar connected expressions are frequent in the names of intercity and international trains. The platform notations, departure and arrival times are also resolved in the text processing phase.

3.3.2. Speech generation

Our corpus based speech synthesizer [16] selects the path with the lowest cost for the sequence of corpus elements (1). There are two components of cost, target (C^t) and concatenation costs (C^c).

$$U = argmin\{\sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=1}^{n+1} C^c(u_{i-1}, u_i)\} \quad (1)$$

The target cost shows the distance between the expected target (t_i) and the possible elements (u_i). The concatenation cost gives the measure of the fluency of element concatenation. The final sequence is given by Viterbi decoding. The search is performed on three levels: sentence, word and phoneme. The search starts on sentence level and it steps down to lower levels, if there are not enough proper candidates (under a cost threshold) at the given level.

3.3.3. Challenges of polyglot synthesis

Our voice talent speaks native Hungarian and Romanian. For the other languages she was adapted to reference pronunciations. Most of the speech database is in Hungarian. As all station names shall be pronounced in the official language of the country in the English system, we have to use several Hungarian names in the English announcements. Both the Hungarian and the English speech database have to be searched in this case. The same phonemes may have different allophone realizations depending on the language. In order to select the right version in a mixed sentence each sound is labeled by a language tag. During text preprocessing the station names and their phonemes are labeled by the respective tag in accordance with the procedures of the database processing stage (c.f. section 3.2.3). Consequently the minimal cost search can be performed on speech databases of different languages.

4. Application environments

There are two basic application scenarios. In large stations (see Figure 1) there is an on-site TTS engine with "hot backup". In case of smaller stations and stops there is a central control station which remotely controls the announcements of several locations (see Figure 2).

4.1. Operating conditions

The application environments are highly variable. There are stops, small stations and the largest Hungarian passenger railway station which is the main international and intercity railway terminal (Budapest-Keleti). The TTS technology is usually installed during station or track reconstruction works. Depending on the rate of reconstructions public address loudspeakers and amplifiers may be replaced by new ones or old sound systems (even 20 years old) are combined with the new TTS output signal.

The load of the TTS engines also depends on the location. It may serve only a single small station where only a couple of trains arrive or depart. In the busiest locations – like railway terminals – the TTS generates the sentences without

interruption in rush hours. Similar load appears at some busy lines, where one TTS engine generates speech for 10-20 smaller stations and stops.

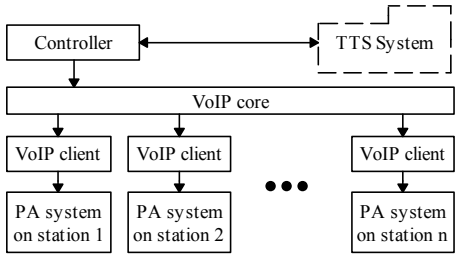

Figure 2: *Block diagram of a remotely controlled installation*

The required quality is also different. It depends on the quality of the stations' sound systems and the speech transmission technology. Typically VoIP technology is used if there is a large distance between the TTS engine and the stations. In this case we use 8 kHz or 16 kHz sampling frequency, which fits the ITU G.711 and G.722 standards. At the stand-alone TTS the sampling frequency is 22 kHz.

After installing a new TTS based announcement system, in most cases the sound system settings are adjusted.

4.1.1. Budapest Keleti (Eastern) railway terminal

This is the largest location where our new TTS system was installed. There are 16 platforms and several waiting halls in it. The main hall dimensions are: 31m height, 188m length, 42m width. After the first settings some adaptation requests came because at daytime use the circumstances changed. The crowd generated noise was measured with specialized noise microphones and the volume was adjusted according to that level. There were some locations where the volume was too low. Most of the complaints were about the system being too loud. For example the cashiers could hear the announcements better than the customers, or the platform staff could not hear the engine driver via the walkie-talkie.

4.2. Text input variations

The source of the synthesized text is just as varying as the environments. The text is determined by the type of traffic (e.g. local, fast, intercity or international trains). There is some independence of regional railway directorates and terminals which increase the variability of texts. For example they use different expressions or sentences for the same event. Unexpected texts were needed when a major terminal was closed for several months because of damage of a train tunnel. The last station before the tunnel was operated as a terminal, and provided extra information (replacement buses, cancelled trains, etc.). The system provided acceptable quality even for strongly out-of-domain texts.

4.3. Speech quality maintenance

The speech quality is maintained by the following process. The systems' logs are checked and new input text and the related speech response is monitored. We also get comments from the staff of the stations and sometimes from passengers. We investigate these comments and create a patch or an upgrade. If it is necessary new sentences are recorded and added to the speech database.

5. Perceptual tests

A perceptual MOS (Mean Opinion Score) test was organized to evaluate the voice quality of the system. An internet based online test was developed. Seven typical sentences produced by the system were evaluated by the test subjects. The following scale was applied: 5 (equal to natural voice quality); 4 (close to natural voice); 3 (well understandable, but machine voice), 2 (robotic voice), 1 (difficult to understand).

The sentences were presented in random order for each subject. They were asked to provide general information, too. The Hungarian version was tested by native Hungarians while the English version was tested by non-native English speakers from 8 countries. The results are presented in Table 1.

Table 1. *Perceptual test results.*

	Hungarian	English
Number of test subjects (**sum**/M/F)	**50**/29/21	**54**/34/20
Age of subjects (**average**/max./min.)	**40**/74/21	**35**/72/22
Test scores (**average**/best/worst)	**4.5**/4.8/4.3	**3.6**/4.1/3.2

These results show that the voice quality of the Hungarian system approaches human performance. The significantly worse – but still highly intelligible – results of the English version may be due to the Hungarian accent of the voice talent, the mixed language of the sentences and the effect of non-native English passengers who mostly do not know the international and especially the Hungarian station names. More detailed studies are required to accurately define the effect of each of these factors.

6. Conclusions

The proposed system could meet the requirements set out in Section 2. It has been in operation at the largest passenger railway station of Hungary since June 2014 and has been installed for more than 60 other stations and stops since then. The greatest difficulty is the proper handling of mixed language sentences. Passengers are satisfied with the solution. Railway operators and officers tend to over-estimate the capabilities of the system and occasionally input extremely long, complicated sentences that may contain out-of-domain elements.

The application domain could be extended to on-board systems and long-distance bus terminals and vehicles. That would require the significant extension of the domain (tourist information, weather, etc.).

This scenario offers a good opportunity for future studies on the perception of mixed language TTS output by non-native listeners in a realistic context. A Hungarian HMM-based general TTS system [18] was adapted on the voice talent database. It offers the possibility of examining quality aspects of a hybrid (corpus-based + HMM) system for special words and texts.

7. Acknowledgements

This research was partly supported be the EITKIC_12-1-2012-0001 project through the Research and Technology Innovation Fund of the National Development Agency of the Hungarian Government, with the contribution of the EIT ICT Labs Hungarian National Associate Node (www.ictlabs.elte.hu).

8. References

[1] E. den Os, L. Boves, L. Lamel, P. Baggia, "Overview of the ARISE project," *Proceedings of Eurospeech '99*, Budapest, Hungary, 1999, pp. 1527—1530

[2] H. Aust, M. Oerder, F..Seide,and V.Steinbiss. "The Philips automatic train timetable informationsystem", *Speech Communication, Vol. 17,* pp. 249-262, 1995

[3] E. Klabbers (1997). "High-quality speech output generation through advanced phrase concatenation", *Proceedings of the COST Workshop on Speech Technology in the Public Telephone Network: Where are we today?,* Rhodes, Greece. 1997, pp. 85-88.

[4] M. Palestri, A. Pacchiotti, S. Quazza, P. L. Salza, S. Sandri, "Choose the Best to Modify the Least: A New Generation Concatenative Synthesis System", *Proceedings. of Eurospeech 99,* Budapest, Hungary, 1999, pp. 2291–2294..

[5] Acapela Transport: Rail, http://www.acapela-group.com /acapela-for-transport/transport/rail/ , 2015

[6] Ivona: Announcement Systems: http://www.ivona.com/en /for-business/announcement-system/ , 2015

[7] J. Švec, L. Šmídl, "Prototype of Czech Spoken Dialog System with Mixed Initiative for Railway Information", *Proceedings of TSD 2010*, Brno, Czech Republic, 2010, pp 568-575

[8] M. Jůzová and D. Tihelka, "Tuning Limited Domain Speech Synthesis Using General TTS System", *Proceedings of TSD 2014*, Czech Republic, 2014, pp. 408–415.

[9] Z. Yu, H. Wu, M. Wu, G. Chen, "Speech Corpus *S*cript Design for TTS System Applied on Railway Passenger Service Information Broadcasting". *Proc. of Oriental COCOSDA 2012.* pp. 97-100.

[10] C. Traber, B. Pfister, "From multilingual to polyglot speech synthesis". *Proceedings of Eurospeech'99*, Budapest, Hungary, pp. 835–838.

[11] J. Lattore, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Communication, vol. 48*, pp. 1227-1242, 2006.

[12] B. Ramani, M.P. Jeeva, P. Vijayalakshmi, T. Nagarajan, " Voice conversion-based multilingual to polyglot speech synthesizer for Indian languages", *Proceedings of TENCON 2013*, 22-25 Oct. 2013, Xi'an, China pp 1 - 4.

[13] Cs. Zainkó, G. Németh, *"*Vasútállomási utastájékoztató (Railway station announcer)*",* In: *A Magyar beszéd (Hungarian Speech),* eds:G. Németh, G. Olaszy, Akadémiai Kiadó, Hungary p. 579. 2010.

[14] J. Kominek and A.W. Black, CMU ARCTIC databases for speech synthesis, *Carnegie Mellon University, Language Technologies Institute Tech Report CMU-LTI-03-177*, 2003

[15] G. Sárosi, B., Tarján, A., Balog, T., Mozsolics, P., Mihajlik, & T. Fegyó, "On modeling non-word events in Large Vocabulary Continuous Speech Recognition", *Proceedings of 3rd IEEE International Conference on Cognitive Infocommunications* (CogInfoCom), Kosice, Slovakia, 2012, pp. 649-653.

[16] M. Fék, P. Pesti, G. Németh, Cs. Zainkó,G. Olaszy,"Corpus-based unit selection TTS for Hungarian," *Proceedings of TSD 2006*, Czech Republic, 2006, pp. 367-373

[17] G. Olaszy, G. Németh, P. Olaszi, G. Kiss, Cs. Zainkó, G. Gordos, "Profivox—A Hungarian Text-to-Speech System for Telecommunications Applications", *International Journal of Speech Technology 11/2000*; 3(3):pp. 201-215. 2000

[18] B. Tóth, G. Németh, "Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis", *Acta Cybernetica-Szeged., Vol., 19.4:* pp. 715-731, 2010.