https://doi.org/10.3176/lu.1998.3.29

MEELIS MIHKLA, ARVO EEK, EINAR MEISTER (Tallinn)

CREATION OF THE ESTONIAN DIPHONE DATABASE FOR TEXT-TO-SPEECH SYNTHESIS*

Introduction

The aim of our text-to-speech synthesis project is to convert the Estonian written text, inserted orthographically into a computer, to an orthoepically correct and natural-sounding spoken text for a wide range of practical application.

Articulated speech flow does not consist of a simple concatenating string of speech sounds. Rather, speech is a continuously overlapping set of transitions from one speech sound to another. Due to the regressive coarticulation, generation of the previous segment contains features from the next speech sound(s). The minimal (elementary) articulatory gesture seems to be a demisyllable (Fujimura 1981) where the movement from a constriction or closed phase to an open phase or vice versa makes up a sonority cycle (Clements 1988). The so-called ahead articulation scans all muscle channels exploited in the generation of the respective minimal articulatory sequence and switches on simultaneously those channels, whose activity is not contradicting the basic state of the movement. On the other hand, the progressive coarticulation, based on the inertia of articulators, leaves traces of the previous sound in the following segment.

These coarticulation phenomena have been the main impediment in a way getting satisfactorily natural speech quality by means of formant synthesizers. Articulatory synthesizers need elaborate computational work to have wider practical applications. Nevertheless, both types of synthesizers are perspective and applicable in research purposes.

The demisyllabic CV- or -VC elementary (transitional) gesture is a relatively non-compressible segment of speech flow and is not subject to considerable durational changes (e.g. Fujimura 1982). The allophonic variants of phonemes generally arise from demisyllabic affiliations. Only the quasi-stationary part of vocalic and consonantal phases of CV- and -VC demisyllables respectively is subject to durational variation. In many languages, including Estonian, the variability of the quasi-stationary part is used to mark the short/long opposition.

Taking into account what was said above, it should not be surprising that a compilative resynthesis, based on concatenating diphones separated from natural speech, may lead to the best results in text-to-speech converting systems. A diphone synthesis has the advantage that the coarticulatory transitions which are controlled

* The support from the Estonian Science Foundation and Open Estonian Foundation has made the present work possible.

with difficulties by rules, are naturally comprised without losses in diphones separated from real speech. A diphone database needs less amount of memory than databases of syllables or words.

Cooperation with the MBROLA project

The MBROLA (multiband resynthesis overlap add) project was initiated by the TCTS Lab of the Faculté Polytechnique de Mons (Belgium). The aim of the MBROLA project is to obtain a set of speech synthesizers for as many languages as possible, free of use for non-commercial and non-military applications. MBROLA consists of a speech synthesizer, based on the concatenation of diphones, and of diphone databases (Dutoit 1997). Several languages are already available in the MBROLA homepage http://tcts.fpms.ac.be/synthesis/mbrola (e.g. Brazilian Portuguese, Breton, British English, Dutch, French, German, Romanian and Spanish).

In order to create a text-to-speech compilative synthesizer for Estonian, the work group consisting of the researchers of the Laboratory of Phonetics and Speech Technology of Institute of Cybernetics and the Institute of the Estonian Language joined the MBROLA project in 1997. Joining this project enables us to use the Mons MBROLA synthesizer (Figure 1: block 3) for concatenating diphones, matching them with each other, changing the duration and fundamental frequency of sounds. In order to convert an Estonian written text into synthesized speech we have to solve the following tasks: (1) to convert an orthographic text into phonetic-phonological (Figure 1: block 1); (2) to compile rules for the control of segment durations and F0 contours (block 2); (3) to compile a prosodic database with data concerning the durations and F0 contours of the segments (block 4); (4) to compile the diphone database (block 5). By today we have compiled the Estonian diphone database (about 1600 diphones; cf. e.g. the corresponding data for other languages: Spanish 800, French 1200, German 1800 diphones). In the nearest future an automatic control algorithm for converting an orthographic text into an orthographic-phonetic-phonological mixed system will be ready.

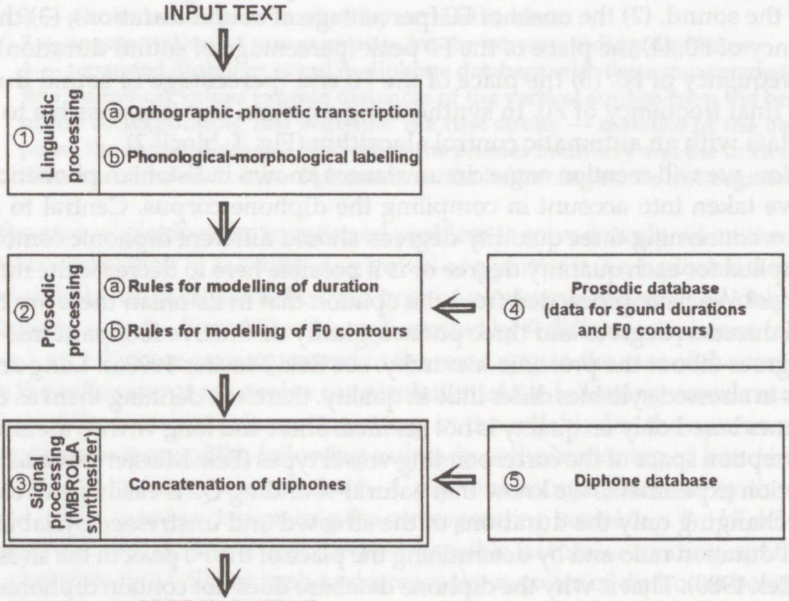
INPUT TEXT
①
Linguistic processing
ⓐ Orthographic-phonetic transcription
ⓑ Phonological-morphological labelling
②
Prosodic processing
ⓐ Rules for modelling of duration
ⓑ Rules for modelling of F0 contours
④
Prosodic database (data for sound durations and F0 contours)
③
Signal processing (MBROLA synthesizer)
Concatenation of diphones
⑤
Diphone database
SPEECH OUTPUT

Figure 1. **A text-to-speech compilative synthesizer based on diphones concatenation.**

Completion of the Estonian diphone database

Our text-to-speech synthesis project uses diphones as the elementary concatenating acoustic units, separated from spoken texts. Diphones are segments of the speech flow consisting of a part of two consecutive phones (sounds) (e.g. pause—consonant #—C, pause—vowel #—V, consonant—vowel C—V, vowel—vowel V—V, vowel—consonant V—C, consonant—consonant C—C, vowel—pause V—# and consonant—pause C—#). A diphone begins with the quasi-stationary part of the first phone and ends in the quasi-stationary part of the following one.

Creating a diphone database consists of the following procedures: (1) possible combinations of all vowels and consonants at the beginning, in the middle and at the end of words were determined taking the three quantity degrees into account; (2) based on these combinations a diphone register was created; (3) text corpus was created so that each diphone would occur in one word at least (the text corpus for Estonian diphones consists of 1270 words); (4) digital recording of the text corpus (the corresponding words were read in the frame sentence *ütlen ... taas* 'I am saying ... again' by one male speaker in an anechoic chamber); (5) speech segments (we call them raw segments here) were separated from the words of the text corpus so that before and after each diphone at least 50 ms of the corresponding sound was retained (see Figure 2); (6) segmenting diphones from the raw segments mentioned in the previous point (see Fig. 2; segmenting is a time consuming task: segmenting and labelling of a 1-minute speech takes approximately 1000 minutes); (7) standardizing the diphones (normalisation of intensities of diphones; made by colleagues from Mons); (8) optimisation of the diphone database (on the basis of listening tests it may turn out to be necessary to correct diphones, to decrease or increase the number of them).

In the diphone database (Fig. 1: block 5) a diphone is characterised in addition to its acoustical pattern by three characteristics which mark (1) the beginning of the diphone, (2) the end of the diphone and (3) the boundary of two phones (all measured from the beginning of the raw segment). In the prosodic database (Fig. 1: block 4) data on each sound in diphones are kept as follows: (1) the whole duration of the sound, (2) the onset of F0 (percentage of sound duration), (3) the initial frequency of F0, (4) the place of the F0 peak (percentage of sound duration), (5) the peak frequency of F0, (6) the place of the F0 end (percentage of sound duration), (7) the final frequency of F0. In synthesizing a certain text it is possible to change these data with an automatic control algorithm (Fig. 1: block 2).

Below we will mention some circumstances known in Estonian phonetics which we have taken into account in compiling the diphone corpus. Central to it is the question concerning three quantity degrees: should different diphonic combinations be compiled for each quantity degree or is it possible here to decrease the number of diphones? We have proceeded from the opinion that in Estonian there are two segmental duration degrees and three phonologically distinctive foot patterns — quantity degrees (about the prosodic hierarchy, see Eek, Meister 1998a). Long and short vowels in stressed syllables differ little in quality, therefore defining them as different phonemes based only on quality is not justified. Short and long vowels are situated in the perception space of the corresponding vowel types (Eek, Meister 1998b). From the perception experiments we know that natural-sounding Q2 is easily generated from Q1 by changing only the durations of the stressed and unstressed syllables to the needed duration ratio and by determining the place of the F0 peak in the stressed syllable (Eek 1980). That is why the diphone database does not contain diphones for the Q2 feet. They are generated automatically by rules (Fig. 1: block 2) based on the data concerning diphones from the Q1 foot. Synthesizing by rules the natural-sounding Q3 foot on the basis of Q1 or Q2 would be a too difficult procedure (Eek, Meister 1997 : 88—90) which would actually be infeasible with the current system due to the prominent reduction in quality, intensity and duration of the vowel of the unstressed syllable, and also due to the peculiarities in V—C transitions of the stressed syllables. This is why we considered it necessary to compile special V—C and C—V diphones for the Q3 foot.

Secondly a question should be mentioned: which consonant taken from a CV-demisyllable should be considered to represent the first part of a consonant in a diphone #—C so that the following syllable would be perceived as a natural-sounding integral unit? The perception experiments with plosives draw the attention to the fact that the syllable is perceived as an integral unit only in the cases when the distance between the strongest burst peak and the onset of the vowel formant representing F2′ does not exceed a critical distance. If the distance is larger than the critical distance, the syllable is perceived as a sequence of discontinuous segments (Eek, Meister 1996).

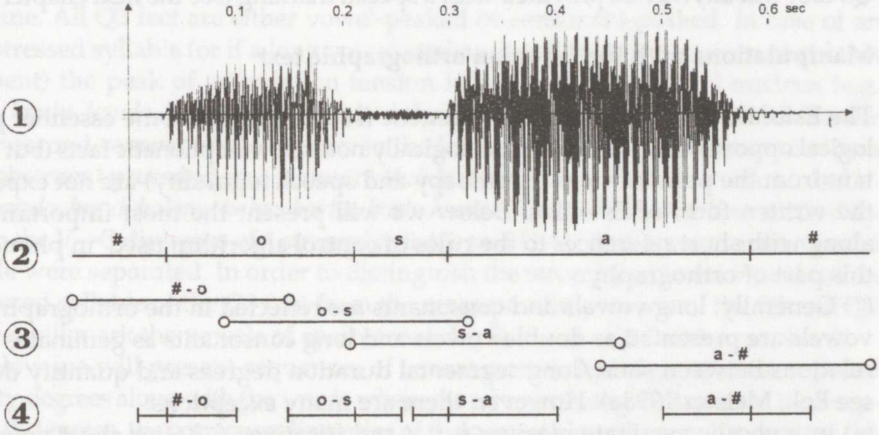

Figure 2. **An example of diphone segmentation (the word *osa* separated from the text corpus).**
1 — oscillogram of the word *osa* (spectrogram used in segmentation is not presented in the figure);
2 — phone boundaries marked by vertical strokes;
3 — segmentation of raw segments for the corresponding diphones;
4 — separated diphones saved in diphone database with three measurement values (the numerical values express distances of the vertical strokes from the beginning of the corresponding raw segment: the first stroke — distance of the beginning point, the second stroke — distance of the phones boundary and the third stroke — distance of the end of the diphone from the beginning of the raw segment).

In the case of plosives the mentioned problem is solved easily when the burst is added to the transition of the following vowel as a part of it (i.e. a plosive is then represented only by an occlusion). So each type of plosives is represented only by one #—C diphone (e.g. #—*k*) and C—V is represented by 9 different diphones (e.g. *k—i*, *k—e*, *k—ä*, etc.). But we cannot proceed in the same way in the case of the continuants because the influence of regressive coarticulation of the following vowel reaches the beginning of the word-initial consonant (e.g. in the syllable *ha-* the spectrum of *h* is similar to the spectrum of the following *a* already at the beginning of *h* and in *hi-* *h* is similar to *i*). That is why we mark the #—C diphone being influenced by the following vowel with a number designating the corresponding vowel (e.g. #—*h7*, *h7—a*; #—*h1*, *h1—i*, etc.). Our aim is to get the best possible natural-sounding synthesized speech, therefore, we will not rush to decrease the number of diphones before it has not been convincingly proved by perception experiments that taking the beginning of a consonant into account does not affect the quality of synthesized speech.

The special marking of word-initial continuants enables to distinguish them from the diphones of the unstressed syllable beginning with the same consonant (e.g. in the word *mõmin*: *#—m6, m6—õ, õ—m, m—i, i—n, n—#*). In such a way the acoustical structure of the unstressed syllable of a foot (i.e. of the unit essential in the identification of quantity degrees) is more adequately represented. There cannot arise a disagreement with the word-initial unmarked plosives because in Q1 and Q2 feet Q1 diphones V—*g* and *g*—V, V—*b* and *b*—V, V—*d* and *d*—V, separated from the Q1, will be used, i.e. the control algorithm directs, e.g., in the word *tuki* (Q2) the intra-word sequence of *ki*, (which otherwise could coincide with the word-initial *k—i* diphone) as well as the *uk* sequence to the diphones *g—i* and *u—g* of the word *tugi* (Q1); the algorithm changes the word-initial *g, b, d* occurring in foreign words into word-initial *k, p, t*; it is possible to use a word-initial C—V diphone from the Q1 foot as the corresponding C—V diphone for the Q3 foot without impairing the naturality, whereas the backmost diphones of the Q3 foot will anyway be provided with a special marking (see the next chapter).

Manipulations with the Estonian orthographic text

The Estonian orthography is not phonetic (see also EKG). Some essential phonological oppositions as well as phonologically non-relevant phonetic facts (but important from the point of view of orthoepy and speech naturality) are not exposed in the written form of Estonian. Below we will present the most important cases along with short references to the rules of control algorithm used in phonetizing this part of orthography.

(1) Generally, long vowels and consonants are reflected in the orthography: long vowels are presented as double vowels and long consonants as geminates (about relations between short/long segmental duration degrees and quantity degrees, see Eek, Meister 1998a). However, there are many exceptions:

(a) intervocalic geminate plosives *k, p, t* and fricatives *f, š* after short vowels are written by one letter in the Q2 foot, e.g. *tuki, tuti, tupe, tufi, tuši*; only in the Q3 foot these geminates are written by two letters, e.g. *tukki, tutti, tuppe, tuffi, tušši*;

(b) after long vowels or diphthongs, irrespective of the quantity degree of a foot, long (geminate) obstruents (except *s*, e.g. *poiss*, Q3) are written by one character, e.g. *saate* (Q2), *saate* (Q3), *laat* (Q3); to this group belongs also a geminate *h* in the Q2 foot;

(c) after a sequence of short vowel + sonorant, irrespective of the quantity degree of a foot, long (geminate) obstruents (except *s*, e.g. *varss*, Q3) are written by one character, e.g. *narta* (Q2), *karta* (Q3), *kart* (Q3);

(d) the same is valid for long (geminate) plosives in the Q2 and Q3 feet after long vowels and diphthongs, e.g. *kaarte*, (Q2), *kaarte* (Q3), *kaart* (Q3);

(e) plosive geminates when followed by a sequence of voiced consonant + vowel, is also written by one character, e.g. *rütmi* (Q2), *rütmi* (Q3), *kitli* (Q3).

All the above mentioned geminate obstruents in the Q2 feet written by one character do not need an extra marking. The intervocalic geminate plosives *k, p, t* written by one character will be directed in the linguistic processing block 1a (Fig. 1) to the diphones consisting of the corresponding short consonants *g, b, d* (e.g. *muki* → *mugi*). Passing untouched through the automatic diacritics rules in block 1b, *muki* will be changed in the prosodic processing block 2a, b according to the data of the diphones *u-g* and *g-i* from the word *mugi* in block 4 so that the duration of the occlusion phase of the corresponding short consonant will be doubled. This economical manipulation decreasing the number of diphones (about the argumentation, see above) cannot be used in the case of the geminates *f* and *š* written by one character. In these cases different diphones (e.g. *u—f* from *tufi* and *u—š* from *tuši*) have been recorded because Estonian lacks an intervocalic short (lenis) *f*, the short (lenis) *š* (orthographically *ž*) could be pronounced in a non-Estonian way as too voiced. Therefore after the transcription change *ž* → *š* in block 1a the short *š* (e.g. orthographically *ž* in *Kiži*) will be directed to the diphone of the corresponding geminate (e.g. *i—š* from *niši*) where in block 2a the stationary part of *š* will be shortened.

(2) Generally, the Q2 and Q3 opposition, a central phenomenon of the Estonian prosody, is not revealed in the orthography (e.g. *saada* (Q2) — *saada* (Q3), *laulu* (Q2) — *laulu* (Q3), *kalla* (Q2) — *kalla* (Q3), *kalmu* (Q2) — *kalmu* (Q3), etc.). The only exception is made by the intervocalic or word-final obstruents following a short vowel (e.g. *tugi* (Q1) — *tuki* (Q2) — *tukki* (Q3), *tukk* (Q3)).

Q1 and Q2 need no special marking. We will mark Q3 by a colon placed after the peak of the Q3 foot (Eek, Meister 1998a). The colon does not denote a Q3 phoneme but indicates that the whole foot is in Q3, and at the same time it signalises the potential but not obligatory duration increment (peakedness) of the preceding syllable-final phoneme. All Q3 feet are either vowel-peaked or consonant-peaked. In case of an open stressed syllable (or if a long vowel nucleus is followed by a sonorant or lenis obstruent) the peak of articulation tension falls on the long vowel nucleus (e.g. *saa:da, lau:lu, lau:da, kaa:rdu, pea:lse*). In a closed syllable the tension peak falls on the first or second consonant of the coda; in the latter case the peak consonant can be a fortis obstruent preceded by a sonorant (*tuk:ki, tuk:k, saat:a, laat:, laut:a, laut:, kal:la, hal:l, kar:da, kar:d, kal:mu, kal:m, kart:a, kart:, kaart:i, kaart:*). In the Q3 foot marked by a colon the V—C diphones of a stressed syllable and C—V diphones of an unstressed syllable were separated. In order to distinguish the strongly reduced vowels of an unstressed syllable of the Q3 foot from the corresponding vowels of the Q1 and Q2 feet, we will mark the vowels of an unstressed syllable of a Q3 foot by number 3.

Below we will present sequences of concatenative diphones of some types of quantity degrees along with the words where the corresponding diphone has been separated (* denotes the corresponding rules of the control algorithm and 2x denotes the lengthening of the duration):

sada		*saada*		*saa:da*		*laulu*		*lau:lu*	
#—s7	(sada)	#—s7	(sada)	#—s7	(sada)	#—l7	(laba)	#—l7	(laba)
s7—a	(sada)	s7—a	(sada* 2×)	s7—a	(sada)	l7—a	(laba)	l7—a	(laba)
a—d	(sada)	a—d	(sada)	a—a:	(laa:bu)	a—u	(lauda)	a—u:	(lau:da)
d—a	(sada)	d—a	(sada)	a:—d	saa:da)	u—l	(ulu)	u:—l	(kuu:la)
a—#	(iha)	a—#	(iha)	d—a3	(pat:ta*)	l—u	(ulu)	l—u3	(hul:lu*)
				a3—#	(tap:pa)	u—#	(uhu)	u3—#	(tip:pu)

maki		*mak:ki*		*saate*		*saat:e*		*auto*	
#—m7	(madu)	#—m7	(madu)	#—s7	(sada)	#—s7	(sada)	#—a	(ahi)
m7—a	(madu)	m7—a	(madu)	s7—a	(sada* 2×)	s7—a	(sada* 2×)	a—u	(lauda)
a—g	(kagu* 2×)	a—k:	(kak:ku)	a—d	(sada*, 2×)	a—t:	(pat:ta)	u—d	(pude* 2×)
g—i	(nõgi*)	k:—i3	(käk:ki*)	d—e	(pude*)	t:—e3	(pet:te*)	d—o	(pedo*)
i—#	(ahi)	i3—#	(top:pi)	e—#	(ehe)	e3—#	(tup:pe)	o—#	(Leho)

laut:a		*saa:d*		*mak:k*		*laat:*		*laut:*	
#—l7	(laba)	#—s7	(sada)	#—m7	(madu)	#—l7	(laba)	#—l7	(laba)
l7—a	(laba)	s7—a	(sada)	m7—a	(madu)	l7—a	(laba* 2×)	l7—a	(laba)
a—u	(lauda)	a—a:	(laa:bu)	a—k:	(kak:ku)	a—t:	(pat:ta)	a—u	(lauda)
u—t:	(rut:tu)	a:—d	(saa:da)	k:—#	(palk:*)	t:—#	(alt:)	u—t:	(rut:tu)
t:—a3	(pat:ta*)	d—#	(põl:d)					t:—#	(alt:)
a3—#	(tap:pa)								

(3) Distinctive palatalisation is not revealed in the orthography. In order to mark palatalisation we use an apostrophe (e.g. *pan'i — pan'ni — pan':ni*).
(4) It will be determined by the rules that the long syllable-final *üü* will be pronounced as *üi* (both in a Q2 and Q3 foot) if the following unstressed syllable begins with a short vowel (e.g. *püüa püia; lüü:a lüi:a*).
(5) The phoneme /n/ is realised as palato-velar nasal [ŋ] before palato-velar plosives (except in the case of morpheme boundary before *g, k*). To bring forth the exception we use a comma as the morpheme boundary (cf. e.g. *istungi — istun,gi*).
(6) The boundaries of the components of a compound word are marked with +.
(7) As the short intervocalic *h* has become voiced, in order to avoid unnaturality we cannot derive geminates by doubling the duration of a short consonant. For that purpose we exceptionally use different diphones.
(8) Word-initial *g, b, d* are changed into *k, p, t* in block 1a; *z* and *ž*, not depending on their position in a word, are changed into *s* and *š* respectively.

Adding diacritics (a colon, apostrophe, comma, plus sign) to the orthography manually will certainly give the best results. But then the general applicability of the synthesizer decreases because not everyone is able to add the additional marks. The soon-to-be-tested automatic system for adding diacritics (block 1b) will likely need to be developed by adding syntactical data. It will be dangerous to leave it half done because the heard defective synthesized speech (especially with an inadequate presentation of Q3 and palatalisation) could eventually lead to the inadequate pronunciation of the listeners, which in the current case would finnishize Estonian.

REFERENCES

Clements, G. N. 1988, The Sonority Cycle and Syllable Geometry. — The Sixth International Phonology Meeting, The Third International Morphology Meeting. Abstracts, Krems, 21.

Dutoit, T. 1997, An Introduction to Text-to-Speech Synthesis, Dordrecht.

Eek, A. 1980, Estonian Quantity: Notes on the Perception of Duration. — Estonian Papers in Phonetics, Tallinn, 5—30.

Eek, A., Meister, E. 1996, The Perception of Stop Consonants: Linking the Strongest Spectral Region of the Burst to the Following Vowel. — Fonetiikan Päivät, 28.—29. 08. 1996, Joensuu (In print).

—— 1997, Simple Perception Experiments on Estonian Word Prosody: Foot Structure vs. Segmental Quantity. — Estonian Prosody: Papers from a Symposium. Proceedings of the International Symposium on Estonian Prosody, Tallinn, Estonia, October 29—30, 1996, Tallinn, 71—99.

—— 1998a, Estonian Speech in the Babel Multilanguage Database: Phonetic-Phonological Problems Revealed in the Text Corpus. — Proceedings of the Workshop on Speech Database Development for Central and Eastern European Languages. The First International Conference on Language Resources and Evaluation, Granada.

—— 1998b, Quality of Standard Estonian Vowels in Stressed and Unstressed Syllables of the Feet in Three Distinctive Quantity Degrees. — LU XXXIV, 226—233.

Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., H. Saari, H. Tael, K., Vare, S., Eesti keele grammatika II. Süntaks. Lisa: kiri. Tallinn 1993 (= EKG)

Fujimura, O. 1981, Temporal Organization of Articulatory Movements as a Multidimensional Phrasal Structure. — Phonetica 38, 66—83.

—— 1982, Relative Invariance of Articulatory Movements. An Iceberg Model. — The XIIIth International Congress of Linguists. Tokyo, August 1982. Working Group on Speech Production, Tokyo.