ForumAcusticum
2 0 0 5

Phonetics and Speech Technology

Gábor Olaszy
Research Institute for Linguistics, Kempelen Farkas Speech Research Laboratory, Budapest, Hungary, olaszy@nytud.hu

Is there a need to apply phonetics in speech technology development? How can phonetic thinking influence the quality of the final product (synthesised speech, speech recognition)? What happens if phonetic aspects are not used? What branches of phonetics are used, and what is to be used in speech technology? How phonetic thinking can be embedded into the development procedure of a speech technology product (the compromise between technical demands and acoustical quality)? What kind of phonetic research may be interesting for the forthcoming speech technology?

1 Introduction

Speech science, phonetics, has witnessed revolutionary changes in the 20th century due to the continuously developing experimental possibilities. Another new branch of speech science, speech processing, has also developed rapidly from the middle of the 20th century [1]. These two branches have been integrated into a new science field, speech technology, which has made rapid progress in the last two decades of the 20th century. The question is, whether the classical phonetic science should be applied in modern speech technology research and application development or not. The answer is rather yes, if a proper, interdisciplinary dialogue can be formed and realized between engineers (exact science) and phoneticians (human science) [2]. If a "bridge" can be constructed between the two disciplines, the interdisciplinarity may result in scientifically better-backed speech technology solutions (the speech quality of speech synthesisers will be better, the recognition rate will be higher as well, special demands in speech technology, such as applications for medical and rehabilitation fields, will be solved with more success).

2 Differences in thinking

Phoneticians regard speech as the verbal tool of human communication. Speech serves as a way of expressing ourselves in the given language, using the biological mechanism of human speech production. This mechanism comprises two main production factors, glottal activity, and articulation, both controlled by the brain. At the glottal level, voice is determined by the physiological properties of the vocal cords (measures, tension etc.). At the level of articulation several factors determine the final product (the moving speed of the articulatory organs, the configuration of the tongue, the jaw, and the uvula, the level of humidity in the mouth etc.). While the brain controls only the main configuration functions during speaking (voiced, unvoiced, lip rounding etc.), the movement details of the mechanism are determined by the person (voice timbre is hard or silky, articulation is precise or not, speech speed is fast or slower, speech is produced from reading or spontaneous etc.). Thus, these changes will be present in the produced speech as well. Therefore, **speech is always unique and personal**, even if the same sound or sound sequence is re-pronounced. A speaking person can-not produce exactly the same speech wave twice. This fact makes speech individual and human. As to the language side of it, phoneticians regard speech as the verbal product of the language at the segmental and suprasegmental levels. The segmental one incorporates the basic building items: the production of speech sounds, the sound combinations, the specific sound durations, the specific sound intensity levels and the sound timbre. The suprasegmental level contains the prosody parameters (speech melody, accent distribution, rhythm, emotion). In sum: phoneticians regard speech mainly from the point of view physiology and the system of language.

Application engineers, who design speech based technology for automatic information systems, regard speech as an acoustic wave and associates this wave mostly with the written form of it, the text. The wave can be characterised by physical data such as intensity, duration and spectral content. The written form, the text, represents both the letters corresponding to the speech sounds and the word forms defined by the language.

2.1 Fundamental differences

One of the basic contradictions between the two conceptions is that while, on the one hand, text is a two level discrete representation (certain number of letters (graphemes) are used, certain word forms are formed from the graphemes defined by the language, and the words are separated by spaces), speech, on the other hand is a continuous acoustic product. Speech sounds are the result of the continuous articulation movements they have no discrete acoustical content (Fig.1). The

words are not separated inside the speech wave either there are not "spaces" between them. In short, a speech wave cannot be derived directly from the written text, however, this is what engineers try to accomplish generally.

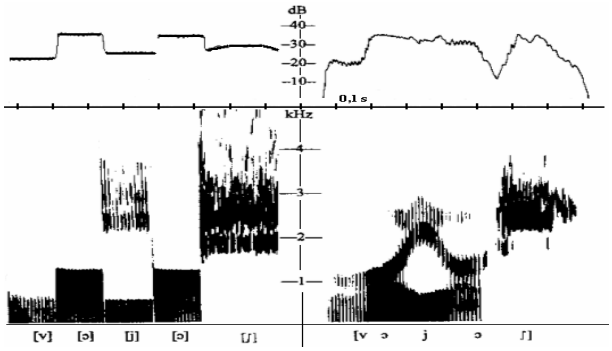

Figure 1: Left: imitation of speech as the series of discrete sounds of the word [v ɔ j ɔ ʃ] 'buttery'; right: the normal pronunciation of this word

The perceptual system of the mind accepts (or understands) certain types of acoustic events that may correspond to a certain letter or a group of letters in the text. This correspondence is defined by the perceptual basis of one's mother tongue. Speech sounds have their basic acoustical content when pronounced separately, as single sounds. In real speech the acoustic content of the sounds may be changed in the function of the surrounding sounds. Three theoretical parts are characteristic of a speech sound, the transient parts at the beginning and at the end, and the central part, which is always the most similar to the basic acoustic content of the sound. Thus, the acoustic content of the same sound may change depending on the preceding and following sound. In the process of hearing we do not realise these small acoustic changes, but they are important from the point of view of understanding continuous speech. The transition phases incorporate already the forthcoming next sound; they indicate it for perception. Therefore it is advisable to apply sound sensitive parts in the analysis algorithms as well.

The other basic difference between engineers and phoneticians contradiction is that engineers measure the spectral content of the speech wave precisely, however they do not generally take into consideration the articulation and the linguistic background, that belongs to the waveform.

3 Without phonetics?

Without using phonetics, the voice of a speech synthesiser may be more unnatural, the speech recognition rate of a recognizer may be lower. As to other application fields, such as teaching speech technology and phonetics, or giving speech technology support to certain medical fields (disorders in speech production, speech therapy) the technological solution cannot be designed successfully without phonetic aspects, i.e. it is important to integrate the phonetic knowledge with engineering and vice versa.

Let us take some examples. In speech synthesis, the pronunciation of numbers may be important in many cases, for instance, the number readers in Interactive Voice Response (IVR) applications). The example is taken from an existing bank information system. The task is well defined for an engineer: produce voice from written numbers (e.g. 37724 = thirty-seven thousand seven hundred and twenty-four). The engineer takes the text form of the number, defines about 25 different number elements (one, two, three, hundred etc.) and designs a wave concatenation system for these elements. The wave forms are recorded as single items (the announcer reads the numbers as 1, 2, 3, ….10, ….100 …). The completed system will be capable of pronouncing any number, but the quality of the synthetic voice will be very low, different from the natural human pronunciation (see on Fig. 2).

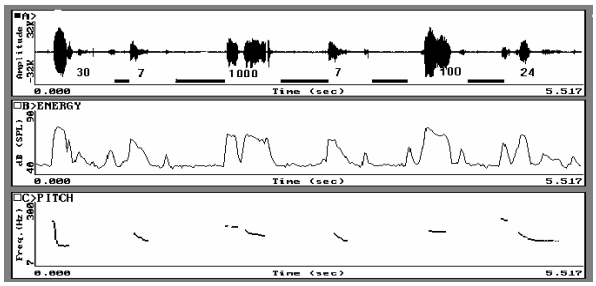

Figure 2: The number "37724" concatenated from single pronounced speech items (wrong conception)

It can be seen that the pronunciation is not continuous (horizontal thick lines show pauses). The energy level of the number elements do not correspond to the human pronunciation (e.g., the element "1000" has higher intensity than the "7"), the melody of the uttered number is unnatural (there are jumps in the Fo curve).

4 If phonetician is involved

As an attempt to solve these problems, a phonetician was then involved in the project. The phonetician has designed the number elements again, taking into account the important processes is speech production: continuity in the spectral content and Fo function, the realization of the necessary rhythm, inserting pauses only at the desirable places, and realising the right intensity levels characteristic of continuous pronunciation. The result was a number reader [3] that produced natural sound quality identical to a human

announcer (Fig. 3). What was the price of this quality rise? Totally new technology? No. Only the number of wave elements was increased to more than 200 different items (the basic technology, i.e. concatenation of stored wave forms did not change) and the concatenation algorithm became more complicated (not only the structure of numbers were taken into consideration, but the proper realization of the spectral continuity, the rhythm and the melody).

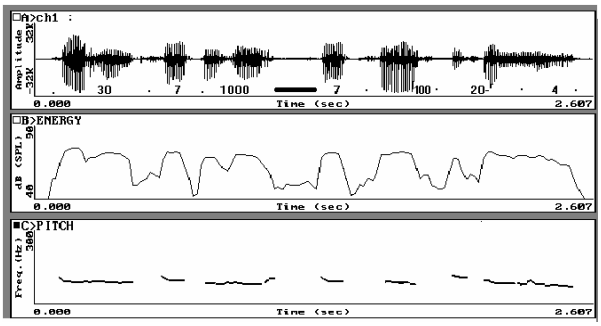

Figure 3: The number "37724" concatenated by the phonetically designed system

Comparing the pronunciation time of the two samples one can see that, in the second version the time has been reduced from 5.517 s to 2.607 s as a result of continuity and the correct rhythm. This number reader was embedded into an experimental IVR system. Perceptual tests confirmed that the users could not make any difference between the quality of the human pronounced prompt (bold) and the synthesised number (italic): **The balance of your account is:** *37724* **Forints**. The system was put into real application 6 years after its basic design, since system providers could not be convinced of the importance of the good acoustic quality, that is, the necessity of creating a good acoustic image for the company [4].

5 Sounds and features

Both in speech technology and in phonetics, in many cases, it is needed to define the sound characters and the boundaries of speech sounds. There can be a variety of reasons and goals: the automatic recognition of a sound from a sound sequence, the precise selection of cutting points in a corpus based speech synthesizer, to measure sound durations, to define the number of sounds in a certain part of speech, or to give an educational presentation about the articulation and its acoustic projection. The development of special measuring tools to support medical decisions may also need to define sound boundaries. The engineer can apply digital speech processing algorithms to detect (more or less) the physical changes in the speech wave (voiced/unvoiced detection, silent parts, energy and spectral changes, speech/no speech detection etc.). The results given by these procedures are sometimes enough for making further decisions (e.g., in certain speech recognition tasks), but in many cases they are not precise enough. In that case, phonetic research results have to be used to improve the algorithm. Such results are, for example, the knowledge of sound concatenation properties and the realization of language level items in the acoustic wave, the perception of certain acoustic forms, the explanation of the articulation positions and their acoustic projection, etc. In order to obtain more precise results by software tools, language dependent parts have to be added to the general purpose digital processing techniques that take into consideration the different sound realisation of the language, the behavior of sounds in the function of surrounding sounds and their types (VV, CV, VC, CC, CCC combinations) and in addition, the higher level prosody properties. For example, a voiceless affricate consists of two main parts: a closure (silent phase) and a voiceless fricative component. This structure is valid in VCV combinations. However, there are certain CC combinations (Fig. 4) in which the closure part is not realised, only the fricative part is. If there is no silence part, theoretically, it should be a spirant, but it is perceived as an affricate.

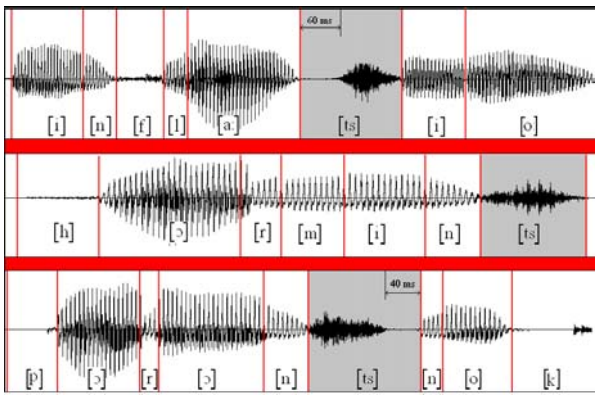

Figure 4: The realization of the affricate [ts] in Hungarian words. The total realization (the closure and the fricative parts) in the the VCV position (above 'inflation'), the lack of the closure part in certain CC clusters (middle, 'thirty') or totally reversed structure (the fricative component precedes the silent part) in a CCC combination (down, 'commander')

The most surprising fact is that, in certain CCC combinations, the usual structure of the affricate is reversed totally. First, the fricative component is present, followed by the silence period (Fig. 4. down). Despite this change, the sound is perceived as an affricate. In sum, the acoustic feature of certain sounds is determined by the sound environment. Detailed knowledge of such cases may help the algorithm making the right decision.

6 Internal sound structures and the articulation movements

Phoneticians say that a speech wave contains everything that happens during speech production in the physiological mechanism. Both in speech synthesis and in speech recognition it is important to define the characteristic acoustic categories in order to make more precise sound selections. If such categories can be identified the target of the decision algorithm may be narrower, the time consuming comparison with the possible candidates may be reduced. Let us take now the C1-V-C2 combinations as most common building elements of the speech wave. A phonetic theorem says that the articulation place of a C has its own acoustic projection (in the function of V). This fact defines the formant movements in the function of time in the vowel during the transient phase. If we take the parallelism that this transient phase is the result of certain articulatory movements, it can be said that every intermediate position of the articulatory organs (between the two target points) has its acoustic projection in the transient phase of the vowel. Thus the articulatory movements can be predicted from the acoustic content and vice versa. This predictability of movements can be equally well used in the analysis of speech disorders, in prediction of sounds during speech recognition, or by selection of cutting points in a corpus based speech synthesis system). Figure 5 shows the articulation places for Hungarian consonants. It can be said that, in this table, each every row represents an identical articulation configuration from where the formant movements of the joining vowel begin to move in the transient phase.

	stops							affricates				fricatives							nasals			laterals	
	[b]	[p]	[d]	[t]	[ɟ]	[c]	[g]	[k]	[ʦ]	[ʣ]	[ʧ]	[ʤ]	[v]	[f]	[z]	[s]	[ʒ]	[ʃ]	[j]	[h]	[m]	[n]	[ɲ]	[l]	[r]
bilabial	☒	☒																			■				
labio-dental													☒	☒											
denti-alveolar			☒	☒					☒	☒					☒	☒						■		☒	☒
alveolar											☒	☒					☒	☒						☒	☒
palatal					☒	☒													☒				■		
velar							☒	☒																	
laryngeal																				■					

Figure 5: The articulation configurations of the Hungarian consonants

From our point of view the denti-alveolar (9 consonants), the alveolar (6) and the palatal (4) rows are interesting, because in these rows are the most of the consonants (17). So the CV or VC combinations realised with these 17 consonants can be sorted into 3 acoustic categories. The theorem says that the vowels joining to any consonant that belongs to the same row in this table share a similar transient phase. This means that consonants of the same row (having the same source function) in a CV combination can be changed and the perceptual quality of the sound sequence will be as good as it was before with the previous consonant. Let us take an example. If the [a s a] sound sequence represents the original speech item, the [a ts a], the [a t a], and even the [a r a] item can be produced using sound surgery methods on consonants, as witnessed by Figure 6.

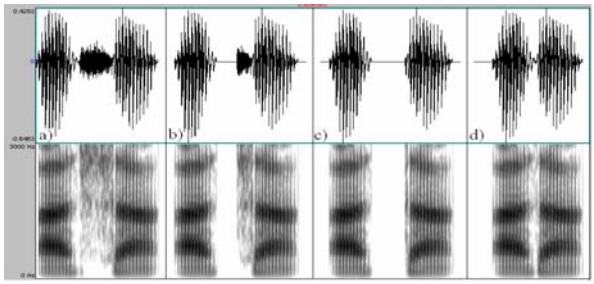

Figure 6: The acoustic content of the transient phase of the [a] vowels as the projection of the denti-alveolar articulation point of the [s], [ts], [t] and [r] consonants (see the VCV sequences in: a), b), c) and d)

The figure shows the following types of acoustic content: a) the original VCV sequence; b) the case where the amplitude is reduced in the first part of the [s] into zero, in other words, a silent period is produced, the fricative consonant is changed into an affricate and this is heard in the VCV item; c) the case where we reduce the amplitude in the first 90% of the consonant and a stop [t] is heard; d) the case where we cut out the whole [t] sound (i.e. we bring close to each other the two vowels, and an [r] consonant is heard in the CVC item. The detailed explanation of this case d) is the following. The Hungarian [r] is apical and voiced. The most characteristic parameter of the sound (in addition to the place of articulation, the denti-alveolar point) is the intensity minimum at the middle of the sound (when the apex of the tongue touches the denti-alveolar part). In our case the consonant of the original [a s a] sequence was formed at the denti-alveolar point, the intensity of the first vowel was decreasing because of the source change from voiced into unvoiced. The case was the same with the second vowel as well. The intensity was gradually increased from the beginning of the vowel. Thus, if we put close to each other (10-20 ms) of the wave forms of the two vowels, all the requirements for the production of an [r] sound will be met and an [r] consonant is heard in this VCV sequence. In sum, the articulation point is really reflected in the acoustic content. Similar examples can be shown concerning the rest of the rows of the table in Figure 5. Making use of this phonetic categorization the production of speech building unit inventories of speech synthesizers can be made more optimal. In addition the searching algorithm in speech recognition systems can also be formed in a more optimal manner.

Also, diagnosing disorders in articulation can be better supported by studying the acoustic representation of the speech sequences.

7 Hearing loss detection

Speech technology can be used to support certain medical fields as well. In the late 80-ies of the 20th century a Hungarian phonetician invented a new method for the detection of hearing loss using synthetic speech [5]. Hearing loss is similar to filtering. The method is based on the following idea: if we know the frequency components of speech sounds and sound combinations, and if we know the characteristic of the filter used, then the change in the speech wave form may be predicted. For example if we eliminate the high frequency F2 formant from a high vowel, the characteristic of it will be changed (due to the filtering) into another vowel. By means of selecting of well-defined speech items, the working of the speech perception (auditory) mechanism (from 300 Hz to 8000 Hz) can be measured. The procedure is simple and can be well applied to small children. Monosyllabic words (with governed acoustic content) are transmitted into the ear of the child by a headphone, and he/she is asked for to repeat the heard word (the procedure is performed as a play with words). If the auditory mechanism is adequate then the original word is given as an answer. If some hearing loss is present (the filtering function is realised) then the answered word will have different sounds. Giving only ten words, the measurement of the whole frequency band can be covered. The imitation of hearing loss and the change of the meaning of the speech item is shown in Fig. 7.

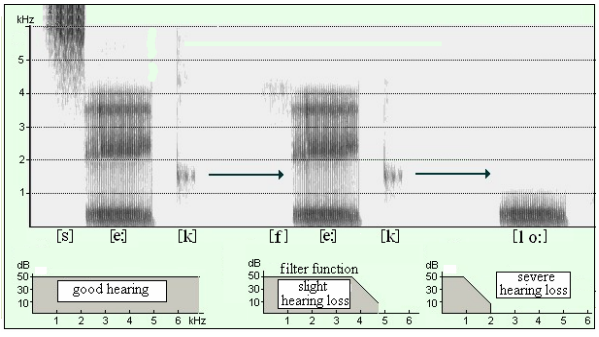

Figure 7: Example about the method for measuring hearing loss by synthetic speech

The explanation of the example on Figure 7 is as follows: the Hungarian word [s e: k] 'chair' has three sounds. The articulation place of the first consonant is the denti-alveolar point. This point defines the formant movements in the next vowel (as defined on Fig.5). The most important frequency components of the [s] consonant are high (about 6000 Hz). If the hearing loss filters out these high frequency components, the [s] sound will loose its basic frequency components and will be changed into [f]. If the hearing loss is more severe, then this first consonant cannot be heard at all. The repeated sequence will begin with the [e:] vowel, i.e. [e: k]. The more severe the hearing loss is the more parts of the higher frequency components are filtered out from the [s e: k] word. In the case of very severe hearing loss the answer may be [l o:]. In this case only the first formant of the vowel will be present, and it represents an [o:] like sound. But what about the [l]? The transient phase of the original [e:] vowel was produced under the influence of a denti-alveolar voiceless consonant, the [s]. According to the table in Figure 5 the [s] and the [l] has the same influence on the following vowel. The transient part of the vowel in the [s e:] sequence and in the [l e:] one is identical. Thus the [l o:] answer indicates a severe hearing loss.

8 Conclusion

The paper demonstrates that phonetic knowledge is indispensable in speech technology. The integration of engineering with traditional phonetics is a way to improve the quality of speaking and speech recognition systems. Integrating engineering and phonetics is promising in terms of creating more precise analysis tools for phoneticians, the two disciplines can profit from mutual help. New research results in phonetics, as in the field of the acoustic structure of spontaneous speech, as well as the research of emotional characteristics of speech may give new support for speech technology in the future.

References

[1] L. R. Rabiner – R.W.Schafer, '*Digital Processing of Speech Signals'* Prentice –Hall Inc. New Jersey ISBN 0-13-213603-1 (1978)

[2] K. Stevens, '*Acoustic Phonetics*'. The MIT Press, Cambridge, ISBN 0-262-19404-X (1998)

[3] G. Olaszy – G. Németh 'IVR for Banking and Residential Telephone Subscribers Using Stored Messages Combined with a New Number-to-Speech Synthesis Method' In: Human Factors and Voice Interactive Systems. Ed.: D. G.Bonneau. Kluwer Academic Publishers, pp.237-256. (1999)

[4] G. Németh, 'Acoustic Company Image and Telecommunications Services. In this proceedings (2005).

[5] M. Gósy, 'Synthesised speech for evaluation of children's hearing and acoustic-phonetic perception' In: Human Factors and Voice Interactive Systems. Ed.: D. G. Bonneau. Kluwer Academic Publishers, pp.123-135. (1999)