BMBF

Verbmobil
Verbundvorhaben

A Mixed Inventory Structure for German Concatenative Synthesis

Thomas Portele
Florian Höfer
Wolfgang Hess

Universität Bonn

Vm **Report 149**
25. September 1996

25. September 1996

Thomas Portele
Florian Höfer
Wolfgang Hess

Universität Bonn
Institut für Kommunikationsforschung und Phonetik
Poppelsdorfer Allee 47, 53115 Bonn

Tel.: (0228) 73 - 5638
Fax: (0228) 73 - 5639
e-mail: wgh@ikp.uni-bonn.de

Gehört zum Antragsabschnitt: 4.1

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 D 08 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

Published in *Progress in Speech Synthesis,* ed. by Jan Van Santen, Joe Olive, Richard Sproat, and Julia Hirschberg. New York: Springer; to appear in November 1996.

A Mixed Inventory Structure for German Concatenative Synthesis

Thomas Portele
Florian Höfer
Wolfgang Hess

ABSTRACT In speech synthesis by unit concatenation a major point is the definition of the unit inventory. Diphone or demisyllable inventories are widely used but both unit types have their drawbacks. This paper describes a mixed inventory structure which is syllable oriented but does not demand a definite decision about the position of a syllable boundary. In the definition process of the inventory the results of a comprehensive investigation of coarticulatory phenomena at syllable boundaries were used as well as a machine readable pronunciation dictionary. An evaluation comparing the mixed inventory with a demisyllable and a diphone inventory confirms that speech generated with the mixed inventory is superior regarding general acceptance. A segmental intelligibility test shows the high intelligibility of the synthetic speech.

1 Introduction

Demisyllables [PS60] and diphones [KW56] are the two main paradigms in the design of inventories for concatenative speech synthesis. Both have their advantages and their drawbacks. The demisyllable paradigm claims that coarticulation is minimized at syllable boundaries and only simple concatenation rules are necessary. However, this assumption is only partially valid for German. Heavy coarticulation effects occur at syllable boundaries (for instance nasal/lateral plosive releases in "Mitleid" /mɪtl̯ɪt/ or "abmachen" /apmaxən/, or initial devoicing in "Aufsatz" /a̯ʊfzats/ or "Hausbau" /ha̯ʊsba̯ʊ/) [Ko90]. The diphone model, on the other hand, assumes that coarticulative effects occur only between adjacent sounds in a domain limited by the centers of the pertinent sounds. Coarticulation is largely planned [Wh90], and the resulting phenomena (for instance lip rounding) may extend over several segments. The partial invalidity of the diphone assumption led Olive [Ol90] to augmenting a diphone inventory with additional units.

Our previous synthesis system HADIFIX [Po92] used a combination of initial demisyllables, diphones (vowel to postvocalic consonant transitions) and suffixes (postvocalic consonant clusters). In our own experience, effects such as the ones described above could not be adequately modeled with HADIFIX, and this difference from natural speeech not only degrades the naturalness but also the intelligibility of the synthetic speech. Moreover, we had serious problems with postvocalic sonorants synthesized by two units meeting in the middle of the sound. A formal evaluation of segmental intelligibility [Po93] revealed that unit boundaries inside postvocalic sonorants lead to undesirable spectral jumps and to misperceptions of the pertinent sounds. These experiments convinced us that neither diphones nor demisyllables are sufficient for high quality synthesis of German utterances.

2 Investigating Natural Speech

2.1 Material

German phonotactics is governed by syllable-domain phenomena, final devoicing being just one example. Therefore, a syllable-based inventory definition [Fm75] is appropriate. However, two phenomena must be taken into account.

- The number of syllables in German is very large. Demisyllables [PS60] can be used instead [FML77], but vowel reduction in unstressed syllables might require a number of additional demisyllables.
- Coarticulation at syllable boundaries must be treated.

Regarding the first point there is some evidence that a complete set of reduced demisyllables is not necessary [DC91] [Po94]. In the experiment described here we examined the second point, namely coarticulation at syllable boundaries.

All possible consonant combinations were studied with an in-between syllable border, and also, if not prohibited by German phonotactics, within the syllable in onset or coda. The items were extracted from 48 specifically designed passages read by a male and a female speaker (the latter was also the speaker for the female voice of our synthesis system). About 4200 items were investigated, 2100 from each speaker. For each sound combination the range of possible realizations was determined, and a standard realization was established as well as hypercorrect and reduced forms. Temporal aspects were also investigated.

2.2 Devoicing

Voiced fricatives and plosives are often devoiced when preceded by an unvoiced obstruent, regardless of their position (see Figure 1). With the ex-

FIGURE 1. Different levels of devoicing for fricatives extracted from fluent speech (left - completely devoiced; middle - partially devoiced; right - voiced). The arrows mark the fricatives.

ception of /z/ which becomes /s/, the devoiced sounds are distinguishable from their unvoiced counterparts: devoiced /b,d,g/ are less aspirated and have a weaker burst than /p,t,k/ [St71] [Ke84], a devoiced /v/ has less intensity than an /f/. The liquids /l,r/ are also subject to devoicing with /r/ more than /l/ [Ko77]. For liquids, the position of the syllable boundary determines the degree of devoicing. A syllable-initial /l/ preceded by a /k/ (as in "weglaufen" /vɛklaʊ̯fən/) is less likely to be devoiced than an /l/ in the second position of the onset (as in "Wehklage" /ve:kla:gə/). Nasals are only devoiced when preceded by a homorganic stop in reduced position.

2.3 Assimilation

Assimilation usually involves manner or place of articulation. Plosives are often assimilated to a following nasal, lateral, or fricative. The combination /tʃ/ is produced with a retracted /t/ to match the place of articulation of the /ʃ/ [Wa60]. The stop in the combinations /pm/ and /tl/ (and in similar pairs) is released by lowering the velum or lifting the tongue blades, the articulatory gesture to produce the following sound [MM61]. In all these cases the two sounds are effectively articulated together and cannot be separated. They are frequent in German, and a syllable boundary is often present between the two sounds (see Figure 2).

FIGURE 2. Stop-nasal combinations extracted from fluent speech. The arrows mark the nasal releases. As with plosive-plosive combinations, the stop is not released when both sounds are homorganic (middle and right).

2.4 Position of the Syllable Boundary

Within-syllable combinations are more likely to be reduced. Devoicing is more common, and stops are less aspirated. It is generally held that a /t/ after a syllable-initial /ʃ/ in words like "Drehstuhl" /dre:ʃtu:l/ is unaspirated, while a syllable-initial /t/ in the same context in a word like "Tischtuch" /tɪʃtu:x/ is aspirated [Fm79]. This difference is statistically significant (U-test, $p < 0.05$), however, among the data there are aspirated stops in the second position of the onset as well as unaspirated syllable-initial ones. The realization depends on the syllable's prominence [Ko90].

2.5 Pre- and Postvocalic Consonants

Differences between pre- and postvocalic realizations of a phoneme can be observed, especially for liquids; a vowel with a following postvocalic liquid is often produced like a diphthong [Ra36], and, sometimes, postvocalic liquids are just indicated by a slight change in the formants of the preceding vowel (and an elongated duration) [He79]. In general, the preceding vowel has a strong influence on liquids and nasals and also on dorsal obstruents. Intervocalic consonants bear more resemblance to their prevocalic counterparts, at least in the acoustic domain investigated here (see Figure 3) [SKT84] [Bo88].

FIGURE 3. Context-dependent rounding of the /ʃ/ and its acoustic manifestation as lowered formant-like structure between 1.5 and 2 kHz. The two leftmost examples show an intervocalic /ʃ/ that is articulated compatible to the following vowel. The four rightmost examples demonstrate the influence of the syllable boundary on the combination /ʃl/ (left - after /ʃ/; middle left, middle right and right - before /ʃ/) The thick line inside the spectrogram denotes the formant-like strucure afflicted by lip rounding.

2.6 Conclusions

A unit inventory based on a phonological syllable definition [Tw38] will inevitably lead to unnatural and hypercorrect synthetic speech. On the other hand, there are noticeable differences between phoneme realizations in the coda compared to those in the onset. A syllable based inventory structure is appropriate for the synthesis of German speech as long as a "syllable" is defined in a phonetical and not phonological way. However, there are doubts whether such a definition can be accomplished [Ko77]. The solution proposed here is designed to avoid the explicit placement of a syllable boundary between two nuclei.

3 Inventory Structure and Concatenation Rules

3.1 Concatenation Methods

There are three ways to concatenate two units. Their differences are exemplified by the construction of the /n/ in /ana:/:

- *Diphone concatenation* takes place inside a sound. A sound suitable for diphone concatenation must have some kind of stable part insensitive to contextual influences. The /n/ in /ana:/ is synthesized by combining units /an/ and /na:/ in the middle (or a suitable place [Kr94] [CI94]) of the /n/.
- *Hard concatenation* is the simplest case of putting two sounds together. This happens at each syllable boundary in standard demisyllable systems. In the example /ana:/ the units /a/ and /na:/ are concatenated.
- *Soft concatenation* also takes place at segment boundaries. However, the concatenation is smoothed by including the transitions that would appear in natural speech. A certain overlap between the two units is necessary. As anticipatory coarticulation is assumed to be more important than persistent coarticulation the overlap is primarily provided by the first unit. The /n/ in /ana:/ is generated by the units /an/ and /na:/, but, unlike diphone concatenation, the concatenation takes place at the beginning of the /n/. While the transition to the nasal and the anticipatory lowering of the velum is preserved inside the first /a/, the realization of the intervocalic /n/ is determined by the following vowel. Furthermore, the concatenation point is located at a point of change in the spectrum. Spectral jumps at this point are not as offending as inside a supposedly stable part of speech.

3.2 Inventory Structure

The mixed inventory structure (MIS) is designed to avoid hard concatenation whenever possible. Inside a vowel diphone concatenation is performed, while soft concatenation is the method of choice whenever consonants are involved. Seven types of units are used:

1. *Initial demisyllables* (1086 elements) with voiced and unvoiced versions for the initial consonants that can be devoiced. Here, "initial demisyllable" means "sequence of consonants that share articulatory gestures with a following vowel." Consequently, units like /pni:/ or /tlu:/ are included.
2. *Final demisyllables* (577 elements). These units are like classical demisyllables but without voiceless fricatives (except /x/ and /ç/) or combinations of obstruents. Due to the German syllable final devoicing

rule only voiceless obstruents appear in a coda. These sounds are synthesized by soft concatenation of suffixes (see below).

3. *Suffixes* (88 elements) consisting of voiceless fricatives and obstruent combinations in rounded and unrounded versions.

4. *Consonant-consonant diphones* (167 elements) to smooth problematic syllable boundaries and to allow synthesizing words that do not obey the phonotactics of German.

5. *Vowel-vowel diphones* (67 elements) to smooth transitions between vowels whenever the glottal stop is missing or eliminated due to reduction.

6. *Syllables or syllable parts containing syllabic consonants* (122 elements) because context dependency was found to be especially prominent in such syllables. A pilot study with 15 of these units has confirmed the perceptual salience of such units [Po94].

7. *Syllables with /ə/ and voiced initial plosives* (75 elements). These syllables are very common prefixes in German, and diphone concatenation inside the /ə/ is especially difficult because of the context dependency of the /ə/.

3.3 Inventory Definition

The exact inventory definition was achieved by analyzing a machine readable pronunciation dictionary with more than 90,000 entries. In this dictionary, inflected forms appear only when irregular. Due to the flexible suffix concept, even the most complicated German inflections like /rpsts/ in "Herbsts" /hɛrpsts/ or /mpfst/ in "kämpfst" /kɛmpfst/ are supported whereas, in a genuine demisyllable system, every possibility will have to be covered by a special unit. Many foreign words even with nasal vowels or non-syllabic vowels are supported, for instance "Chance" /ʃãsə/, "Szene" /stse:nə/, "Dschungel" /dʒʊŋəl/ or "Skorpion" /skɔrpĭo:n/. The complete inventory consists of 2182 units, which is in the order of standard inventories for German (about 2700 diphones in the German version of the CNET synthesizer [Bo93], and more than 2000 demisyllables in a demisyllable inventory [KA92]).

3.4 Concatenation Rules

The complex structure of the inventory requires a complex set of rules describing the unit selection. Only the major principles are described here (see [Po94] for a complete explanation of the concatenation). A phonemic string like /ʔapmʊrksən/ is converted into a list of synthesis units in two steps:

- For each syllable nucleus an environment is determined. If no special nucleus (e.g. a syllabic consonant) is present, the longest matching initial demisyllable and the longest matching final demisyllable are selected. Final obstruent clusters are represented by a list of suffixes that are combined using soft concatenation. The principle is to maximize the environment for each nucleus with no respect to any phonologically defined syllable.

  In our example, /ʔapmʊrksən/ the following syllable nucleus environments are obtained:

  1. /ʔap/ with /ʔa/ and /ap/
  2. /pmʊrks/ with /pmʊ/, /ʊr(k)/ (the /k/ is deleted) and /ks/
  3. /ksn̩/ with syllabic [n̩] as one unit (the /ə/ is deleted to simulate its elision due to reduction)

- The nucleus environments are concatenated. If there is an overlap between two adjacent environments, the pertinent sounds of the first environment are deleted. This ad-hoc rule assumes the validity of the maximum-onset principle. The transition to the following (deleted) sound is still present in the last sound of the first nucleus environment. The results described in Section 2, however, indicate that such a simple procedure may sometimes fail. Work is in progress to determine the best places for concatenation by minimizing the spectral distances in the overlapping areas and by appropriate rules describing phenomena related to the position of the syllable boundary, such as lip rounding (Figure 3) or aspiration. In our experience, in more than 80% of all cases there is an overlap of at least one sound.

  If two nucleus environments meet without overlap, two possibilities exist.

  1. Hard concatenation is allowed. This holds, for instance, when the second environment begins with a glottal stop.
  2. Hard concatenation is prohibited. For instance, concatenating /n/ and /g/ is not permitted because of the partially velarized /n/ before /g/ in natural speech. This phenomenon is instead modeled by the use of an intermediate diphone (diphone concatenation). The appearance of a sound pair in a special diphone table determines whether hard concatenation between these sounds is allowed or not.

  In the unlikely case (less than 1%) that there are sounds not represented by one or the other of two adjacent nucleus environments, a set of "exception diphones" is defined. These units, in part extracted from other units, allow synthesizing every possible combination of German sounds.

In our example, the following will happen:

1. /ʔa(p)/ and /pmʊrks/ - the /p/ from the first environment is deleted
2. /pmʊr(ks)/ and /ksn̩/ - the /ks/ suffix from the first environment is deleted

The final result is then:

1. Initial demisyllable /ʔa/
2. Final demisyllable /a(p)/ (/p/ is deleted in this unit)
3. Initial demisyllable /pmʊ/
4. Final demisyllable /ʊr(k)/ (/k/ ist deleted in this unit)
5. Syllable with syllabic consonant /ksn̩/

Whenever a vowel is involved, diphone concatenation takes place. Everywhere else soft concatenation is performed.

4 Perceptual Evaluation

To check whether the new mixed inventory structure MIS meets our expectations, two perception experiments were carried out: a pair comparison test with standard inventory structures, and a segmental intelligibility test.

4.1 Pair Comparison

A test set of 22 words was used. It mimics the properties of the German sound architecture to a large extent, i.e., sound frequency, sound number per word, and syllable number per word. Demisyllables, diphones, and mixed units necessary to synthesize these words were defined, spoken by a male speaker, recorded with 32 kHz sampling rate, and segmented by hand. A complete diphone method was assumed, i.e., all sound pairs are represented by their own unit. We decided to treat combinations of a long vowel and a following /ɐ/ as two sounds. This decision is questionable [Ko77] but reduces the number of units considerably. Moreover, as our results indicate (see below), at least the combinations of a vowel and a following /l/ must receive the same treatment. The demisyllable definition was adopted from [KA92]. Altogether, 400 units were generated. Whenever possible, "unit-sharing" was performed insofar as, for instance, the diphone /na/, the demisyllable /na/ and the MIS unit /na/ were extracted from the same utterance. Prosodic manipulations were carried out using the TD-PSOLA-algorithm [MC90]. The original temporal structure was used, and all versions of a word had the same intonation contour. A pair comparison

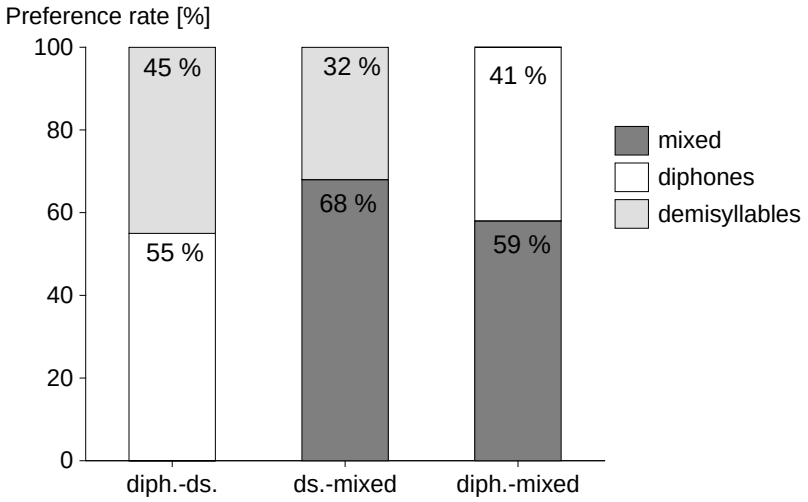

FIGURE 4. Results of a pair comparison between the mixed inventory, a demisyllable inventory, and a diphone inventory.

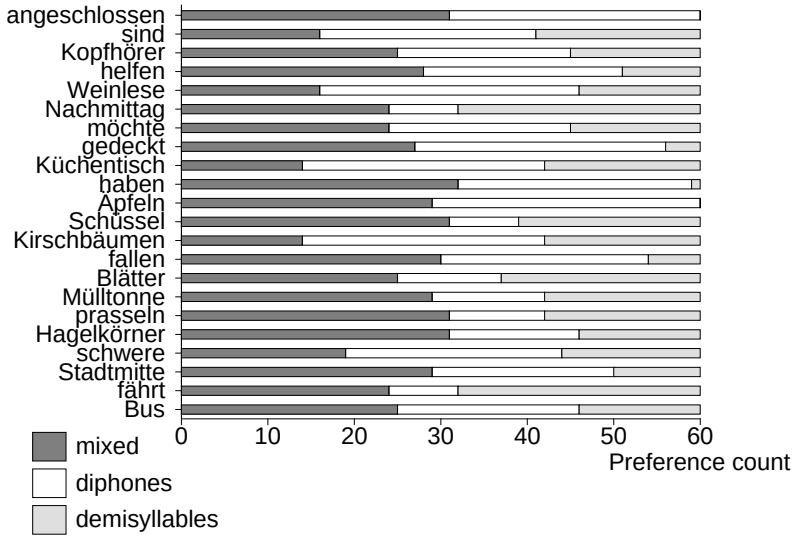

FIGURE 5. Results of a comparison between the mixed inventory, a demisyllable inventory, and a diphone inventory. The preference number is displayed for each test word.

FIGURE 6. Comparison of two versions of the word "fährt" /fɛ:ɐt/ (left - diphone version; right - demisyllable/mixed version). The dashed lines indicate positions where two units meet.

test was used to assess quality differences. Ten subjects participated. The outcome of the test (Figure 4) is a significant (χ^2-test, $p < 0.0005$) advantage of the new mixed inventory over demisyllables (68% vs. 32% preference rate), and over diphones (59% vs. 41% preference rate). A result in its own right is the preference of diphones to demisyllables (55% vs. 45%), which is also significant (χ^2-test, $p < 0.04$).

A separate analysis for each test word (Figure 5) shows that the diphone versions of some words performed poorly, especially for words with postvocalic liquids (for instance "fährt" /fɛ:ɐt/ (Figure 6) or "Mülltonne" /mʏltɔnə/), while the demisyllable versions of some other words with large coarticulatory effects across syllable boundaries had very bad preference values (for instance "angeschlossen" /angəʃlɔsən/ (Figure 7) or "Stadtmitte" /ʃtatmɪtə/). The mixed inventory structure produced no severe outliers and turned out to avoid the weaknesses of the standard paradigms.

4.2 Segmental Intelligibility

The segmental intelligibility was explored by the SAM segmental test [CG90, Po93]. The test assesses consonant intelligibility in CV, VC, and VCV contexts. All German consonants are combined with the vowels /a,i,u/. In the VC, the CV, and the VCV contexts 36, 48 and 56 items were included,

FIGURE 7. Comparison of two versions of the word "angeschlossen" /angəʃlɔsən/ (left - demisyllable version; right - mixed version). The dashed lines indicate positions where two units meet. The final syllabic nasal on the right picture could not be modeled with the demisyllable inventory used in this test.

respectively. The open response form allows analyzing the confusions between the sounds without external constraints. The synthesized stimuli were recorded on a DAT tape and presented over headphones. Eighteen subjects participated.

Figure 8 displays the results for the three contexts in comparison with the values obtained in an earlier investigation [Po93] for a human voice and the synthesis system HADIFIX [Po92]. There is a noticeable decrease in the error rates for the new synthetic voice with the mixed inventory (VC: 1.9%, CV: 6.0%, VCV: 6.9%) compared to the HADIFIX voice. Indeed, in the VC and the CV context the differences between human and synthetic voice are not significant (t-test, $p > 0.4$). In the VCV context, however, the error rate for the MIS voice is still three times as high as the one for a human voice. Possible sources of these errors for the MIS voice are:

- *VC context.* There is only one case of systematic misinterpretation. The stimulus /an/ was understood as /aŋ/ in one third of all responses. A more carefully pronounced unit could solve this problem. Otherwise, the synthetic voice is as intelligible as a human voice.

- *CV context.* More than 50% of all errors were caused by misunderstanding /hi:/ as /ti:/ or /çi:/, and /hu:/ as /bu:/, /tu:/, /ku:/, or

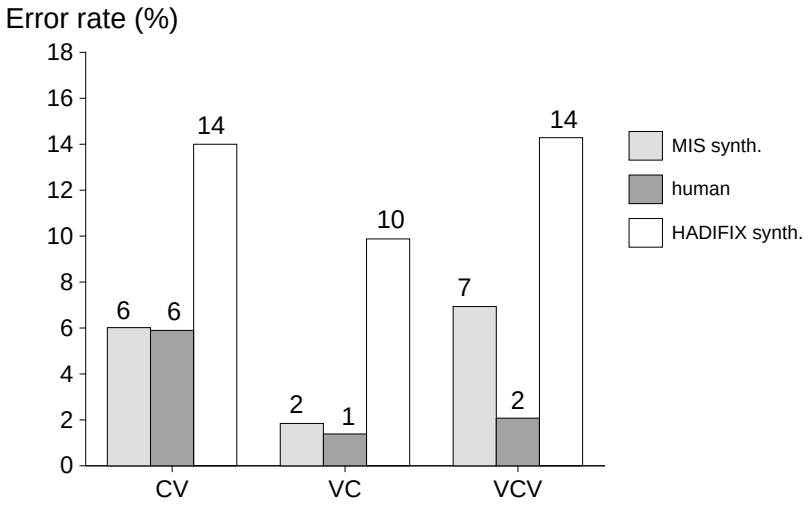

FIGURE 8. Segmental intelligibility of VC, CV, and VCV consonants. The mixed inventory (MIS) is compared to the HADIFIX inventory and to a human voice.

/fu:/. Similar results were obtained for the human voice. The high error rates for /h/ seem not to be caused by bad synthesis, but by difficult discrimination of /h/ with closed vowels; /ha:/ was always recognized correctly. Other error sources were /b/ (9.8% error rate), /ʃ/ (11% error rate), and /ç/ (25% error rate). Most of these errors occured in combination with /i:/. No other systematic problems were detected.

- *VCV context.* In this context, /h/ was also problematic (44.4% error rate). Voiced apical obstruents were perceived as voiceless (6.1% error rate). Nasals were sometimes confused (13.0% error rate), /ana:/ with /aŋa:/, /ini:/ with /imi:/, and, very often, /uŋu:/ with /unu:/. Synthetic intervocalic nasals and /h/ must be modelled in a better way to reach the segmental intelligibility of their natural counterparts. Intervocalic /h/ is, in fact, the only intervocalic consonant for which hard concatenation is performed, and this may be the primary source of the increased confusion rate compared to the other consonants.

The segmental intelligibility of the MIS inventory almost meets the standard set by the intelligibility of human voices (at least for this simple test under laboratory conditions).

5 Summary

This paper has described an inventory structure based on seven types of units. The definition was based on experiments investigating the relevant acoustic and phonetic facts. Elaborate concatenation rules allow the synthesis of natural sounding speech as confirmed in two different evaluations: a pair comparison test with diphones and demisyllables, and a segmental intelligibility test.

Acknowledgments: This research was supported within the language and speech project VERBMOBIL by the *Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie*, and by the *Deutsche Forschungsgemeinschaft*. We thank two anonymous reviewers for their helpful comments, Dieter Stock for his support, Volker Kraft for valuable discussions, and all the participating subjects for their patience.

6 References

[Bo93] O. Boeffard, B. Cherbonnel, F. Emerard, and S. White. Automatic segmentation and quality evaluation of speech unit inventories for concatenation-based, multilingual PSOLA text-to-speech systems. In: *Proceedings EUROSPEECH'93*, pages 1449–1452, Berlin, Germany, 1993.

[Bo88] V.J. Boucher. A parameter of syllabification for VstopV and relative timing invariance. *Journal of Phonetics* 16: 299–326, 1988.

[CI94] A. Conkie and S. Isard. Optimal coupling of diphones. In *Second ESCA/IEEE-Workshop on speech synthesis*, pages 119–122, New Paltz, U.S.A., 1994. ESCA.

[CG90] R. Carlson, B. Granström, and L. Nord. Segmental evaluation using the ESPRIT/SAM test procedures and monosyllablic words. In *First ESCA-Workshop on Speech Synthesis*, pages 257–260, Autrans, France, 1990. ESCA.

[DC91] R. Drullman and R. Collier. On the combined use of accented and unaccented diphones in speech synthesis. *J. Acoust. Soc. Am.* 90: 1766–1775, 1991.

[Fm75] O. Fujimura. Syllable as the unit of speech synthesis. Unpublished paper.

[FML77] O. Fujimura, M.J. Macchi, and J.B. Lovins. Demisyllables and affixes for speech synthesis. In *9th ICA*, page 513, Madrid, 1977.

[Fm79] O. Fujimura. An analysis of English syllables as cores and affixes. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 4/5: 471–476, 1979.

[He79] G. Heike. Prerequisites of speech synthesis on the basis of an articulatory model. *AIPUK* 12: 91–99, 1979.

[Ke84] K.A. Keating. Phonetic and phonological representation of stop consonant voicing. *Language* 60: 286–319, 1984.

[Ko77] K. Kohler. *Einführung in die Phonetik des Deutschen.* Erich Schmidt, Berlin, 1977.

[Ko90] K. Kohler. Segmental reduction in connected speech in German: phonological facts and phonetic explanations. In *Speech Production and Speech Modeling* (ed.: W.J. Hardcastle and A. Marchal), pages 69–92, Kluwer, Dordrecht, 1990.

[KA92] V. Kraft and J. Andrews. Design, evaluation, and acquisition of a speech database for German synthesis-by-concatenation. In *Proc. SST-92*, pages 724–729, Brisbane, Australia, 1992.

[Kr94] V. Kraft. Does the resulting speech quality improvement make a sophisticated concatenation of time-domain synthesis units worthwhile? In *Second ESCA/IEEE-Workshop on speech synthesis*, pages 65–68, New Paltz, U.S.A., 1994. ESCA.

[KW56] K. Küpfmüller and O. Warns. Sprachsynthese aus Lauten. *Nachrichtentechnische Fachberichte* 3:28–31, 1956.

[MM61] C. Martens and P. Martens. *Phonetik der deutschen Sprache.* Hueber, Munich, 1961.

[MC90] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9: 453–467, 1990.

[Ol90] J. Olive. A new algorithm for a concatenative speech synthesis system using an augmented acoustic inventory of speech sounds. In *First ESCA-Workshop on Speech Synthesis*, pages 25–30, Autrans, France, 1990. ESCA.

[PS60] G. E. Peterson and E. Sievertsen. Objectives and techniques in speech synthesis. *Language and Speech* 3: 84–95, 1960.

[Po92] T. Portele, B. Steffan, R. Preuss, W.F. Sendlmeier and W. Hess. HADIFIX - a speech synthesis system for German. In *Proceedings ICSLP'92*, pages 1227–1230, Banff, Canada, 1992.

[Po93] T. Portele. Evaluation der segmentalen Verständlichkeit des Sprachsynthesesystems HADIFIX mit der SAM-Testprozedur. In: *Fortschritte der Akustik - DAGA'93*, pages 1032–1035, Frankfurt, Germany, 1993. DPG GmbH.

[Po94] T. Portele. *Ein phonetisch-akustisch motiviertes Inventar zur Sprachsynthese deutscher Äusserungen.* Diss., Univ. Bonn, 1994.

[Ra36] K.M. Rapp. *Versuch einer Physiologie der Sprache nebst historischer Entwicklung der abendländischen Idiome nach physiologischen Grundsätzen.* Cotta, Stuttgart-Tübingen, 1836.

[SKT84] A.G. Samuel, D. Kat, and V. Tartter. Which syllable does an intervocalic stop belong to? A selective adaptation study. *J. Acoust. Soc. Am.* 76: 1652–1663, 1984.

[St71] D. Stock. *Untersuchungen zur Stimmhaftigkeit hochdeutscher Phonemrealisierungen.* Buske, Hamburg, 1971.

[Tw38] W. F. Twadell. A phonological analysis of intervocalic consonant clusters in German. In: *Actes du 4e congrès int. des linguistes*, pages 218–225, Copenhagen, Denmark, 1938.

[Wa60] H.-H. Wängler. *Grundriss einer Phonetik des Deutschen.* Elwert, Marburg, 1960.

[Wh90] D. H. Whalen. Coarticulation is largely planned. *Journal of Phonetics* 18: 3–35, 1990.