AGH University of Science and Technology
Faculty of Mechanical Engineering and Robotics

Bachelor Thesis

Sebastian Dziadzio

Unit Selection Text to Speech System for Polish

Supervisor: Bartosz Ziółko, Ph.D.

Kraków 2014

AGH University of Science and Technology

Faculty of Mechanical Engineering and Robotics

Field of Study: Acoustical Engineering

Sebastian Dziadzio

Engineer Diploma Thesis

Unit Selection Text to Speech System for Polish

Supervisor: Bartosz Ziółko, Ph.D.

SUMMARY

The thesis presents consequent stages of building a unit selection text to speech system. The first two chapters treat the definition of speech synthesis and present the history and current state of this research field. Thereafter, a thorough discussion of the distinctive features of Polish are discussed, including its phonology, morphology and prosody. The main part is a report from the process of extracting, balancing, recording and segmenting a speech database. The penultimate chapter presents the unit selection algorithm used in the Festival environment, along with an overview of the most important signal processing algorithms. The closing chapters comprises of a discussion, conclusions and plans for further development.

DECLARATION OF ORIGINALITY

I hereby declare that this thesis has been independently prepared, solely with the support of listed literature references. I certify that, to the best of my knowledge, it does not infringe upon anyone's copyright nor violate any proprietary rights in view of the Act on Copyright and Related Rights of 4th February 1994 (Journal of Law 2006, no. 90, heading 631, as amended) and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices.

CONTENTS

FIGURES

TABLES

Chapter 1

Introduction

In 1965, Herbert Simon, a famous American polymath, wrote in his book called "The Shape of Automation for Men and Management":

> *Machines will be capable, within twenty years, of doing any work that a man can do [45].*

Three years later, 2001: A Space Odyssey was released, a classic Stanley Kubrick's masterpiece based on Arthur C. Clarke novel, establishing a timeless and unforgettable icon for natural human-computer communication. HAL 9000, apart from being one of the most complex and memorable villains in the history of cinema, is also an example of then remarkable optimism in the field of artificial intelligence and man-machine interaction. It is capable of speech recognition and synthesis, natural language processing, understanding and simulating emotional behaviours, lip reading, facial recognition, autonomous reasoning, playing chess and even appreciating works of art.

It took about ten years for the greatest and most influential researches of that time to realise that their estimations were far off. The initial enthusiasm was based on the false premise that since it is possible to write programs using logic, solving hard algebraic and geometric problems, or effectively proving mathematical theorems, it is only a matter of time before machines could equal humans in "simpler" tasks, such as language understanding, visual and auditory perception, commonsense reasoning, and motor skills. It turned out, however, that our distinction between hard and easy problems was evidently based on our own perspective as human beings. In the course of natural selection, a certain set of skills proved to be more desirable than others as far as the question of survival and genetic success was concerned. Not surprisingly, the ability to calculate a complex integral would not dramatically increase the chances of remaining in the gene pool, as opposed to face or voice recognition, interpreting intentions, dodging a moving object, keeping balance, understanding speech, paying attention to rapid changes in the closest surroundings and so forth. As a consequence, our brain is excellent at skills which gave evolutionary advantage from the dawn of our species. Since high-level, abstract reasoning required for mathematical thinking is a relatively new ability, its biological implementation has not yet evolved. At the same time, tasks which are deemed basic and can be effortlessly and automatically performed by our brain, turned out to be extremely complex in terms of their mathematical description. This cognitive phenomenon is known as the Moravec's Paradox and can be stated as follows:

> *The main lesson of thirty-five years of AI research is that the hard problems are easy and the easy problems are hard. The mental abilities of a four-year-old that we take for granted – recognizing a face, lifting a pencil, walking across a room, answering a question – in fact solve some of the hardest engineering problems ever conceived* [39].

The above realisation was one of the landmarks in the history of human-machine interfaces. It became clear that reverse-engineering a complete set of communication skills is impossible and instead the effort should be put into separate tasks, such as speech recognition, question answering, computer vision and so on. Although we now know that constructing a fully functional dialogue system able to reliably pass the Turing test[1] may not be feasible in decades to come, natural language processing, along with speech analysis and synthesis became intensively investigated fields of research. Their applications are almost endless and include search engines, information extraction, machine translation, automatic summarisation, online assistants, biometrical security and more. In many of those systems the most desirable form of output is a synthetic voice.

1. 1. Outline

The following thesis is a report from the development of a unit selection text to speech system for Polish. The described process consists of two main stages: recording a speech corpus and building a working voice in Festival framework – a speech synthesis environment created at the University of Edinburgh and distributed under a license allowing unrestricted commercial and non-commercial use [4].

In Chapter 1, the notion of speech synthesis is introduced and defined in a descriptive manner, by means of establishing requirements that should be fulfilled by an ideal text to speech system. The inevitable tradeoff between vocabulary flexibility and overall quality is

[1] The Turing test evaluates machine's ability to exhibit intelligent behaviour indistinguishable from that of human. The original formulation describes a human judge involved in a natural language conversation with a human and a machine. All participants are separated from each other and the interaction is limited to text-only channel. The machine is said to have passed the test if the judge is unable to reliably distinguish between two interlocutors [54].

also taken under consideration. Then a historical perspective is outlined, including various approaches and strategies over the years of advancements in electrical engineering, computer science and linguistics. Most influential conceptual contributions are discussed: Kratzenstein's realisation that speech is an acoustical signal and is governed by the same laws of physics as any other sound, Von Kempelen's hypothesis attributing the main role in speech production to the vocal tract, and the acoustic theory of speech production allowing to describe speech with a source-filter model and thus giving rise to articulatory and formant synthesis. Particular emphasis is put on analogies between acoustical, electrical and computational models of speech generation.

In Chapter 2, a detailed overview of modern techniques is given. Firstly, diphone synthesis is described. Both PSOLA and MBROLA techniques are mentioned and their assets and shortcomings are discussed. Different approaches to designing and recording the diphone database are presented. The next subsection deals with unit selection, giving a formal definition of the search problem according to Black and Hunt [3]. The questions of base unit and feature systems are also elaborated upon. The penultimate subsection treats the statistical parametric synthesis. Hidden Markov models are introduced and the advantages of this approach are outlined. Finally, other techniques are mentioned, along with the discussion of current challenges and opportunities in the field of speech technology.

Chapter 3 is a thorough analysis of the distinguishing properties of Polish, including its syntax, phonology and morphology. Although some questions and techniques are universal, speech synthesis is obviously a language-specific task. This means that unique traits of a certain language must be considered very early in the process of building a synthesiser. Understanding phonetic phenomena and pronunciation rules is crucial for implementing a functional grapheme-to-phoneme module, which in turn is one of the key factors in obtaining high quality synthesis. In the introduction, a brief description of vocal tract and articulation mechanisms is given, followed by a classification of Polish phones in terms of their articulatory and acoustic features. The International Phonetic Alphabet notation is also introduced, thus enabling a precise description of phonetic phenomena, such as palatalization and voicing. The last subsection deals with prosody, with a particular focus on stress.

In the bulk of this thesis, the most laborious part of the project is addressed, namely designing, recording and labeling the speech corpus. Audio database greatly affects the quality of synthesized speech, consequently being the main determinant of system's overall performance. In short, the more carefully the utterances are selected, recorded and labeled, the more natural sounding voice can be achieved [7]. The third chapter covers in detail all subsequent steps of this process, first one being the search for a large and varied text collection. A sub-corpus of the National Corpus of Polish [42], described in subsection one, was a perfect choice – not only is it abundant and diverse, but also manually annotated. This substantially expedited the task of rejecting foreign words, non-word segments, abbreviations and punctuation. Remaining texts were then rewritten in phonetic notation by means of dedicated software. The transcription served as a ground for automatic balancing of the corpus in terms of phoneme and diphone coverage, eventually leaving a set of prompts for the voice talent. Recording them required specialised and reproducible conditions, professional equipment and adequate software, all described in the penultimate subsection. Finally, the audio files needed to be phonetically labeled. Segmentation is an arduous and time-absorbing task, but its precision directly influences the quality of concatenation. Semi-automating the process is a reasonable compromise between workload and quality.

The first subsection of Chapter 5 is dedicated to Festival environment and FestVox project. Afterwards, important signal processing techniques are introduced, such as preemphasis, framing, windowing, discrete Fourier transform, mel-frequency cepstral coefficients, and linear predictive coding. In the last subsection, two implementations of unit selection algorithms in Festival are presented.

In Chapter 6, the current stage of the project is presented, along with recent plans for further development. Moreover, a short summary is given, discussing the hitherto encountered challenges, impediments, applied solutions and conclusions.

1.2. Speech synthesis

The core task of speech synthesis can be reduced to "taking a series of text words and producing as output an acoustic waveform [28]." Such description is rather vague and does not fully capture the essence of the notion in question, so a more descriptive approach is perhaps more applicative. Consider an ideal text to speech system. According to the aforementioned definition, its input comprises of an arbitrary sequence of words and its output is an utterance in form of an acoustic waveform indistinguishable from a real human voice. Such formulation of the problem entails several substantial ambiguities that require further elaboration. Let us analyse the process of speech synthesis in a step-wise manner.

To begin with, there is a rather unclear view among linguists of what a word is and the perception of this term is manifold. At first glance a phonological definition appears to be most appropriable for the purpose of speech synthesis:

> *A phonological word is a piece of speech which behaves as a unit of pronunciation according to criteria which vary from language to language [52].*

Above requirements, however, provide no clear and universal distinction between words and other units such as phonemes, syllables or breath groups [2]. Moreover, the input of speech synthesis system is in written form, so all the information that could be helpful in identifying phonetic words, such as stress, tone patterns, accent, prosody and pauses [2] are being inevitably lost in transcription. This leads to a conclusion that an orthographic approach may be more useful:

> *An orthographic word is a written sequence which has a white space at each end but no white space in the middle [52].*

This definition would be adequate, if not for its broadness. A text to speech system is not expected to synthesize an arbitrary sequence of characters, because there is often no way of extrapolating a correct pronunciation from phonetic rules. Practical applications of speech synthesis narrow the domain of possible inputs down to words existing in a certain language. Hence it may be worth consideration to define the input of a synthesiser as plain text, understood intuitionally as "a form of unstructured data based on written or printed language [55]".

One of the main challenges in processing plain text is the presence of abbreviations, contractions, logograms, acronyms, and punctuation. Moreover, due to homographs[2], heteronyms[3] and numerals, there is no one-to-one relationship between orthographic words and their pronunciation. Disambiguation is usually done by means of heuristic techniques for part-of-speech tagging, be it decision trees, n-grams, Bayesian classifiers [56] or hidden

[2] Word that shares the same written form as another word but has a different meaning, e.g. bear (an animal or a verb meaning 'carry', 'be equipped with). One Polish example is the word *zamek*, meaning 'a castle' or 'a lock'.

[3] Words that are written identically but have different pronunciation and meaning, e.g. number /nəm-bər/ – a numeral and number /nʌmər/ – comparative form of 'numb'. In Polish they are rather rare, one example could be *cis* /ɕis/ 'yew' and cis /tsʲis/ 'C-sharp'.

Markov models [15]. Analogous methods are used for recognition and classification of foreign words. This entire process is known as text normalisation.

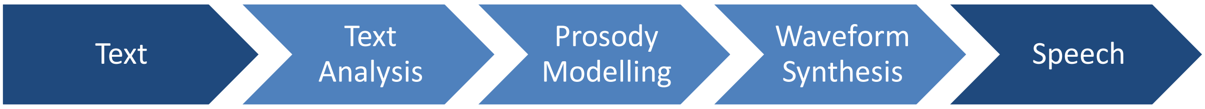


Figure 1. Schematic representation of a concatenative text to speech system.

The next step, determining the pronunciation of a word based on its spelling is often called text-to-phoneme or grapheme-to-phoneme transcription. There are two main strategies in use, both having their own advantages and disadvantages [13]. Their adequacy depends mostly on the properties of a certain language. The first one is the dictionary-based approach, where a large dictionary containing pronunciation of all the morphemes from the input domain is stored by a program. Morphophonemic rules are then applied to predict the pronunciation of derivative, inflectional and compound words. Some regularities are also extracted to provide for out-of-vocabulary words. The alternative is the rule-based approach based on the combination of letter-to-sound rules and a relatively small exception dictionary. The main advantage of the former of two choices is the speed and straightforwardness of the process. Finding a transcription is a matter of looking up a certain word in the dictionary and applying additional rules if necessary. An obvious shortcoming is the domain limitation, significant memory usage required for storing a large pronunciation database and potential complexity of morphophonemic patterns. The latter solution allows to easily take new words into account, but formulating functional letter-to-sound rules may be a demanding task. As has already been said, choosing a right approach mostly depends on language features. For example, 95% of input words in English can be covered by a set of 12 000 morphemes [1], so a dictionary-based approach is easily appropriable. This is possible because English grammar has relatively low inflection. In case of fusional languages using a dictionary is impractical because of the variety of inflected and derived forms. Such languages can be quite reliably modelled by human constructed letter-to-sound rules, as long as they have a phonemic orthography[4]. Otherwise a hybrid solution is required [13]. The data-oriented approach is also

[4] Phonemic orthography is a system for writing a language such that graphemes (written symbols) correspond to phonemes (smallest contrastive linguistic units which may bring about a change of meaning [22]). There is no natural language with an ideal phonemic orthography, i.e. one-to-one relationship between graphemes and

worth mentioning, because it could conceivably be used language-independently [56]. The core concept of this method is extracting letter-to-sound rules from large databases instead of having them built by language experts.

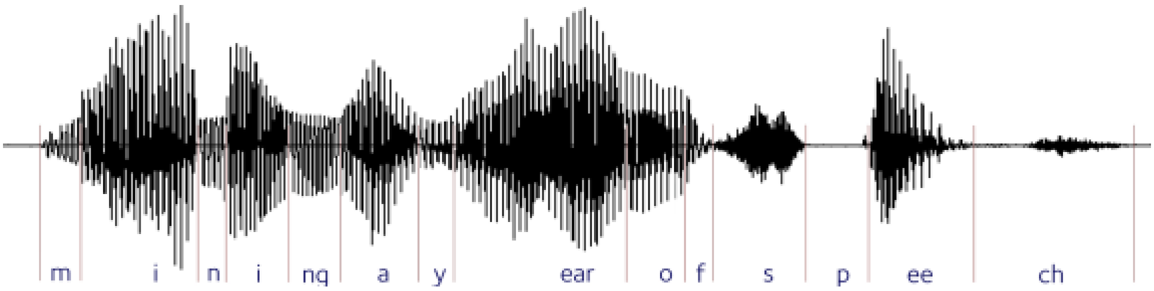


Figure 2. An example speech waveform representing a phrase “mining a year of speech”.

During the process of transcription, information concerning syllabification and accents is added. Later it is fed into the prosodic module, which generates prosodic parameters for each utterance. This is innately a one-to-many problem, because there is never a single correct set of prosodic and accentual features. On the other hand, slight variations of prosody may entirely change the meaning of a certain word or an interpretation of a sentence. For instance, incorrectly placed pitch maximum may turn a verb present /pri-ˈzent/ into a noun /ˈpre-zənt/.

The last stage of the described process is the actual synthesis. In a perfect world, the output is an utterance in form of an acoustic waveform, identical to a human voice in terms of pitch, rhythm, intonation, timbre, diction, stress, pauses and emotional prosody. Achieving such naturalness is not yet possible as there are still multiple challenges that need to be taken into account. Let us briefly elaborate upon them on the example of unit selection synthesis. In this case, final quality depends mostly on speech database[5] and cost function[6]. Nevertheless, even in the ideal case of carefully prepared recordings, accurate labeling and optimised unit selection function, the synthetic voice may be at most clear, fluent and intelligible, but not natural. One of the reasons is the fact that as humans we tend to perceive inerrancy as artificialness. While speaking we often make unnecessary pauses, errors and slips of the tongue, we stammer, produce non-speech sounds, and falter. Most importantly though, we constantly modify our tone, pitch, loudness, and syllable length to convey information that

phonemes. Deviations include multigraphs, historical pronunciations, different graphemes for the same phonemes (for example <u> and <ó> in Polish), rules of pronunciation depending on adjacent letters, and loanwords.

[5] See Chapter 4.

[6] See subsection 2.2.

may not be encoded by the choice of vocabulary, such as the emotional state of the speaker, character of the utterance (question, statement, command), presence of sarcasm, emphasis, contrast, and focus. Although no one really expects a speech synthesiser to deliberately make mistakes, models of semantic and emotional prosody are a promising field of research and may be a big step towards natural-sounding speech synthesis bridging the vocal uncanny valley[7].

However, almost complete naturalness can be achieved for limited-domain synthesis. Prior knowledge of the input allows to adjust the speech databases accordingly. In extreme cases, entire words or phrases are pre-recorded. The process of synthesis is then straightforward, as it reduces to filling the gaps in an utterance template. Such method is inherently limited, so it can be employed only for certain tasks, such as simple dialogue systems, weather reports, or passenger information. Since many interactive voice response technologies are intrinsically limited, it is often possible to obtain an explicit list of utterances that can be generated and an estimation of their frequency [5]. This enables a more general approach – instead of pre-defined prompts, a common unit selection database may be used. By ensuring good coverage of most frequent prompts in several prosodic contexts, it is possible to gain flexibility with no decrease in quality. Notwithstanding, this solution is ineffective in the long run, because the size of the database grows rapidly with the problem [5]. On the whole, there is currently an unavoidable trade-off between naturalness of a synthetic voice and its vocabulary flexibility. As useful as limited domain systems may be, some applications depend on more dynamic response generation, which in turn requires natural-sounding, general-purpose speech synthesis [12].

1.3. History of speech synthesis

The dream of machines able to mimic human or animal behaviour has fascinated mankind for a considerable length of time. The analogy between living organisms and inanimate mechanisms was first explicitly pointed out by Descartes:

[7] Uncanny valley is an aesthetical hypothesis holding that human observers have a tendency to feel revulsion and uneasiness when confronted with almost (but not fully) human-like features. It is primarily applied to appearance of robots or computer animated characters, but a dispassionate and unusually calm voice also evokes anxiety and antipathy. Notable examples include aforementioned HAL 9000, Hannibal Lecter and GLaDOS – the main villain in *Portal* computer game series.

> *It seems reasonable since art copies nature, and men can make various automata which move without thought, that nature should produce its own automata much more splendid than the artificial ones. These natural automata are the animals [16].*

These views seemed to resonate among the great artisans of XVII and XVIII century[8], who produced a variety of clockwork toys. Descartes himself believed that language is the main reason why it is impossible to build a humanoid automaton:

> *A machine could never arrange its speech in various ways in order to reply appropriately to everything that could be said in its presence [16].*

Nonetheless, numerous attempts of reproducing speech with inanimate mechanisms were made in XVIII century. The pioneer in this field was Christian Gottlieb Kratzenstein, a German-born physician, engineer and physicist. Around 1769 he constructed a vowel synthesiser consisting of a free reed and a set of resonators, essentially playing the role of acoustic filters.

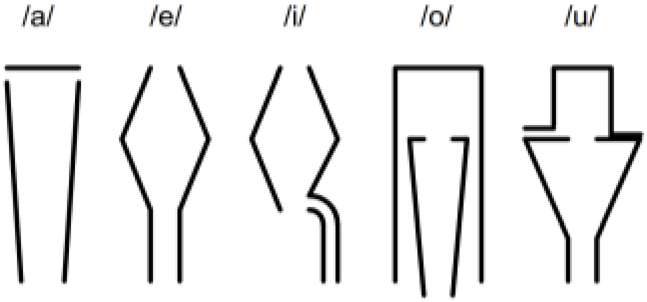


Figure 3. Kratzenstein's resonators for vowels. Each sketch is a section showing the varying diameter through the longitudinal axis of a cylindrical tube. The input end is on the bottom.

Although his accomplishment does seem modest from the present perspective, it contributed to the idea that speech is subject to the same physical principles that govern the generation of other sounds, for example in musical instruments. Kratzenstein's views on phonetics were far from correct and it is apparent from his writing that he was designing the resonators mostly by trial and error, but the importance of his work is mainly conceptual [35].

Another milestone in the history of speech synthesis was the machine of Johann Wolfgang von Kempelen, a Hungarian-born inventor. His apparatus, when in hands of a skilled operator, allegedly allowed not only to produce simple speech sounds, but also entire

[8] Athanasius Kircher, Jacques de Vaucanson, Pierre Jaquet-Droz, Henri Maillardet, John Joseph Merlin.

words and sentences. It consisted of a pair of bellows for lungs, a vibrating reed acting as vocal cords, a hand-varied leather resonator simulating the vocal tract and even special apertures in place of nostrils. The quality of its voice, as demonstrated with modern replicas, was rather poor. Nonetheless, von Kempelen's studies led to formulation of the theory that the vocal tract is the main site of acoustic articulation, contrary to the prior belief attributing key role to the larynx.

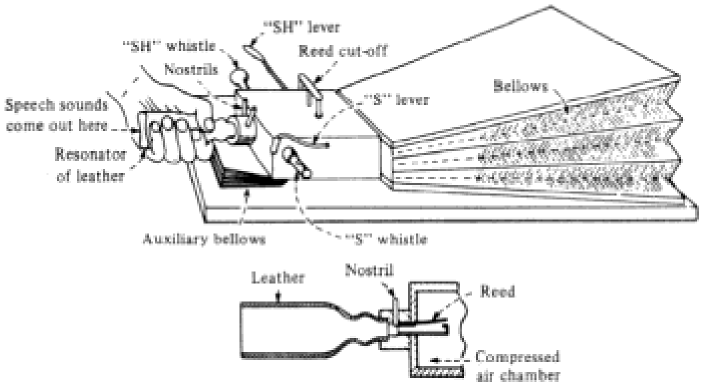


Figure 4. 4. Von Kempelen's apparatus (from [21]).

It was not until the twentieth century when a first electronic counterpart of von Kempelen's machine is built. Called the Voder and developed in Bell Telephone Labs by Homer Dudley, it was the first attempt to synthesise speech by breaking it into frequency bands. It used two sound sources, a microphone, a bank of ten analog band pass filters, an amplifier, and a loudspeaker. The device was operated through a set of keys and foot pedals, giving control over the type of sound (hissing noise for sibilants and buzz tone for vowels and nasals), pitch, prosody and resultant frequency response. Special keys allowed to generate plosives and affricatives[9]. Although quite complex to operate (it was said to require a year of constant training), it was not only able to produce understandable speech, but also sing or imitate animal and industrial sounds.

The fifties mark the formulation of the acoustic theory of speech production, holding that speech wave can be viewed as the response of the vocal tract filter systems to one or more sound sources. Consequently, any speech signal can be uniquely described in terms of source and filter characteristics [20].

[9] For an explanation of phonetic terms, see Chapter 3.

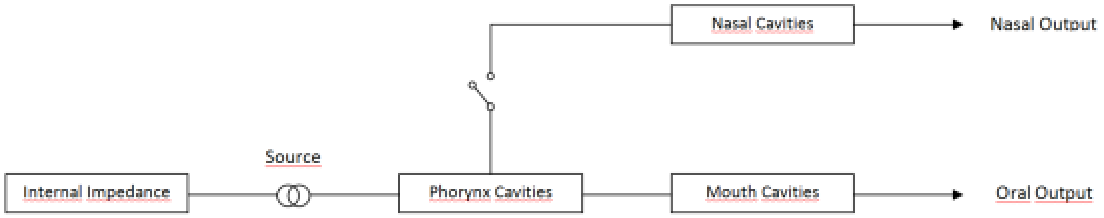


Figure 5. Speech production as a filtering process. Schematic representation of the production of sounds employing a glottal source (based on [20]).

Application of this theory resulted in emergence of two methods, namely formant[10] and articulatory synthesis. In the former technique, the output is created by means of additive synthesis and acoustic models. This allowed to produce robotic, yet smooth and intelligible speech by avoiding artifacts peculiar to concatenative synthesis. Significant formant synthesis systems were: Walter Lawrence's PAT (1953), OVE I-III [30] (1952-1967), and GLOVE [44]. In contrast, articulatory synthesis is based on simulating the effects of the vocal tract, which can be modelled acoustically (as in von Kempelen's machine), electrically, or computationally. Examples include George Rosen's DAVO (1958) and a system created by Noriko Umeda and his associates in 1968.

Inability to achieve natural speech with acoustic approach gave rise to concatenative method, in which the output signal is constructed by stringing together elementary segments of recorded speech. In 1958, diphones were first proposed as basic units. It was argued that they enable concatenation in the stable part of the phone, which reduces glitches and discontinuities [33] [38].

[10] Formants are the spectral peaks of the sound spectrum of the voice [20].

Chapter 2

State of the art in speech synthesis

Over the last 80 years, due to advancements in computer science and linguistics, synthesised speech went a long way from low quality, unintelligible and audibly artificial to easily comprehensible and almost natural-sounding. There are numerous commercial solutions available and speech synthesis is nowadays virtually ubiquitous due to popularity of online assistants and knowledge navigators in smartphones (Siri for Apple iOS and Samsung's S-voice). New techniques such as statistical parametric synthesis, based on Hidden Markov Models, proved to be very effective in generating acceptable speech [9]. Unit selection is a more conventional method that has dominated speech synthesis for over a decade. With a carefully prepared database and a well-designed cost function, it is capable of producing a natural and intelligible voice [14].

2.1. Diphone synthesis

Rapid development of computers in terms of memory sizes and processing power redounded to the expansion of concatenation techniques. The idea behind them is to record a speech database ensuring a single appearance of each basic unit (usually a diphone[11]). Theoretically it is then possible to generate any sequence of phones with recombination techniques. The first challenge for this type of synthesis is to determine a full list of legal diphones for a certain grammar. Given that sequences both within and across word boundaries have to be included, the diphone inventory is usually just slightly smaller than the theoretical maximum (all possible pairs). Afterwards, a representative real-speech example for every basic unit needs to be found. The first strategy for this task is to search for diphones in a large speech corpus from a single speaker. Another approach is to deliberately design and record a minimal speech corpus providing full diphone coverage. Since some units are extremely rare

[11] Diphones are speech segments usually extending from the mid-point of one phone to the mid-point of the next one. In a language with N unique phonemes there are just less than N^2 diphones, as not all combinations occur in practice.

or can exist only on the word boundaries, it is often easier to create a set of invented words (logotomes) according to some template.

With an appropriately chosen and recorded diphone set it is usually possible to concatenate units without any smoothing at the boundaries [52]. Unfortunately, while it is possible to guarantee full coverage for the range of phonetic effects, taking all prosodic contexts into account is not feasible. This entails the necessity of using signal processing algorithms. Ideally, they should achieve the desired prosodic effect while leaving the rest of the signal as it was. The most common approach is called pitch synchronous overlap and add (PSOLA). It works by dividing the speech waveform in small overlapping segments called epochs. In order to change the pitch, they are moved closer together or further apart. To modify time, they are either duplicated or deleted. After being rearranged, the epochs are recombined using the overlap-add technique. PSOLA comes in several variants, such as time-domain, frequency-domain, and linear-prediction. The main disadvantage of this method is its sensitivity to errors in epoch placements. Insufficient accuracy in the process of marking the segments may result in very poor quality manifesting itself with irregular periodicity, perceived by humans as hoarseness. This can be avoided by using an alternative technique called MBROLA, which uses sinusoidal modelling to decompose each frame and then resynthesise the database at a constant pitch and phase [52].

2.2. Unit selection

Although the intelligibility of diphone systems is usually satisfactory, their naturalness often leaves much to be desired. This is caused mostly by prosodic modifications, limited number of phone variants in the database, and artifacts emerging from inadequate concatenation. Moreover, the assumption that all variation within a single diphone can be accounted for by pitch and time modification, is simply wrong and substantially limits performance of such systems [52].

Above observations gave rise to the unit selection method. The idea is to store several instances of each unit and select the appropriate one during synthesis. For example, the database could contain four versions of each diphone: stressed, unstressed, phrase-final and non phrase-final. Assumedly, with a sufficiently big database, little or no signal processing is required to achieve desired prosody. Unit selection relies on large speech corpora, as they can be used to a full extend, as opposed to the wasteful approach in diphone synthesis, where one diphone required recording a whole utterance[12]. At the core of unit selection lies the notion of cost function, estimating how well a certain unit fits the specification in terms of phonetic and prosodic features (target cost) and concatenation (join cost).

The first formulation of unit selection was proposed by Alan Black and Andrew Hunt in 1996[3]. They defined the technique as a search problem over the space of every possible sequence of units with a heuristic in form of the cost function. The specification is given in the form of a list of diphones $S = \langle s_1, s_2, \dots, s_T \rangle$, each described by a set of features. The database can be viewed as a set of diphone units $U = \{u_1, u_2, \dots, u_M\}$, each of which is also characterised by a feature structure. The target cost $T(u_t, s_t)$ is the distance (calculated in the adopted feature system) between the specification s_t and candidate unit u_t. In contrast, the join cost $J(u_t, u_{t+1})$ estimates the quality of concatenation for two consecutive units u_t and u_{t+1}. The task is to find a sequence of units $\widehat{U}$ such that:

$$(2.1)\quad \widehat{U} = \operatorname*{argmin}_{u} \left\{ \sum_{t=1}^{T} T(u_t, s_t) + \sum_{t=1}^{T-1} J(u_t, u_{t+1}) \right\}$$

There are several possibilities as far as the base unit is concerned: single frames[13], states[14], half-phones[15], phones, diphones, demi-syllables[16], di-syllables[17], syllables, words, or phrases. Apart from homogeneous systems using a single type, there are also heterogeneous, which use a combination of two or more. The choice of base unit is largely discretionary, but sometimes it is dictated by language features. For instance, Chinese is often considered to be syllable-based, as opposed to phone-based European languages.

In the best case, unit selection allows to raise the quality of speech synthesis to that of recorded natural speech, because long contiguous sections of speech are frequently chosen [6]. However, the robustness of such systems is unsatisfactory, as in some cases they are known to give very poor results. A lot of effort has to be put into the design of corpus and selection function to ensure a stable performance. Choosing the feature system and metric for target and join cost is a non-trivial and heavily language-dependant task and several approaches, such as evolutionary algorithms [50], were proposed.

[12] It is impossible to capture a specific diphone in isolation, so carrier words or phrases are necessary, see [51].
[13] During speech processing the signal is decomposed into overlapping frames to ensure it is locally stationary.
[14] Parts of phones, described as the alignment of HMM models.
[15] Half-phones usually extend from phone beginning to mid-point or from mid-point to end.
[16] The syllable equivalent of half-phones.
[17] The syllable equivalent of diphones.

2.3. HMM synthesis

Concatenative synthesis, be it diphone or unit selection, has some obvious limitations. First of all, there is a constraint on how many phonetic and prosodic contexts can be included in the database. Secondly, it is very hard to ensure an even coverage of various effects in the speech corpus. Lastly, although concatenative synthesis can be described as data-driven, there is no learning mechanism. Effectively, all data is memorised, because acoustical parameters are derived directly from recorded waveforms. The alternative is to use the speech database to build a statistical model. This gives rise to two main advantages: smaller memory usage (only the parameters of the model need to be stored) and greater flexibility (it is possible to modify the model in any way one sees fit) [52].

Hidden Markov models (HMM) are general statistical models, originally developed for speech recognition, but currently extensively used in cryptanalysis, handwriting and gesture recognition, machine translation, financial simulations, part-of-speech tagging, biochemistry and bioinformatics. Generally speaking, they allow to model any generative sequence that can be characterised by an underlying process generating a sequence of observations. Formally, a HMM can be defined as follows:

$$(2.2)\ \ \lambda = (A, B, \pi)$$

S is the state set and V is the observation set:

$$(2.3)\ \ S = (s_1, s_2, s_3, \dots, s_N)$$

$$(2.4)\ \ V = (v_1, v_2, v_3, \dots, s_M)$$

Q is the fixed state sequence of length T and O is the corresponding sequence of observations:

$$(2.5)\ \ \ Q = (q_1, q_2, q_3, \dots, q_T)$$

$$(2.6.)\ \ O = (o_1, o_2, o_3, \dots, o_T)$$

The transition array A stores the probability of state j following state i:

$$(2.7)\ \ A = [a_{ij}]\ \ \ a_{ij} = P(q_t = s_j | q_{t-1} = s_i)$$

The observation array B stores the probability of observation k being emitted from a state i:

$$(2.8)\ \ B = [b_i(k)]\ \ \ b_i(k) = P(x_t = v_k | q_t = s_i)$$

It is worth pointing out that both transition and observation probabilities are independent of time.

An exhaustive description of HMM is beyond the scope of this paper, so let us only formulate two important assumptions made by this model. The first one holds that the current state depends only on the previous one. This feature of a stochastic process is known as the Markov property. The second postulate is called the independence assumption. It states that the output observation at time t is dependent only on the current state, i.e. previous observations and states have no effect on it [10]. Because of these two premises, HMM can be used as a declarative representation of the acoustics of phone. In speech recognition, the task is then to find a sequence of states given the observation. Speech synthesis is an opposite process: the sequence of models is given by the specification (usually along with the desired temporal structure) and the challenge is to generate spectral and excitation parameters [58]. The models for synthesis usually consist of vectors of MFCC[18] and logarithmic fundamental frequencies (in recognition MFCC and energy is usually sufficient).

The main advantage of statistical parametric systems is that the voice features (pitch, emotional prosody, accent, emphasis, speed) can be easily modified by changing the HMM parameters. Achieving the same result with a unit selection synthesiser is impossible, as the database would have to cover all desired voice characteristics and speaking styles. Moreover, trained statistical models are compact and provide means to automatically train the specification-to-parameter rules, thus avoiding hand-written ones.

2.4. The future

Although unit selection and statistical parametric synthesis prevail as far as contemporary commercial solutions are concerned, other approaches to modelling speech are still investigated. For example, advancements in biomechanical imaging allowed using ultrasonography and EMMA[19] to build 3D models of the vocal tract, which are then translated to acoustical parameters [25]. This form of articulatory synthesis is not expected to outperform unit selection and HMM-based systems, but it can provide valuable data concerning the mechanisms of speech production, which then can be used to refine other methods of synthesis [37].

[18] Mel-frequency cepstral coefficients. See Chapter 5.

[19] Electro-magnetic midsagittal articulometers.

As far as the opportunities in the field of synthesis are concerned, statistical parametric synthesis is a noteworthy trend. The flexibility, mobility and reliability of such systems is an inestimable advantage. The only hindrance at the moment are the limitations of signal processing algorithms, resulting in various artifacts. However, signal processing in speech synthesis is now receiving more attention and a lot of improvements are to be expected. Progress in machine learning will probably contribute to the further development of data-driven techniques, both for text processing and speech modelling. Dialogue systems in cars, television sets, personal computers, and mobile devices are becoming increasingly popular, although they are still treated as a novelty and serve mostly for entertainment. Considering the fact that there are about 1.4 billion smartphones in use at the moment (compared to 1.2 billion PCs)[20], speech technology was never so ubiquitous. Since two most influential companies producing mobile devices (Samsung and Apple) use speech recognition and synthesis in their products, an increasing interest in speech and language technology can be expected.

[20] Source: Business Insider Intelligence Estimate, Gartner, IDC, Strategy Analytics, company filings, World Bank 2013.

Chapter 3

Polish language

Prior information about the structure of language, its phonetics, syntax, semantics, and morphology can be very useful in the process of building a text to speech system. First of all, understanding syntactic and semantic mechanisms is indispensable for text normalisation, especially for homograph disambiguation and interpretation of numerals, abbreviations and acronyms. It is particularly important for inflecting languages, like Polish. Secondly, extensive knowledge about phonetics is required for the grapheme-to-phoneme conversion. Additionally, proper design of a representative database is impossible without linguistic and phonologic bases. The same applies to prosody, which obviously varies significantly across languages. In case of unit selection, careful choice of the feature system and the metric[21] is essential for achieving reliable results. Again, there is no universal solution to this problem, because some features turn out to be more important for certain languages [50]. In short, every stage of text to speech conversion requires awareness of language-specific features.

3.1. Articulation

Knowledge of the biomechanical aspect of speech production is a prerequisite for understanding phonetic definitions which will be introduced in succeeding subsections. Speech organs are generally divided into two types: passive and active articulators. The former group includes upper lips, teeth, alveolar ridge[22], hard palate[23], soft palate[24], uvula[25], and pharynx[26] wall, while the latter comprises of the tongue, lower lips and glottis[27] (see Figure 6).

[21] The task comes down to finding weights associated with parameters: F0, pitch, energy, spectral features etc.

[22] Upper alveolar ridge is located between upper teeth and hard palate. The lower ridge is located on the bottom of the mouth, behind the teeth.

[23] Thin horizontal bony plate of the skull, located in the roof of the mouth.

[24] Soft tissue constituting the back of the roof of the mouth.

[25] Conic projection from the posterior edge of the middle of the soft palate.

[26] Part of the throat situated immediately posterior to the nasal cavity, posterior to the mouth and superior to the oesophagus and larynx.

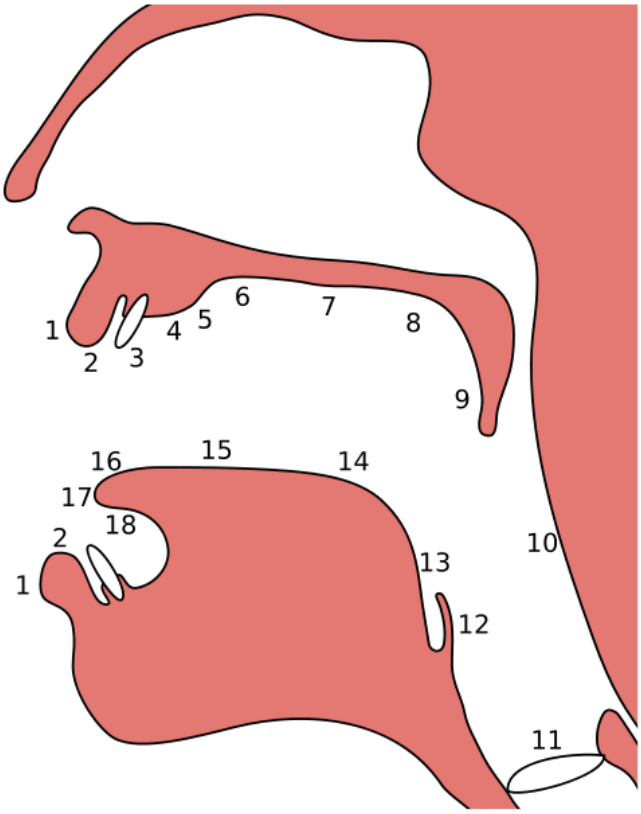


Figure 6. Places of articulation.

1. Exo-labial (outer part of lip)
2. Endo-labial (inner part of lip)
3. Dental (teeth)
4. Alveolar (front part of alveolar ridge)
5. Post-alveolar (rear part of alveolar ridge)
6. Pre-palatal (front part of hard palate)
7. Palatal (hard palate)
8. Velar (soft palate)
9. Uvular (uvula)
10. Pharyngeal (pharyngeal wall)
11. Glottal (vocal folds)
12. Epiglottal (epiglottis)
13. Radical (tongue root)
14. Postero-dorsal (back of tongue body)
15. Antero-dorsal (front of tongue body)
16. Laminal (tongue blade)
17. Apical (apex or tongue tip)
18. Sub-laminal (underside of tongue)

Articulation is usually defined as the movement of the articulators in order to generate speech sounds. In essence, sound production is just expelling air from lungs through the vocal tract. To produce distinguishable speech sounds, pairs of speech organs come close to each other to produce an obstruction, thus shaping the air in a particular fashion. The point of maximal obstruction is known as the place of articulation.

[27] Space between the vocal folds.

In general, when the obstruction is being formed, one of the speech organs (active) is moving, while the other (passive) remains stationary. Hence, to uniquely define a place of articulation, one needs to know the place of active and passive articulation. In most cases, the active speech organ is the tongue. For example, post-alveolar articulation occurs when the tongue blade contacts the region behind the alveolar ridge. Sometimes, the coarctation of the vocal tract is the effect of two active articulators approaching each other. For instance, while producing bilabial speech sounds[28], both lips come together. Another example is laryngeal articulation, where the only obstruction occurs in the vocal cords.

Another important criterion for speech sound classification is the manner of articulation. This notion refers to the general characteristics of speech organs, not just the place of maximal obstruction. Let us briefly define the manners of articulation in English (main differences for Polish will be discussed in the next subsection):

1. Degree of stricture

- stop (complete blockage followed by rapid release): t, d, p, b, k
- fricative (close enough to cause significant airflow turbulence): f, v, s, z, sh, zh, th
- affricate (complete blockage followed by gradual release): ch, j
- approximant (far enough apart that airflow remains smooth): r, y, w, h

2. Alternative air flow

- nasal (air can freely flow only out the nose): m, n, ng
- lateral (blockage of air by the center of the tongue, air can flow out the sides): l

3. Movement of the tongue

- flap (short complete blockage): t, d between vowels in the American pronunciation
- trill (multiple brief blockages in a row): r by some English speakers

The above classification applies to consonants, i.e. speech sounds articulated with complete or partial closure of the vocal tract. The other group of speech sounds are vowels, generated with no constriction of the vocal tract. They are almost always voiced, which means that vocal

[28] /p/, /b/, /m/

cords are vibrating during the articulation[29]. There are several articulatory features that distinguish vowel sounds:

1. Height, named for the vertical position of the tongue. In IPA[30] defined according to the degree of jaw opening.

2. Backness, named for the position of the tongue, defined according to the relative frequency of the second formant. The lower F2, the more retracted the vowel.

3. Roundedness, refers to whether the lips are rounded.

4. Nasalisation, refers to whether some air escapes through the nose.

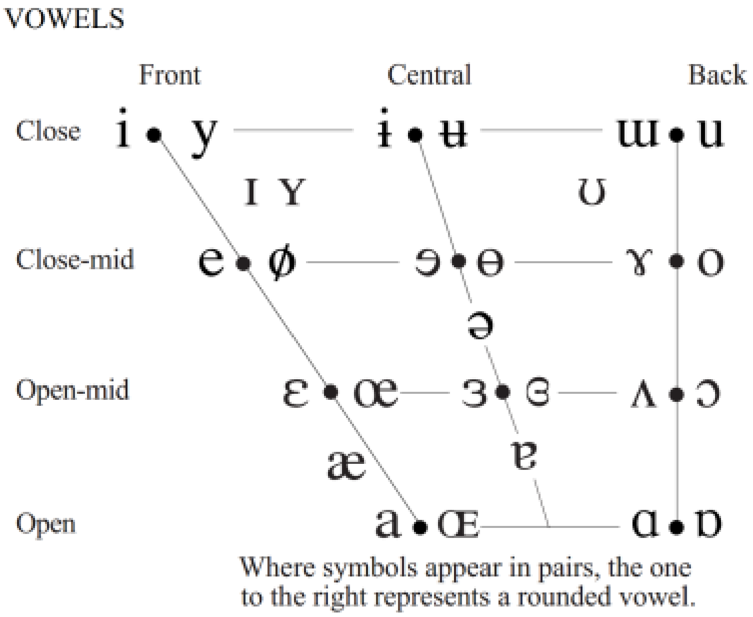


Figure 7. Articulatory features of vowels in IPA (from official IPA chart 2006).

3.2. Classification of Polish phones

Before proceeding with the classification of Polish speech sounds, it is worth to establish a common notation system. For the purpose of the project, we adapted the International Phonetic Alphabet, which is a standardised representation of the sounds of oral language. An important feature of IPA is that it normally has one letter for one speech sound. This means that phonetic symbols are not context- and language-dependent (if no language

[29] Exceptions include several Native American languages. Vowels are also devoiced in whispered speech.
[30] See the next subsection.

makes a distinction between two sounds, they will be denoted by the same symbol). The IPA-based systems of transcription differ in certain ways from those traditionally employed by Slavic and Polish works in phonetics and phonology but the systems are mutually translatable [23].

Vowels

Table 1 presents six oral vowels in Polish, their orthographic transcription, corresponding IPA symbols and example words. Several comments should be made. First of all, the vowel /u/ has two historically justified spellings: <ó> and <u>. There is no difference in articulation between them whatsoever. Secondly, letter <i> can denote a front vowel /i/ or have a purely orthographic role of marking palatalized nature of the preceding consonant, when followed by a vowel, e.g. *siano* /ɕanɔ/ 'hay'. It can also do both at the same time, e.g. *siwy* /ɕiwɨ/ 'grey'. Moreover, it can denote a consonant /j/, as in *hiena* /xjɛna/ 'hyena' or a combination of /j/ and /i/, e.g. *szyici* /ʃɨjiɕi/ 'Shias'.

Vowel		Articulatory features	Example	IPA	Translation
ORT	**IPA**				
a	/a/	front, open	matka	/matka/	mother
e	/ɛ/	front, half-open	wesele	/vɛsɛlɛ/	wedding
i	/i/	front, close	igła	/igwa/	needle
o	/ɔ/	back, half-open, round	most	/mɔst/	bridge
u/ó	/u/	back, high, round	stół, ul	/stuw/, /ul/	table, beehive
y	/ɨ/	front, half-close, retracted	wydry	/vɨdrɨ/	otters

Table 1. Polish vowels (based on [23].

As far as Polish nasal vowels *ę, ą* are concerned, they usually consist of mid-vowels /ɔ/, /ɛ/ followed by a nasalised labio-velar glide /w̃/ or the nasalised palatal glide /j̃/. Moreover, graphemes <ę>, <ą> may correspond to a combination of an oral vowel and a nasal consonant.

Vowel		Example	IPA	Translation
ORT	IPA			
ę	/ɛ$^{\tilde{w}}$/	gęsty	/gɛ$^{\tilde{w}}$stɨ/	thick
	/ɛ$^{\tilde{j}}$/	gęś	/gɛ$^{\tilde{w}}$ɕ/	goose
	/ɛn/	wędka	/vɛntka/	rod
	/ɛɲ/	sędzia	/sɛɲʥa/	judge
	/ɛm/	tępy	/tɛmpɨ/	blunt
ą	/ɔ$^{\tilde{w}}$/	wąs	/vɔ$^{\tilde{w}}$s/	moustache
	/ɔŋ/	bąk	/bɔŋk/	bumblebee
	/ɔɲ/	wziąć	/vʑɔɲʨ/	to take
	/ɔn/	wątroba	/vɔntrɔba/	liver
	/ɔm/	kąpiel	/kɔmp^{j}jel/	bath

Table 2. Example pronunciations of graphemes *ę* and *ą* (based on [23]).

Consonants

	Labial	Labio-dental	Dental	Alveolar	Alveolo-palatal	Palatal	Velar
Plosive	/p/ /b/		/t/ /d/			/c/ /ɟ/	/k/ /g/
Fricative		/f/ /v/	/s/ /z/	/ʃ/ /ʒ/	/ɕ/ /ʑ/		/x/
Affricate			/ʦ/ /ʣ/	/ʧ/ /ʤ/	/ʨ/ /ʥ/		
Nasal	/m/		/n/		/ɲ/		/ŋ/
Lateral			/l/				
Trill				/r/			
Glide	/w/					/j/	

Table 3. Simplified consonantal system of Polish (based on [34]).

The above table illustrates the classification of Polish consonants in terms of their articulatory features. Let us present some examples of their orthographic transcription: /p/: *para* /para/ ‘pair’, /b/: *bar* /bar/ ‘bar’, /t/: *tor* /tɔr/ ‘track’, /d/: *dar* /dar/ ‘gift’, /c/: *kiwać* /civatɕ/, /ɟ/: *ogier* /ɔɟɛr/ ‘stallion’, /k/: *kat* /kat/ ‘executioner’, /g/: *gad* /gat/ ‘reptile’, /f/: *fan* /fan/ ‘fan’, /v/: *wat* /wat/ ‘watt’, /s/: *ser* /sɛr/ ‘cheese’, /z/: *zupa* /zupa/ ‘soup’, /ʃ/: *szał* /ʃaw/ ‘rage’, /ʒ/: żal /ʒal/ ‘sorrow’, /ɕ/: *śmiech* /ɕm^{j}jex/ ‘laughter’, /ʑ/: *zima* /ʑima/ ‘winter’, /x/: hak /xak/ ‘hook’, /ʦ/: *car* /ʦar/ ‘tzar’, /ʣ/: *dzwon* /ʣvɔn/ ‘bell’, /ʨ/: *ćma* /ʨma/ ‘moth’, /ʥ/: *dziób* /ʥup/ ‘beak’, /m/: *mak* /mak/ ‘poppy’, /n/: *nad* /nat/ ‘above’, /ɲ/: *nic* /ɲiʦ/ ‘nothing’, /ŋ/: *pstrąg* /pstrɔŋk/ ‘trout’, /l/: *las* /las/ ‘forest’, /r/: *raj* /raj/ ‘paradise’, /w/: *łyk* /wɨk/ ‘sip’, /j/: *jak* /jak/ ‘how’.

Although the orthography of Polish is highly phonemic (certainly to a greater extent than in case of English), there are a few inconsistencies. They can be useful in text processing, as they allow to differentiate between homophones, but for text to speech systems they are an additional setback:

1. The sound /ʒ/ can be written orthographically either as <ż>, e.g. *może* /mɔʒɛ/ ‘(s)he can’ or <rz>, as in *morze* /mɔʒɛ/ ‘sea’. However, in some cases <rz> denotes a combination of /r/ and /z/, e.g. obmierzły /ɔbm^{j}jɛrzwɨ/ ‘detestable’.

2. The sound /x/ is rendered orthographically either as <ch>, e.g. *kocha* /kɔxa/ ‘(s)he loves’, or <h>, e.g. *waha* /vaxa/ ‘(s)he hesitates’.

3. The affricatives /ʣ ʧ ʤ/ are spelled <dz cz dż>, e.g. *rdza* /rʣa/ ‘rust’, *czar* /ʧar/ ‘spell’, *dżuma* /ʤuma/ ‘plague’. In contrast, transitions /t-ʃ d-ʒ/ are recorded as <drz> and <trz>, e.g. *trzoda* /t-ʃɔda/ ‘flock’, *drzewo* /d-ʒɛwɔ/ ‘tree’. There are very rare exceptions, e.g. *nadżerka* /nad-ʒɛrka/ ‘laceration’.

4. The alveolo-palatals /ɕ ʑ ʨ ʥ ɲ/ are spelled <ś ź ć dź ń> pre-consonantally and word-finally, e.g. *myśl* /mɨɕl/ ‘thought’, *łoś* /wɔɕ/ ‘moose’, *jaźń* /jaʑɲ/ ‘ego’, *maź* /maʑ/ ‘slime’, *ćma* /ʨma/ ‘moth’, *dźwig* /ʥvig/ ‘crane’. However, they are spelled <si zi ci dzi ni> before a vowel, as in the following examples: *sito* /ɕitɔ/ ‘sieve’, *zima* /ʑima/ ‘winter’, *cisza* /ʨiʃa/ ‘silence’, *dziwny* /ʥivnɨ/ ‘strange’, *nic* /ɲiʦ/ ‘nothing’. There are some exceptions, especially in loanwords, e.g. *sinus* /sinus/ ‘sine’, *siwert* /sivert/ ‘sievert’.

3.3 Phonetic phenomena

Coarticulation

Coarticulation occurs when a speech sound is influenced by adjacent speech sounds. It may have anticipatory[31] or carryover[32] character. In Polish, it alters the place of articulation, which may result in voicing or devoicing of some speech sounds. Plosives, fricatives and affricates located at the end of an isolated or sentence-final word are devoiced. This causes following substitutions: /b/→/p/, /d/→/t/, /g/→/k/, /v/→/f/, /z/→/s/, /ʒ/→/ʃ/, /ʑ/→/ɕ/, /ʣ/→/ʦ/, /ʥ/→/ʨ/. Coarticulation takes place also in case of consonant groups containing both voiced and unvoiced speech sounds. The character of such cluster depends on its last element[33] (anticipatory coarticulation). Devoicing occurs in case of the following groups: /bk/→/pk/, /wʃ/→/fʃ/, /zɕ/→/sɕ/, /bs/→/ps/, /bx/→/px/, /bc/→/pc/, /ʣk/→/ʦk/, /wp/→/fp/, /ʒk/→/ʃk/, /vc/→/fc/, /zʨ/→/sʨ/, /bsk/→/psk/, /dsk/→/tsk/, /bʃʧ/→/pʃʧ/, /dɕʨ/→/tɕʨ/, /vsp/→/fsp/, /vʃʧ/→/fʃʧ/, /wɕʨ/→/fɕʨ/, /zdk/→/stk/, /ʑʥs/→/ɕʨʦ/, /ʒʤk/→/ʃʧk/, /vsx/→/fsx/. The opposite process (voicing) is less frequent: /kʒ/→/gʒ/, /ɕb/→/ʑb/, /ʧb/→/ʤb/. In some rare cases, coarticulation has a carryover character. This happens when either /v/ or /ʒ/ (but only rendered as <rz>) is located at the end of the consonant clustered and preceded by an unvoiced phone. In such case, /v/ and /ʒ/ are devoiced.

Palatalisation

Palatalisation takes place when a speech sound (usually a consonant) is pronounced with a palatal secondary articulation. In general, the meaning of this notion is twofold. It may refer to an articulatory feature: a speech sound is palatalised if during pronunciation the tongue is raised toward the hard palate and alveolar ridge. It also denotes an assimilatory coarticulation process involving a front vowel /i/ or a palatal approximant /j/, causing adjacent phones to shift towards palatal articulation. The differentiation between palatal and palatalised speech consonants can be therefore a complicated question [23]. Polish linguists [51] usually introduce a distinction between soft and hard consonants, where the former group includes /ɕ

[31] Occurs when the features of a speech sound are anticipated during the production of the preceding speech sound.

[33] This also applies to clusters extending across word boundaries.

ʑ tɕ dʑ ɲ/. Remarkably, they are all speech sounds of their own, not just the palatalised versions of /s z z dz n/. In contrast, /p^j b^j f^j v^j m^j c ɟ/ are not treated as separate phones.

3.4 Prosody

Prosody refers to the rhythm, stress[34], tone[35], accent[36] and pitch accent[37] of speech. It can facilitate speech perception, express emotions, give cues about intentions, convey additional semantic information or even be essential for proper interpretation of the utterance. It serves to draw receiver's attention to the tenor or temper of the sender's nominal verbal message. To realise its importance, consider a text-only channel, e.g. Facebook chat. Despite the extensive use of emoticons, a lot of messages remains ambiguous, especially when irony is concerned. It is also difficult to draw receiver's attention to certain aspects of the message. This is done in speech by means of stress. Consider the fragment "she noticed me at the bar" in the following sentences:

> ***She*** *noticed me at the bar (of all the people that could have noticed me).*
>
> *She* ***noticed*** *me at the bar (despite me being terrible with women).*
>
> *She noticed* ***me*** *at the bar (though she might have preferred my friend).*
>
> *She noticed me* ***at the bar*** *(although I promised to stop drinking)***.**

The above example proves that the same text accompanied by different meta-communication can slightly change its meaning. This is especially important in case of questions:

> ***What*** *have you done? (I can tell you did something.)*
>
> *What* ***have*** *you done? (I think you have not done anything. Prove me wrong.)*
>
> *What have* ***you*** *done? (What was your contribution?)*
>
> *What have you* ***done****? (You stupid person!)*

In Polish, prosody is also necessary to determine the form of the utterance, as there is sometimes no other difference between a question and a statement, e.g. the phrase *zrobili to* may mean 'they did it' or 'did they do it?', depending on intonation.

[34] Lexically specified distinction between strong and weak syllables. Stressed syllables are usually louder and longer than unstressed ones.

[35] Lexically specified pitch shift, property of a syllable.

[36] Post-lexical pitch movement, linked to a stressed syllable.

[37] Lexical pitch movement, property of a word.

Generating adequate and natural sounding intonation remains one of the biggest challenges in TTS systems. Its accuracy and authenticity (or lack thereof) greatly affects the overall naturalness of synthesised speech. At the same time, it is rather difficult to capture and describe prosody as a feature of an acoustic waveform, so a lot of research is carried out to build functional models for speech synthesis.

Polish has a strong tendency towards fixed lexical stress on the penultimate syllable (paroxytone) and a secondary stress on the initial syllable in words containing more than three syllables [36]. Exceptions include:

1. Proparoxytones (words with lexical stress on the antepenultimate syllable).

   - first and second person plural forms of verbs in the indicative past, e.g. *zrobiliśmy* 'we did'
   - first, second and third person singular forms in the conditional mood, e.g. *zrobiłbym* 'I would do', *poszedłbyś* 'you would go', *zjadłaby* 'she would eat'
   - some Greek loanwords, e.g. *matematyka* 'mathematics', *muzyka* 'music', *polityka* 'politics'
   - some numerals ending with -*set*, e.g. osiemset 'eight hundred', *dziewięćset* 'nine hundred'

2. Words with lexical stress on the preantepenultimate syllable.

   - first and second person plural forms in the conditional mood, e.g. *zrobilibyśmy* 'we would do'

3. Oxytones (words with lexical stress on the last syllable).

   - some words with the prefix *anty-*, e.g. *antygen* 'antigen'
   - some loanwords, e.g. *à propos*
   - some fixed phrases, e.g. *do cna* 'fully', *na skos* 'diagonally'
   - abbreviations, e.g. *ONZ (Organizacja Narodów Zjednoczonych*) 'UN (United Nations)'

As far as intonation is concerned, modern studies [17, 18] based on statistical analysis lead to the formulation of a theory holding that there is a finite number of melodic patterns, each forming an intonation phrase. They were divided into eight categories according to the F0[38] contour:

Class	**Description**	**F0 contour**
HL	full falling	reduced from highest level to lowest, global maximum at the beginning and global minimum at the end
ML	low falling	reduced from mid to lowest, global minimum at the end
HM	high falling	reduced from high to mid level, global maximum at the beginning
XL	extra low falling	reduced from lower mid to below minimal
LM	low rising	increases till mid level, global minimum at the beginning
MH	high rising	increases from mid to highest level, global maximum at the end
LH	full rising	increases from lowest level to highest, global minimum at the beginning and maximum at the end
LHL/MHL	rising-falling	increases to global maximum and then is reduced to global minimum

Table 4. Classes of nuclear accents (based [18]).

[38] Fundamental frequency of the vocal cords, approximately 200 Hz for adult women and 125 Hz for adult men.

Chapter 4

Speech corpus

Almost every modern text to speech system, whether it uses concatenation or statistical models, relies on the quality of speech database. The design and recording of the speech corpus was the most important and laborious part of the project. It required a thorough theoretical preparation: choosing the most appropriate phonetic alphabet, finding an applicable balancing method and doing a lot of research on recording practices. It also involved more businesslike tasks: obtaining a large enough text collection for input, writing lots of text processing scripts, acquiring all the necessary equipment, gaining access to the anechoic chamber, finding a volunteer voice talent, and finally making the recordings. The following chapter is a detailed account of this process and a presentation of its outcome.

4.1. NKJP

One of the most important requirements that a speech corpus must meet is representativeness. In order to achieve this, it should be generated based on a large and diverse set of texts. Such collections are obtained from various sources: newspaper archives, transcripts of parliament speeches, web-crawlers and so on. The main problem with this approach is usually the quality of text data. There is no easy way of filtering out undesired segments (abbreviations, foreign words, punctuation) to obtain "pure" text data.

The National Corpus of Polish (NKJP, *Narodowy Korpus Języka Polskiego*) is the biggest corpus of modern Polish, containing over fifteen hundred million words. It was developed as a shared initiative of four institutions: Institute of Computer Science at the Polish Academy of Sciences, Institute of Polish Language at the Polish Academy of Sciences, Polish Scientific Publisher PWN, and the Department of Computational and Corpus Linguistics at the University of Łódź. It has been carried out as a research-development project of the Ministry of Science and Higher Education. The list of sources contains classic literature, daily newspapers, scientific periodicals and journals, transcripts of conversations, and a variety of short-lived and Internet texts. The most important feature, which makes

NKJP especially fit for our purposes is the fact that it is searchable by means of advanced tools that analyse Polish inflection and sentence structure.

The one million manually-labeled sub-corpus, used for our work, is available for non-commercial use on the GNU general public license. It has the form of XML files containing the texts and morphosyntactic tags. Thanks to this we were able to automatically reject all foreign words, incomplete or corrupted segments, punctuation, and non-alphanumeric characters. This was done by means of Ruby scripts, provided by courtesy of Aleksander Pohl [41].

4.2. Grapheme-to-phoneme transcription

In order to balance the corpus, i.e. ensure a satisfactory coverage of phonetic units, the entire text collection mentioned in the previous subsection needed to be transcribed into phonetic form. This of course must have been done automatically, by means of a dedicated software.

Ortfon 2.0. is a software developed by the Digital Signal Processing Group at the Chair of Electronics at AGH University of Science and Technology. Thanks to its rule-based approach it allows to obtain a phonetic transcription of any text. It takes an orthographic version as input and renders it into phonetic notation of choice (SAMPA[39], IPA, AGH[40], Corpora[41]). It is compatible with several encoding standards, including UTF-8, ISO-8859-2 and CP-1250. Currently it is still under development, so before it could be used it in this project, it had to be carefully tested and improved. First of all, the transcription rules had to be updated. Initially they were based on a publication by Maria Steffen-Batóg [48]. The task was to change some of them according to the more recent work by Andrzej Pluciński [40]. Afterwards, they had to be experimentally validated. The manual inspection of the results (about 100 randomly chosen short paragraphs from NKJP) allowed to find five incorrect rules.

Another important question was the choice of phonetic alphabet. Several criteria had to be taken into consideration. The SAMPA and AGH alphabets have the important one-byte-

[39] Speech Assessment Methods Phonetic Alphabet, a mapping of IPA into 7-bit printable ASCII characters.

[40] Internal phonetic alphabet for AGH systems. It is a one-byte-per-phoneme version of SAMPA.

[41] Phonetic alphabet used by Stefan Grocholewski in the CORPORA project, a speech database for Polish diphones [24].

per-symbol property. The main and decisive shortcoming of the AGH alphabet was the fact that its use is restricted only to internal systems. As regards SAMPA, despite its universality it is only a partial encoding of IPA. This means that a lot of possibly important information would be lost in transcription. Hence, we finally decided to use a set of 78 IPA phones:

i	u	m̥	v^j	ɣ	b^j
ĩ	ũ	m^j	f	x	p
ɨ	j	n	f^j	ç	p^j
ɨ̃	j̃	n̥	z	ʣ	d̯
ɛ	w	n^j	z^j	ʣ^j	d^j
ɛ̇	w̥	n̳	ʑ	ʥ	ḍ
ɛ̃	w̃	ɲ	s	ʦ	t
ɛ̇̃	l	ɲ̊	s^j	ʦ^j	t^j
a	l^j	ŋ	ɕ	ʨ	ṭ
a̟	r	ŋ̩	ʒ	ʤ	g
ɔ	r̥	ŋ^j	ʒ^j	ʧ	ɟ
ɔ̃	r^j	ŋ̩^{j}	ʃ	ʧ^j	k
ɔ̟̃	m	v	ʃ^j	b	c

Table 5. A set of IPA phones used in this project.

4.3. Balancing

The transcription of the entire input text collection served as a basis for the process of extracting a smaller, representative corpus according to some criteria. From practical reasons and due to limited technical possibilities we decided to build a database of 2500 utterances, each consisting of at most eighty phonemes. The target coverage was set to at least forty appearances of each phoneme and at least four appearances of each diphone. The balancing was done with CorpusCrt, a toolkit developed at Universitat Politècnica de Catalunya and freely available for non-commercial use. It uses a greedy iterative algorithm to extract a text database meeting predefined requirements. The solution will always be suboptimal, but with a big enough input the results are satisfactory.

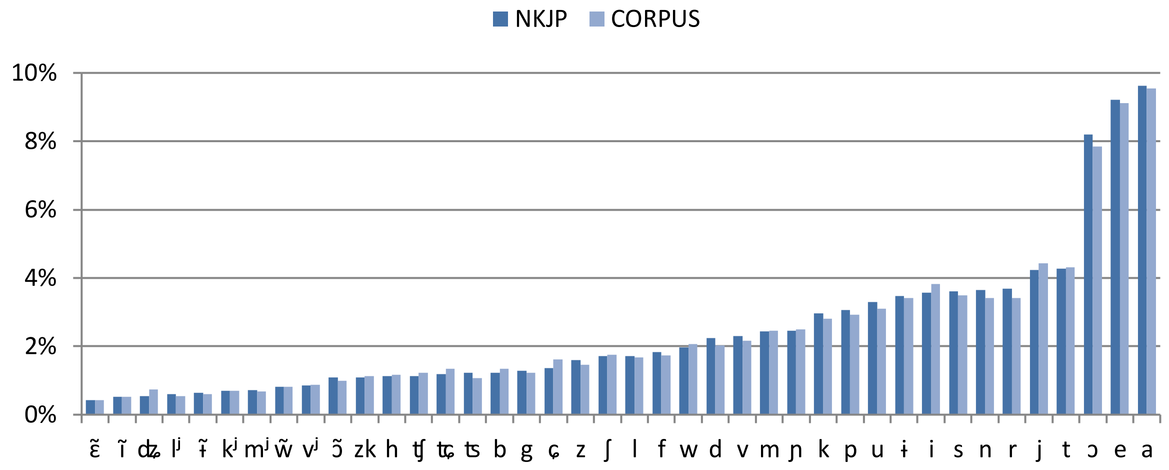


Figure 8. Comparison of the coverage of forty most popular phonemes between the initial text collection and the final corpus.

It is worth to make a comment about the desired distribution of phonemes. Generally, two approaches have been proposed [31]: the first one is to select sentences which contain phonetic events with respect to their frequency in natural speech (phonetic balancing). This results in poor representation of less frequents phonemes, but ensures better coverage of units which will appear more often during synthesis. An alternative approach is uniform balancing, where a target distribution is uniform. Meeting this criterion is of course impossible, but it will increase the number of rare units in the database which in turn may significantly improve the quality of synthesis in some exceptional cases.

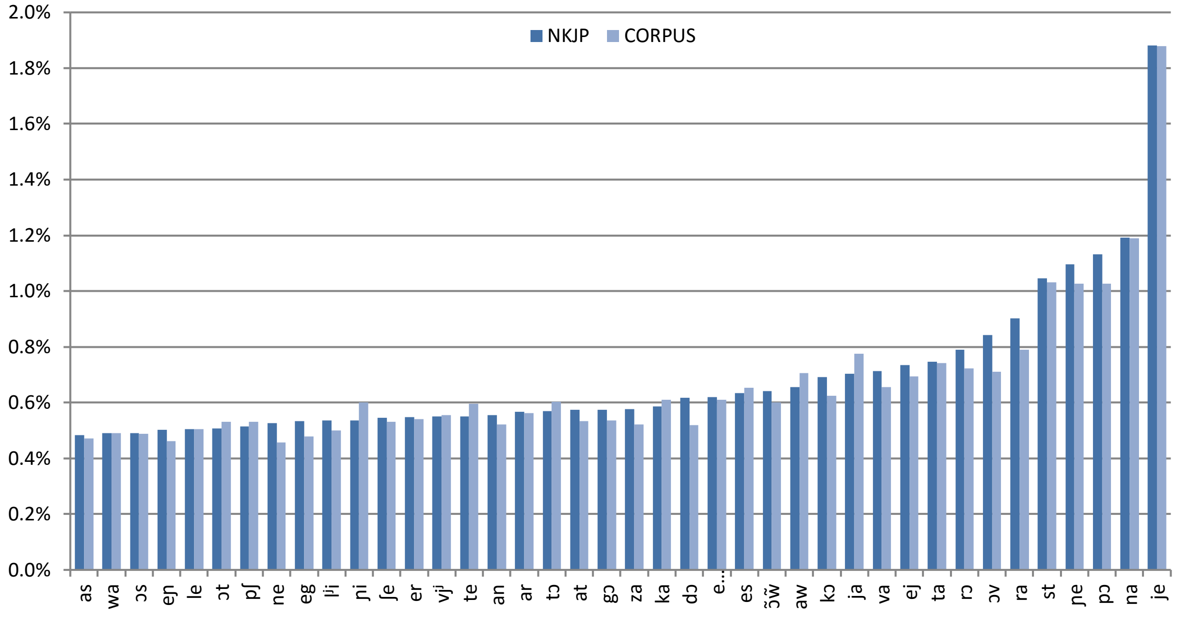


Figure 9. Comparison of the coverage of forty most popular diphones between the initial text collection and the final corpus.

4.4 Recording

In order to ensure smooth concatenation, every recording session should be carried out in identical acoustic environment, introducing little or no noise. The sessions should feature a trained voice talent recorded with professional equipment. This is naturally impossible to achieve in most cases, but thanks to the courtesy of Technical Acoustic Laboratory and the Chair of Electronics, we were able to partially fulfill those requirements.

Anechoic Chamber

The recordings were conducted in the bigger anechoic chamber of the Technical Acoustic Laboratory at AGH UST. The chamber is a 600 ton, ten-meter wide ferroconcrete cube with a functional volume of 342 m^3. It is fixed on a set of 25 heavy-duty mechanical vibration isolators, each consisting of four strings designed to damp all vibration caused by the urban surroundings. The measured sound pressure level of the ambient noise is 1.5 dB_A[42] during daytime and even 0 dB_A at night [29].

Microphone

We decided to use AKG C 4000 B model, multi-pattern condenser microphone with an electret capsule, made available by courtesy of the Digital Signal Processing Group. Thanks to its dual-diaphragm assembly, it is possible to achieve either an omnidirectional, hypercardioid or cardioid pattern. Its output level is 25 mV per Pascal with a standard 48V phantom power supply. It also incorporates a switchable 10 dB pre-attenuator to reduce the likelihood of overload and a switchable high-pass filter, which may be employed to avoid high-energy, low-frequency distortions. We decided to use the cardioid polar pattern to avoid distortions caused by the equipment and recording personnel. We also used the high pass filter, and a supplied H100 elastic suspension to eliminate low-frequency acoustical[43], electrical[44] (energy network) and mechanical[45] distortions. To attenuate the energy of aspirated plosives (/p/, /b/, /d/ etc.), which could exceed the design input capacity of the microphone, we resorted to a pop-filter.

[42] A-weighted sound pressure level.

[43] Breath, humming noise from the lights, ventilation systems, working computer etc.

[44] The energy network can produce distortions [48].

[45] The chamber floor takes the form of a suspended net made of metal ropes, so its transmits mechanical distortions easily.

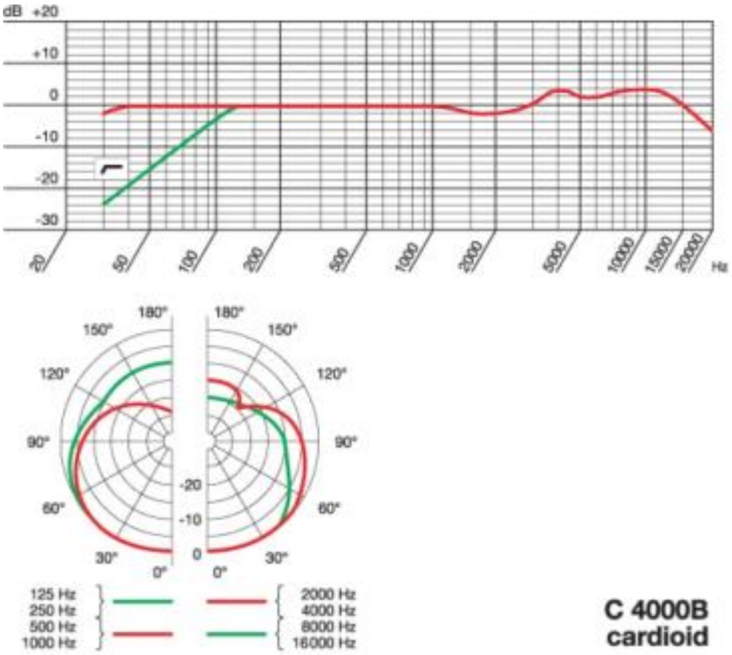
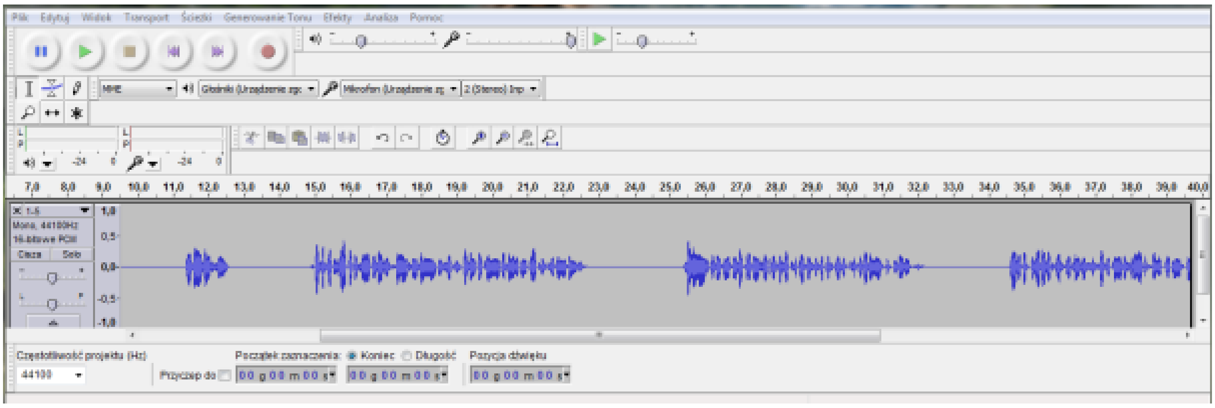


Figure 10. Frequency response of AKG C 4000B microphone set to cardioid polar pattern (from AKG website).

Software

For recording and processing we used Audacity, a free open source digital audio editor and recording application. It was used to capture, pre-process and export the input from the Lexicon interface (I-O|82 and Lambda). The resolution was set to 16 bit, with the sampling frequency of 44.1 kHz. The recordings were then exported to waveform audio format and saved on the hard drive. No signal processing was done at that stage, apart from some audio file manipulation (cut, copy, paste, delete).



Częstotliwość projektu (Hz): 44100

Przyczep do □

Początek zaznaczenia: ● Koniec ○ Długość

00 g 00 m 00 s

00 g 00 m 00 s

Pozycja dźwięku: 00 g 00 m 00 s

Figure 11. Audacity main window.

Recording sessions

Because of the non-commercial character of the project, we were unfortunately unable to hire a professional voice talent. However, we managed to find three volunteers, two men and one woman, each having some experience with voice acting. We then auditioned all the candidates and assessed them in terms of their vocal qualities: intelligibility, timbre, diction, pronunciation. We also evaluated their ability to maintain unvarying intensity, pitch and speed during long and exhaustive recording sessions. We finally chose to record a female voice.

Each session started with a vocal warm-up, followed by several test recordings, which were then compared to previously captured audio files in order to give some suggestions or correct the equipment setting. Then the proper recording session started, in form of 45 minutes cycles with at least 10 minutes breaks between them. At the beginning of each cycle, the voice talented listened to previously recorded utterances to maintain the same voice qualities.

Altogether, recording the database took six sessions, each lasting from five to nine hours (including a longer break). Four and a half hours of records were made at approximately five hundred tracks, each containing five utterances separated with silence. The total duration of solely the utterances is approximately three hours. It should be noted that the excessive length of recording sessions was caused by limited access to the anechoic chamber and recording equipment.

4.5. Segmentation

Labeling the recordings will be the most laborious part of the project, despite the planned semi-automation of the process. First of all, all tracks need to be split into separate files, each one containing a single utterance. Afterwards, the word-level segmentation will have to be performed by hand, using labeling software developed by the Digital Signal Processing Group. The phone-level segmentation will be done with Hidden Markov Models, trained on a manually segmented learning set.

Chapter 5

Building a voice

The final stage of the project will be building a functional TTS system based on the Festival environment. At present, there are two techniques for building a unit selection voice. The first one, cluster unit selection is a reimplementation of the technique described by Alan Black and Paul Taylor [4]. The alternative, called Multisyn, is a more general-purpose unit selection algorithm. Both are described in detail in the third subsection.

5.1. Festival and FestVox

The Festival Speech Synthesis System is a general multi-lingual speech synthesis system developed at the Centre for Speech Technology Research at the University of Edinburgh. As authors claim, it is designed for at least three levels of users:

> *Festival is designed as a speech synthesis system for at least three levels of user. First, those who simply want high quality speech from arbitrary text with the minimum of effort. Second, those who are developing language systems and wish to include synthesis output. In this case, a certain amount of customization is desired, such as different voices, specific phrasing, dialog types etc. The third level is in developing and testing new synthesis methods [4].*

It is distributed under a free software license. It offers full text to speech through a number of APIs, but more importantly it provides a general framework for building speech synthesis systems. It is written in C++ and uses the Edinburgh Speech Tools Library for low level architecture and has a Scheme based command interpreter for control. It is designed to support many languages, at present there exist packages for English, Welsh, Spanish, Czech, Finnish, Hindi, Italian, Marathi, Polish, Russian and more.

The FestVox project, carried out by Carnegie Mellon University's speech group aims at advancing the state of speech synthesis by providing better documentation for the Festival environment, sharing specific scripts, example databases, links, demos and repositories. All the resources are available for free without any restrictions.

5.2. Speech signal processing

The electrical signal generated by the microphone is the effect of transducing the vibration of the diaphragm, which in turn is caused by the sound wave. This means that the resultant waveform is a time-domain representation of the speech signal. However, the human perception of speech (and sound in general) is based on its frequency spectrum. That is why numerous techniques were designed to enable frequency-domain analysis of sound signals.

Framing and Windowing

One of the key assumptions made during spectral analysis of a speech signal is that it can be regarded as stationary (having constant spectral characteristics) over a short interval, typically of a few milliseconds. The first stage of processing is therefore dividing the signal into blocks in order to derive smoothed spectral estimates of each of them. The blocks are overlapped to give an analysis window of approximately 25 ms. The spacing between blocks is typically 10 ms. It is usual to apply a tapered window function (e.g. Hamming) to each block. Pre-emphasis is also often used, in form of high-pass filtration. It aims at compensating for the attenuation caused by the radiation from the lips [58].

DFT and FFT

Discrete Fourier Transform (DFT) converts a finite list of equally spaced samples of a signal into the list of coefficients (finite combination of complex sinusoids ordered by their frequencies). The input samples are complex numbers (in most practical applications they are real numbers), as well as the output coefficients. The sequence of N complex numbers $x_0, x_1, x_2, \dots, x_{N-1}$ is transformed into a N-periodic sequence of complex numbers $X_0, X_1, X_2, \dots, X_{N-1}$ according to the following formula:

$$(5.1)\ \ X_k = \sum_{n=0}^{N-1} x_n \cdot e^{\frac{-i2\pi kn}{N}}$$

DFT is widely used across a large number of fields, but its utility for spectral analysis is of greatest interest for us. When DFT is used for this purpose, $\{x_n\}$ represents a set of time-samples of the microphone output waveform x(t). A plot presenting DFT calculated for each frame of the signal is called a spectrogram.

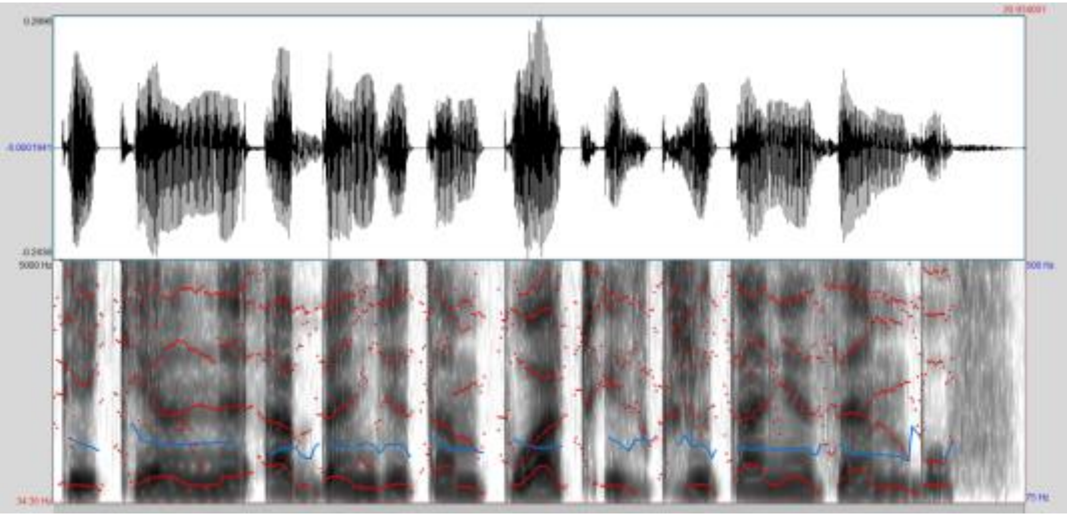


Figure 12. A spectrogram of one of the recordings. Visible formants (red) and fundamental frequency (blue).

Fast Fourier Transform (FFT) is a collective term describing a number of algorithms designed to effectively compute the DFT and its inverse. Because of their speed[46] and accuracy, they are widely used for many applications in engineering, science and mathematics.

Mel-frequency cepstrum

Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on linear cosine transform of a log power spectrum on a nonlinear scale of frequency. To compute MFC coefficients the Fourier spectrum is smoothed by integrating the spectral coefficients within triangular frequency bins arranged on the mel-scale[47]. Afterwards, log compression is applied to the filter-bank output. The final stage is to apply the Discrete Cosine Transform (DCT) to the log filter bank coefficients. This has the effect of compressing the spectral information into lower order coefficients and it also decorrelates them. Usually the signal energy and first 12 cepstral coefficients are combined to form a 13-element basic acoustic vector. First and second differentials are then appended to incorporate dynamical information about the signal, thus creating a 39-element acoustic vector.

[46] There are FFTs with O (N $\log_2$ N) complexity for all N. This is a substantial improvement when compared to the original O (N^2).

[47] Perceptual scale designed to approximate the frequency resolution of the human ear being linear up to 1000 Hz and logarithmic thereafter.

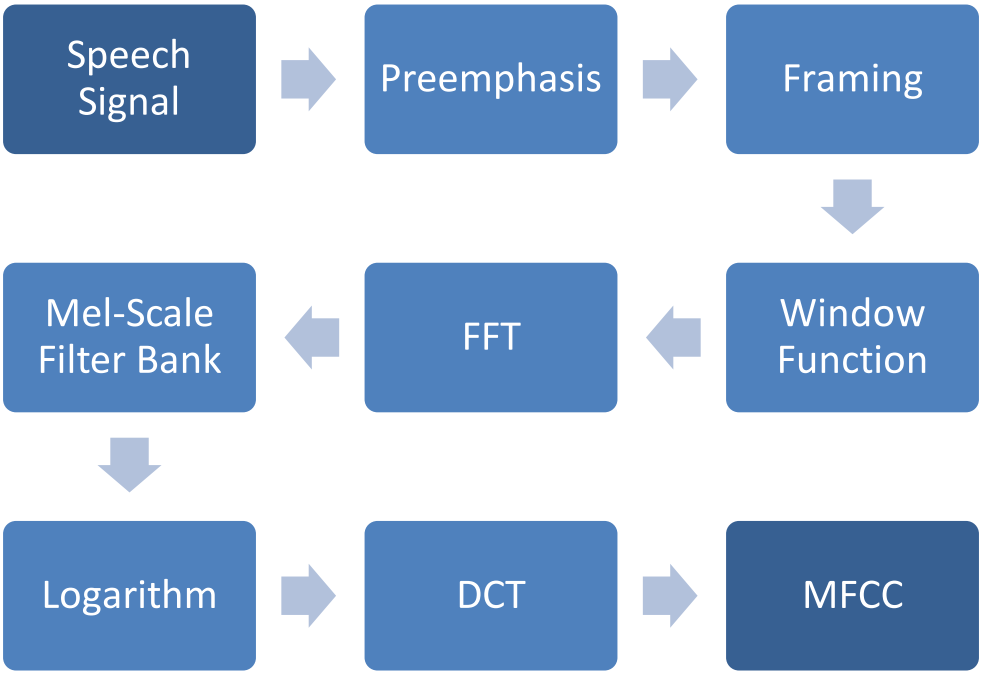


Figure 13. The full process of calculating MFC coefficients.

Linear Predictive Coding

Linear Predictive Coding (LPC) is a digital method for encoding an analog signal in which a particular value is predicted by a linear function of the past values of the signal. In case of modelling speech, the LPC model is based on a mathematical approximation of the vocal tract represented by a tube of varying diameter. At a particular time t the speech sample $x(t)$ is represented as a linear sum of the p previous samples [11].

5.3. Unit selection in Festival

In the introduction to this chapter, two techniques for unit selection were mentioned. The first one is called cluster unit selection (hereinafter referred to as clunits) and was proposed in 1997 by Alan Black and Paul Taylor [4]. The idea is to cluster a database of general speech into groups of acoustically similar units based on the information available at synthesis time. This may include phonetic context, F0, duration and higher-level features such as stressing, accents and word position. A group of candidates is therefore selected and finding the best overall selection can be reduced to finding the best path through each set of candidates for each target phone. The feature space and the metric used for clusterisation can be easily modified. The process of building a voice (with a collected database of general speech at disposal) can be divided into following stages:

- building utterance structures
- calculating coefficients for acoustic distances, e.g. LPC, cepstrum + F0
- building distances tables, precalculating the acoustic distance between each unit of the same phone type
- dumping selection features (phone context, prosodic, positional and other) for each unit type
- building cluster trees

The second method was proposed by Robert Clark, Korin Richmond and Simon King [14]. It is capable of building open-domain voices, as opposed to clunits, performing best in limited domains. It is implemented as a module of Festival, so it just replaces some modules from the pipeline for the standard diphone method. In particular, it shares the common data structure called the Utterance. Multisyn is currently a homogenous system with diphone as a base unit. The authors believe that selection of bigger units should result from the search (through selection of contiguous sequences of smaller units, in this case diphones), rather than be predefined [14]. Nevertheless, the system is implemented in a way that supports using other units with little programming effort.

There are several substantial differences between Clunits and Multisyn methods. Most importantly, the basic unit of the former is a single phone, while the latter is based on diphones. The advantage of the diphone method is a smaller probability of bad joins, because of the context being taken into account. Its major shortcoming is a bigger feature space and squared size of the inventory (there is almost N^2 diphones in a language with N phones). The second major difference is the target cost implementation. The Multisyn approach scores units based upon matches in their linguistic context, whereas the Clunits approach uses the linguistic features to predict a unit’s gross acoustic properties and then performs the scoring in acoustic space [14].

Chapter 6

Discussion

Building a functional TTS system is a complicated and multi-faceted process. It requires multi-disciplinary theoretical knowledge as well as practical skills and a fair amount of arduous labour. The following chapter is a summarisation of all the progress up to now, along with a discussion of key challenges and most recent development plans.

6.1. Final remarks

The task of building an acoustic database requires precision and thoroughness. Thanks to a profound study of available publications on similar projects [32, 47, 48] we were able to avoid some basic mistakes, such as DC distortions, noisy acoustic environment, variable conditions, inconsistent speed, pitch and intensity of the voice, mechanical distortions, incorrect sampling frequency and more.

The recording conditions were as close to professional as we could get, considering the non-commercial nature of the project. Due to limited access to the anechoic chamber we were often stressed for time during the sessions, which probably might have negatively affected the uniformity of the audio corpus. However, it has to be noted that databases recorded in much worse environments by people who have never worked with voice, resulted in high quality synthesis [47].

One of the important issues was the choice of phonetic alphabet. We finally opted for IPA mostly because it still leaves out the possibility of mapping it into any smaller alphabet by means of clusterisation. An opposite operation, i.e. extending from SAMPA or AGH to IPA obviously would not be possible.

Although the main purpose of the recorded database is building a TTS system, a professionally-recorded, well-segmented audio corpus is an asset for the Digital Signal Processing Group, as it can also be used for refining the ASR system currently under development.

6.2. Plans

The next stage of the project is the segmentation of the database. As it was said in Chapter 4, a semi-automation of this process is planned. The idea is to manually label a training set such that every phoneme will appear at least three times. Thereafter a HMM based ASR (Automatic Speech Recognition) system will be trained and used to label the entire speech database. The result of this process will have to be manually inspected and probably corrected. Then the MLF files will have to be modified to describe diphone segmentation instead of phoneme.

As far as the unit selection method is concerned, the Multisyn approach seems to be most appropriable. The quality of concatenation is the priority, especially because a squared feature space is not a problem considering average processing power of modern computers. The system was intended to be open domain, so using Clunits would probably result in poor quality.

Another important task will be the implementation of text-normalisation module and grapheme-to-phoneme rules. The former will probably be limited to the simple lookup table for numerals (without taking inflection or compound numerals into account) and most common abbreviations. As far as the latter is concerned, two strategies are possible. The first one is to use the Festival’s built-in module and rewrite the rules manually from Ortfon. The advantage of this approach is that it would ensure compatibility with other modules. However, the syntax of grapheme-to-phoneme rules in Festival requires exhaustive enumeration, which is almost impossible for 78 units. An alternative solution is to build a separate module in C++ and then access Festival through its C++ API.

References

1. ALLEN James, HUNNICUT Sheri (1987): *From Text to Speech: the MITalk System.* – New York: Cambridge University Press.
2. ALWOOD Jens, HENDRIKSE A.P., AHLSÉN Elisabeth (2009): *Words and Alternative Basic Units for Linguistic Analysis.* – in: HENRICHSEN Peter Juel (ed.): Linguistic Theory and Raw Sound: Copenhagen Studies in Language. Samfundslitteratur, pp. 9-26.
3. BLACK Alan, HUNT Andrew (1996): *Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database.* – in: Proceedings of ICASP-96, vol. 1, pp. 373-376.
4. BLACK Alan, TAYLOR Paul (1997): *Automatically Clustering Similar Units for Unit Selection in Speech Synthesis.* – in: Proceedings of Eurospeech, pp. 601-604.
5. BLACK Alan, TAYLOR Paul, CALEY Richard (1999): *The Festival Speech Synthesis System: System Documentation.* – Centre For Speech Technology Research, University of Edinburgh.
6. BLACK Alan, LENZO Kevin (2000): *Limited Domain Synthesis.* – in: Proceedings of ICSLP, pp. 411-414.
7. BLACK Alan (2003): *Unit Selection and Emotional Speech.* – in: Proceedings of Eurospeech, pp. 1649-1652.
8. BLACK Alan, LENZO Kevin (2007a): *Building Synthetic Voices.* – Carnegie Mellon University.
9. BLACK Alan, TOKUDA Keiichi, ZEN Heiga (2007c): *Statistical Parametric Speech Synthesis.* – in: Proceedings of ICASSP-07, pp. 1229-1232.
10. BLUNSOM Phil (2004): *Hidden Markov Models.* – University of Melbourne.
11. BRADBURY Jeremy (2000): *Linear Predictive Coding.* – University of Ontario.
12. BULYKO Ivan, OSTENDORF Mari (2002): *A Bootstrapping Approach to Automating Prosodic Annotation for Limited-Domain Synthesis.* – University of Washington.
13. BYEONGCHANG Kim, GEUNBAE Lee, JONG-HYEOK Lee (1999): *Hybrid Grapheme to Phoneme Conversion for Unlimited Vocabulary.* – Cambridge University Press.

14. CLARK Robert, RICHMOND Korin, KING Simon (2007): *Multisyn: Open-domain Unit Selection for the Festival Speech Synthesis System.* – in: Speech Communication, vol. 49, pp. 317-330.
15. CUTTING Doug, KUPIEC Julian, PEDERSEN Jan, SIBUN Penelope (1992): *A Practical Part-of-Speech Tagger.* – in: Proceedings of the Third Conference on Applied Natural Language Processing, pp. 133-140.
16. COTTINGHAM John (1978): *A Brute to the Brutes? Descartes' Treatment of Animals.* – in: Philosophy, vol. 53, pp. 551-559.
17. DEMENKO Grażyna, NOWAK Ignacy, IMIOŁCZYK Janusz (1993): *Analysis and Synthesis of Pitch Movements in a read Polish Text.* – in: Proceedings of Eurospeech, pp. 797–800.
18. DEMENKO Grażyna, Jassem Wiktor (1999): *Modelling Intonation Phrase Structure with Artificial Neural Networks.* – in: Proceedings of Eurospeech, pp. 711-714.
19. DUDLEY Homer, TARNOCZY T.H. (1950): *The Speaking Machine of Wolfgang von Kempelen.* – in: Journal of Acoustical Society of America, vol. 22, pp. 151-166.
20. FANT Gunnar (1970): *Acoustic Theory of Speech Production.* – The Hague: Mouton&Co.
21. FLANAGAN James (1972): *Speech Analysis, Synthesis, and Perception*. – Berlin: Springer Verlag.
22. GIMSON Alfred Charles (2008): *Gimson's Pronunciation of English.* – London: Hodder Education.
23. GUSSMANN Edmund (2007): *The Phonology of Polish.* – New York: Oxford University Press.
24. GROCHOLEWSKI Stefan (1997): *CORPORA – Speech Database for Polish Diphones.* – in: Proceedings of Eurospeech, pp. 1735-1738.
25. ISKAROUS Khalil, GOLDSTEIN Louis, WHALEN D.H., TIEDE Mark, Rubin Philip (2003): *CASY: The Haskins Configurable Articulatory Synthesiser*. – in: Proceedings of XV ICPhS, pp. 185-188.
26. JASSEM Wiktor (2003): *Polish.* – in: Journal of the International Phonetic Association, vol. 33, pp. 103-107.
27. JUANG B.H., RABINER Lawrence (2004): *Automatic Speech Recognition – A Brief History of the Technology Development.* – Georgia University of Technology.

28. JURAFSKY Daniel, MARTIN James (2009): *Speech and Language Processing.* – New Jersey: Prentice Hall.
29. KAMISIŃSKI Tadeusz, FLACH Artur, FELIS Józef, PILCH Adam (2013): *The Negative Aspects of Automation of Selected Acoustic Measurements Performed in an Anechoic Chamber*. – in: Archives of Acoustics, vol. 38, pp. 439.
30. LILJENCRANTS J. (1967): *The OVE III Speech Synthesiser*. – in: Speech Transmission Laboratory Quarterly Progress Report, vol. 8, pp. 76-81.
31. MATOUŠEK Jindrich, ROMPORTL Jan (2006): *On Building Phonetically and Prosodically Rich Speech Corpus for Text to Speech Synthesis.* – in: Proceedings of the Second IASTED Conference on Computational Intelligence, pp. 442-447.
32. MATOUŠEK Jindrich, TIHELKA Daniel, ROMPORTL Jan (2008): *Building of a Speech Corpus Optimised for Unit Selection TTS Synthesis.* – in: Proceedings of 6th International Conference on Language Resources and Evaluation, pp. 1296-1299.
33. MERCADO Jordi Adell (2009): *Prosodic Analysis and Modelling of Conversational Elements for Speech Synthesis.* – Universitat Politècnica de Catalunya.
34. NOWAK Paweł Marcin (2006): *Vowel Reduction in Polish.* – University of California, Berkeley.
35. OHALA John (2011): *Christian Gottlieb Kratzenstein: Pioneer in Speech Synthesis.* – in: Proceedings of XVII ICPhS, pp. 156-159.
36. OLIVER Dominika (2007): *Modelling Polish Intonation for Speech Synthesis.* – Universität des Saarlandes.
37. PALO Perti (2006): *A Review of Articulatory Speech Synthesis.* – University of Helsinki.
38. PETERSON Gordon, WANG William, SIVERTSEN Eva (1958): *Segmentation Techniques in Speech Synthesis.* – in: Journal of Acoustical Society of America vol. 30(8), pp. 739-742.
39. PINKER Steven (1994): *The Language Instinct.* – New York: William Morrow.
40. PLUCIŃSKI Andrzej (2000): *Optimisation of Transcription Rules.* – in: Studia Phonetica Posnaniensia, vol. 6.
41. POHL Aleksander, ZIÓŁKO Bartosz (2013): Using Part of Speech N-grams for Improving Automatic Speech Recognition of Polish, Proceedings of 9th International Conference on Machine Learning and Data Mining.

42. PRZEPIÓRKOWSKI Adam, BAŃKO Mirosław, GÓRSKI Rafał, LEWANDOWSKA-TOMASZCZYK Barbara (2012): *Narodowy Korpus Języka Polskiego.* – Warszawa: PWN.
43. RABINER Lawrence (1989): *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* – in: Proceedings of IEEE, vol.77, pp. 257-286.
44. SCHÖTZ Susanne (2006): *F0 and Segment Duration in Formant Synthesis of Speaker Age.* – in: Proceedings of Speech Prosody Conference, pp. 515-518.
45. SIMON Herbert (1965): *The Shape of Automation (for Men and Management).* – New York: Harper and Row.
46. SUENDERMANN David, HÖGE Harald, BLACK Alan (2010): *Challenges in Speech Synthesis.* – in: CHEN Fang, JOKINEN Kristiina (eds.): Speech Technology – Theory and Applications. Springer Verlag, pp. 19-30.
47. STĂNESCU Miruna, CUCU Horia, BUZO Andi, BURILEANU Corneliu (2012): *ASR for Low-Resourced Languages: Building a Phonetically Balanced Romanian Speech Corpus.* – in: Proceedings of 20th European Signal Processing Conference, pp. 2060-2064.
48. STEFFEN-BATÓG Maria, NOWAKOWSKI Paweł (1993): *An Algorithm for Phonetic Transcription of Ortographic Texts in Polish.* – in: Studia Phonetica Posnaniensia, vol. 3. Poznań: Wydawnictwo Naukowe UAM.
49. SZKLANNY Krzysztof, OLIVER Dominika (2006): *Creation and Analysis of a Polish Speech Database for Use in Unit Selection Synthesis.* – in: Proceedings of 5th Language Resources and Evaluation Conference, pp. 297-302.
50. SZKLANNY Krzysztof (2009): *Optymalizacja funkcji kosztu w korpusowej syntezie mowy polskiej.* – Polish-Japanese Institute of Information Technology.
51. TAMBOR Jolanta, OSTASZEWSKA Danuta (2000): *Fonetyka i fonologia współczesnego języka polskiego.* – Warszawa: PWN.
52. TAYLOR Paul (2009): *Text-to-speech Synthesis.* – Cambridge University Press.
53. TRASK Robert Lawrence (2004): *What is a Word?* – University of Sussex.
54. TURING Alan (1950): *Computing Machinery and Intelligence.* – in: Mind, vol. 59, pp. 433-460.
55. WEGLARZ Geoffrey (2004): *Two Worlds of Data – Unstructured and Structured.* – in: DM Review Magazine.

56. VAN DEN BOSCH Antal, DAELEMANS Walter (1997): *Language-Independent Data-Oriented Grapheme-to-Phoneme Conversion.* – in: VAN SANTEN Jan, SPROAT Richard, OLIVE Joseph, HIRSCHBERG Julia (eds.): Progress in Speech Synthesis. Springer-Verlag, pp. 77-89.
57. YAROWSKY David (1997): *Homograph Disambiguation in Speech Synthesis.* – in: VAN SANTEN Jan, SPROAT Richard, OLIVE Joseph, HIRSCHBERG Julia (eds.): Progress in Speech Synthesis. Springer-Verlag, pp. 159-175.
58. YOUNG Steve (1996): *Large Vocabulary Continuous Speech Recognition: a Review.* – Cambridge University.
59. ZEN Heyga et al. (2007): *The HMM-based Speech Synthesis System (HTS) Version 2.0.* – in: Proceedings of 6th ISCA Workshop on Speech Synthesis, pp. 294-299.