SPEECH DYNAMICS

Louis C.W. Pols

Institute of Phonetic Sciences / ACLC, University of Amsterdam, the Netherlands
Louis.Pols@uva.nl

ABSTRACT

This keynote presentation is about the various dynamic aspects of speech (energy envelope, spectral variation, voicing and pitch variation, speaking style and pronunciation variation, and influence of communication channel). The related speech signal characteristics are measured and modeled and are tested in listening experiments. Consequences for speech recognition and speech synthesis are discussed. The Modulation Transfer Function can be used as a global measure for speech intelligibility and leads to interesting applications such as feedback in speech training. We wonder how TE-speakers and CI-users can handle speech dynamics properly.

Keywords: speech dynamics, speech acoustics, formant tracks, speech perception, speech training.

1. INTRODUCTION

In order for speech to be informative and communicative, segmental and suprasegmental variation is mandatory. Only this leads to meaningful words and sentences. The building blocks are no stable entities put next to each other (like beads on a string or like printed text), but there are gradual transitions from one to the other. This leads to what I call ***speech dynamics***. It is manifest in the energy envelope (loud vs. soft, gradual vs. abrupt onset), in spectral variation (CV and VC formant transitions, diphthongs), in voicing and pitch variation (word stress, sentence accent, question intonation and emotion), in speaking style and pronunciation variation (clear vs. sloppy, speech rate, reduction, deletion, insertion and emphasis), and it varies with the communication channel (noise, reverberation, filtering). For automatic speech recognition systems it is hard to properly interpret this variation under variable conditions, just as it is hard for speech synthesis systems to properly generate these speech dynamics in order to reach high intelligibility and greater naturalness.

In human speech production and perception interesting questions are, whether we run against the limitations of the articulatory system f.i. in fast speech, and whether we are able to make good use of local context, such as formant transitions, to optimize speech understanding. Spectro-temporal variation leads to a low-frequency modulation of the intensity spectral envelope in contiguous audio frequency bands. The resulting Modulation Transfer Function (MTF) appears to be a good predictor for speech intelligibility.

Speakers with a speech handicap, such as dysarthric speakers or tracheo-esophageal speakers (TE) can get effective feedback about their speech quality with measures derived from the MTF. This leads to interesting applications in rehabilitation programs. One also wonders how cochlear implant (CI) users can process speech dynamics often so well, despite the deviant speech processing.

These are some of the topics to be discussed in this paper. Most data presented here are based on work done with my PhD students.

2. DYNAMICS IS THE NORM

Dynamic variation is everywhere in the speech signal, from very global up to the finest segmental details. In speech, contrary to printed text, there generally are no clear word boundaries in the signal. In conversational speech, phonemes, syllables and complete words are even regularly deleted. For instance the city name ‘Amsterdam’ is often pronounced as /Ems@dEm/.

In Dutch the common 7-syllable regular expression ‘op een gegeven moment’ is frequently reduced to something like the 4-syllable utterance /Op@xemEn/, see Table 1 and [25]. Up to what

Table 1: Realizations of 474 Dutch face-to-face expressions of ‘op een gegeven moment’ in the 1000-hours CGN corpus [21].

utterance type	count	perc.	dur.(sd)
Op @n x@x’ev@(n) mom’Ent	0	0	-
Op @ (x@)x’ev@ mOm’En(t)	19	4.0	947(185)
(Op) (@) x’ev@m’En(t)	210	44.3	548(126)
Op (@) xem’En(t)	178	37.6	394(90)
Op (@) xev@	11	2.3	362(117)
inappropriate segments	56	11.8	-
Total	474	100	495(170)

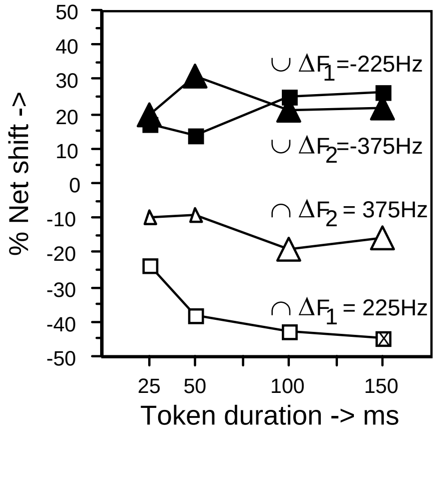
level this happens is a matter of word stress, speaking style and communicative conditions: The speaker speaks as sloppily as the listener allows him to do. Within and between words there is much reduction, coarticulation and assimilation. But one also frequently sees /r w j/ insertion to ease articulation, such as in the word combination ‘the spa[r] is closed’.

2.1. Formant tracks, coarticulation

One way to study coarticulation and reduction in a systematic way is to measure, to stylize and to model formant contours via n-th degree curve fitting, or via Legendre polynomials, or by using a fixed number of points per segment [9]. In his thesis Rob van Son [32, 33] studied formant tracks in an 850-words text, read at normal and fast rate. The seven most frequent Dutch vowels plus the schwa were analyzed. We were interested to see whether, especially in fast-rate speech, there was a tendency for duration-controlled undershoot as proposed by Lindblom [15]. The F1-F2 values in the vowel *centers* gave no indication whatsoever for that, there was only some overall rise in F1 for fast rate (for all vowels). Also the formant track *shape* (time-normalized to 16 points per segment) was the same for normal and fast rate speech, thus indicating that changes in vowel duration do not change the amount of undershoot. This is a strong indication of an active control of articulation speed, at least for this trained speaker [28].

Lindblom and Studdert-Kennedy [16] proposed a ‘compensatory perceptual overshoot’ model based on vowel identification experiments with synthetic /wVw/ and /jVj/ stimuli. We also studied this claim by using stimuli with systematically manipulated formant track shape, duration and context. We compared the vowel identification of the *level* stimuli (stationary F1 and/or F2) with those for *dynamic* stimuli with a formant onglide and/or offglide. As can be seen in Fig. 1, the responses always shifted towards the on/offset of the curved formant track, thus showing a ‘perceptual undershoot’ instead! This ‘weighted perceptual averaging’ appeared to emphasize more the final offglide. Since in the ultimate case of free conversational speech, some form of perceptual overshoot is active all the time, we suggest that this is the result of higher level linguistic processing and complex normalization mechanisms. For more details see [28].

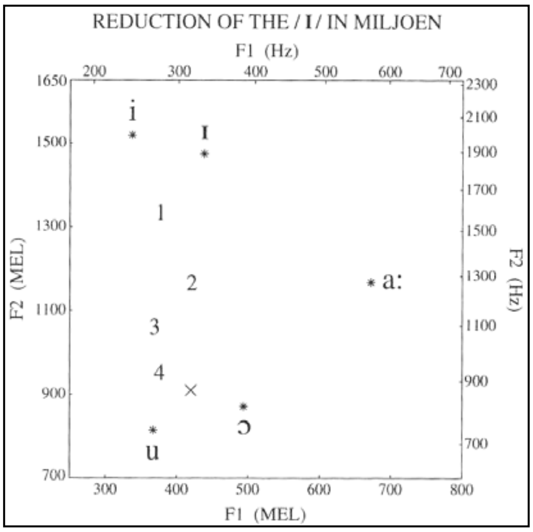
Figure 1: Percentage net shift in vowel identification for dynamic vs. stationary stimuli.



2.2. Formant tracks, vowel reduction

Whenever vowels are reduced because of a more sloppy speaking style and/or faster speech, it is generally supposed that vowel formants are more centralized, thus becoming more similar to the central vowel schwa. When data are averaged over many utterances, this is clearly shown in various languages, see for Dutch f.i. Koopmans-van Beinum [14]. However, Dick van Bergem [2] showed that, if one looks more carefully at individual utterances, spectral vowel reduction can much better be interpreted as the result of increased contextual assimilation of the vowel with surrounding phonemes rather than as a tendency to centralize.

Figure 2: F1-F2 realizations of the vowel /I/ in the Dutch word ‘miljoen’ /mIljun/ by 20 male speakers. These realizations are grouped in four clusters labeled 1 to 4. 20 Listeners have also judged all these 20 /I/ vowels as a full vowel or as a schwa. The percentage schwa-responses for cluster 1 to 4 were 5, 36, 60 and 69%, respectively. The cross indicates the position of the schwa for this /m-l/ context according to the schwa model presented below.

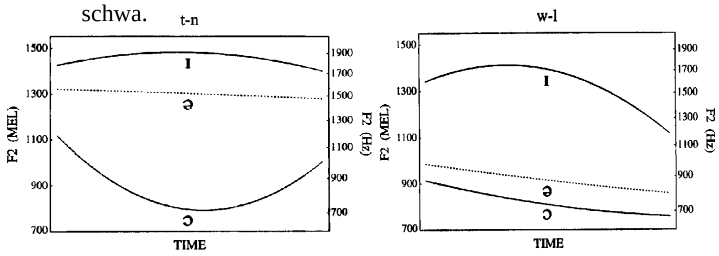


Thus, a reduced /I/ in the Dutch word 'miljoen' /mIljun/ becomes more /u/-like (see Fig. 2), whereas a reduced /O/ in the Dutch word 'bioscoop' /bijOskop/ becomes more /I/-like.

Van Bergem also *modeled* the coarticulatory effects of the schwa in open and closed syllables with a second-order polynomial, taking into account the effect of C_1, C_2, and V in nonsense words of the type $C_1@C_2V$ and $VC_1@C_2$, spoken by three male Dutch speakers. Especially the large variation in F2 is modeled very well: about 90% of the variance is explained.

Fig. 3 illustrates the fact that the schwa contour is rather straight from onset to offset, but even more so that its position highly varies with local context. For instance the final velar /l/-allophone (right panel) has a tremendous lowering effect on F2 for the schwa. All in all we can conclude again that the schwa is not just a centralized vowel but something completely assimilated with its phonemic context.

Figure 3: Stylized F2-contours for /I/ and /O/ in the Dutch words 'tin' and 'ton' (left panel) and 'wil' and 'wol' (right panel) plus the model prediction for schwa.



2.3. Consonant reduction

Articulatory consonant reduction has been reported regularly, although mainly limited to a few classes of consonants, like plosives, see f.i. Sussman et al. [37]. We tried to compare the acoustic consequences of vowel and consonant reduction. We did this for two speaking styles: spontaneous and reading aloud. An experienced Dutch male newscaster spontaneously told some stories and anecdotes. This speech was then transliterated and, after some time, read by him aloud, resulting in roughly 20 minutes of spontaneous and 20 minutes of read speech. From this material 791 VCV-pairs were selected for further analysis. Five aspects of vowel and consonant reduction were measured. Two of them related to coarticulation: 1) F2 slope difference between VC- and CV-boundary, 2) F2 locus equation (F2 onset vs. F2 target). The other three related to speaking effort: 3) duration, 4) spectral center of gravity COG (mean frequency) and 5) intervocalic sound energy difference.

Average F1-F2 positions of the 12 Dutch vowels clearly showed vowel reduction for spontaneous speech. For most consonants the F2 slope difference was lower in spontaneous speech, indicating a decrease in articulation speed. We found no systematic effect in F2 locus equation, indicating that vowel onsets and vowel targets change in concert. This implies that vowel reduction is mirrored by a comparable consonant change. The vowel and consonant durations were shorter in spontaneous speech, while also the COG and the intervocalic sound energy difference were lower, indicating a decrease in vocal and articulatory effort. For many more details see [35].

2.4. Diphthongs in pc1-pc2 representation

Apart from the dynamicity in CV- and VC-transitions, there is of course also much vowel-internal variation (see f.i. [3]), especially in diphthongs and diphthongized long vowels, such as /Au Ei 9y/ and /e: o: 2:/ in Dutch.

This was systematically studied by Irene Jacobi [10, 11]. Her project was initiated by the observation of lowering (getting more /a/-like) of (the onset of) the Dutch diphthong /Ei/ by many young well-educated females, this variant was called Polder Dutch or avant-garde Dutch. Our sociolinguistic study was based on segments from spontaneous speech from 70 speakers (balanced as good as possible in terms of sex, age, regional background and level of education) selected from the existing Dutch corpus CGN [21]. Although formant analyses were also performed, most analyses were based on principal component analyses (PCA) of bark-filtered spectra. This analysis method has proven over the years to be very efficient [e.g. 23, 24, 27]. The pc1-pc2 plane itself was derived from a PCA on all average /a/, /i/ and /u/ bandfilter spectra of all 70 speakers.

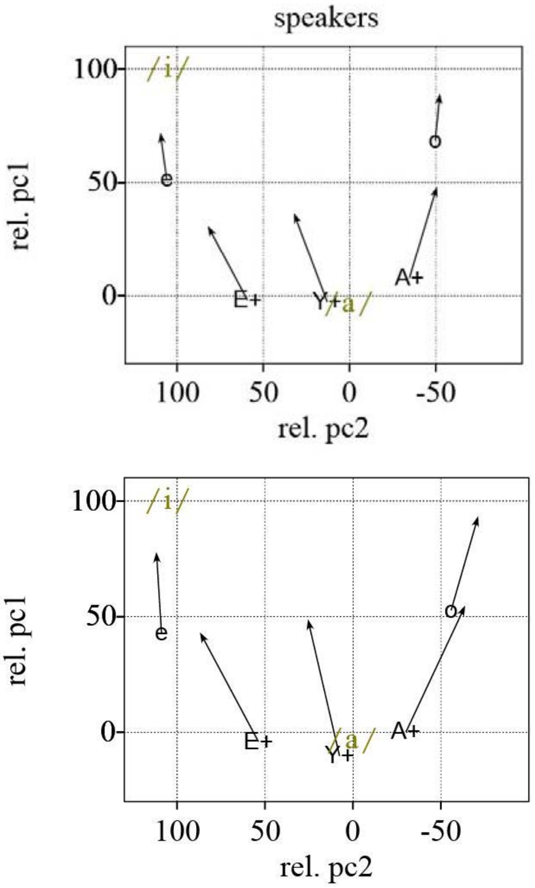
We chose to express all measured *onset lowering* and *amount of diphthongization* relative to the position of the individual corner vowels /a i u/. This way we applied a nice way of speaker normalization and it appeared to make our measurements also less sensitive to the noisiness of many live recordings in CGN. Actually the percentage onset lowering per realization was expressed as the pc1 (first principal component) distance to /a/ relative to the pc1-distance between

/a/ and /i/. A small or even negative value thus represents much lowering. Similarly the percentage degree of diphthongization was expressed as the pc1-distance between onset and offset (at 10 and 90% of the segment duration) relative to the pc1-distance between /a/ and /i/. A 100% score thus represents much dynamics, much diphthongization. It will be clear that there will be much individual variation in the data. Still there were various significant effects, but because of the large number of factors, the picture is rather complicated.

Some of the main findings are: With our way of analyzing and presenting the data, there were no significant differences between male and female data. The higher-educated speakers showed more lowering and more diphthongization than the lower-educated speakers for all vowels, see Fig. 4.

By grouping the continuous parameter *age* into three discrete *age groups* (young (<36 yrs.), mid (36-54), and old (>54)), several more interesting effects showed up. The factor ‘level of education’ appeared to be the most regular main effect, followed by the factor ‘age group’. When split to age groups, significant patterns of change in pronunciation between generations were found for the higher-educated generations, most salient for /o:/ and /e:/. For many more interesting details, see [10]. There, also a same-different perception experiment is described that compares speaker realizations of vowels within words. It is shown that the most salient spectro-temporal differences in the vowel realizations are indeed perceptually relevant. Also the age of the *listeners* appeared to have an influence upon the judgment of the (dis)similarities of the vowel pronunciations of the *speakers*.

Figure 4: Relative pc1-pc2 values of the mean vowel pronunciations of the 35 lower-educated (top panel) and the 35 higher-educated (bottom panel) Dutch speakers. The latter show more lowering and more diphthongization. By definition /a/ has the position (0,0) and /i/ the position (100,100).



3. PERCEIVING SPEECH DYNAMICS

By using a measure of listener confusion based on the perplexity of the confusion matrix, it is possible to quantify the amount of information extracted by the listener from different parts of the speech signal. This allows us to measure the importance of local (dynamic) context for consonant and vowel identification. The vowel and consonant identification study presented here was based on 120 CVC speech fragments taken from a long Dutch text read aloud by a male speaker (see also sect. 2.1). Only fragments containing vowel realizations with a duration of 100 ms or longer were used. Fig. 5 specifies for one of the CVC fragments (/Sa:l/) the various tokens used for vowel and consonant identification.

Figure 5: Various tokens (with labels like CV and TCC and median durations) for V and C identification.

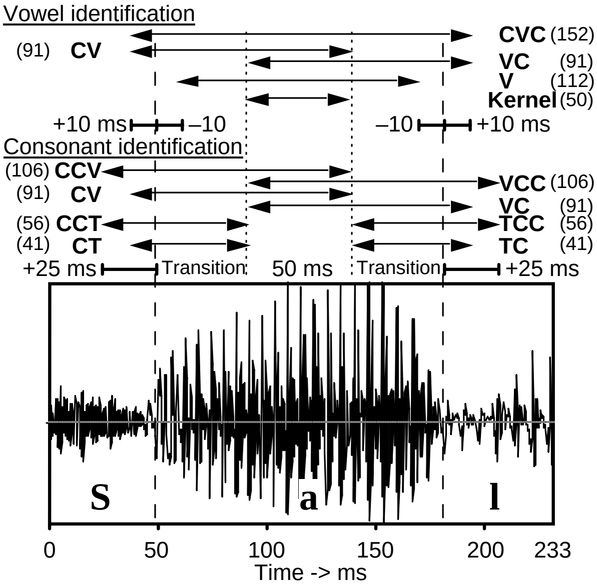
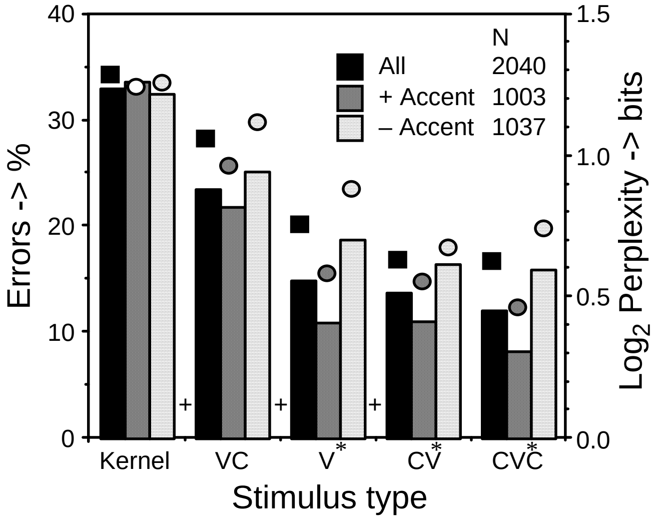


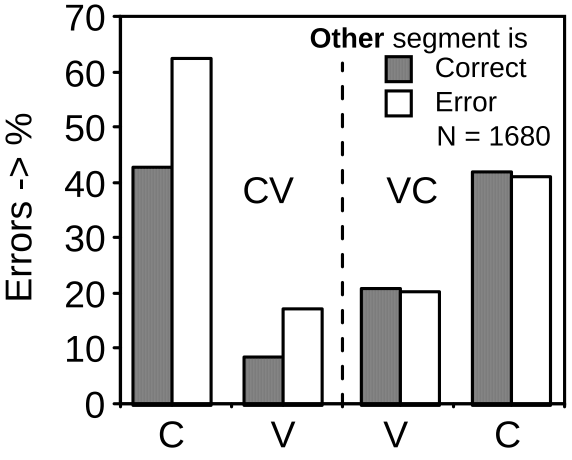
Figure 6: Results of the V identification experiment.



We ignored voicing errors in obstruents, since these are quite weak in Dutch, as well as long/short errors in vowels. Fig. 6 shows the error rate and the $\log_2$ perplexity of the vowel identification of the various tokens. All data pooled are presented, as well as those for accented and unaccented separately. The central 50 ms of the vowel tokens (kernel) elicited quite high error rates (33%). When more of the vowel realizations and their context were included, the differences between the error rates of accented and unaccented vowels grew. The error rate in the longest condition (CVC) gets as low as 12%.

Human subjects can extract around half of the information needed to identify a phoneme from only a short, 50 ms, fragment of running (read) speech from the ‘segment proper’. However, this is only enough for a ‘rough’ first guess and nowhere enough for correct identification. Further speech from the transitions and neighboring segments is needed to refine this first guess sufficiently for correct identification. A considerable fraction, around a third, of the information needed for perfect identification could very well be extracted from beyond the neighboring segments. All of this points for perception of connected speech towards the importance of cues that are not localized inside the phoneme proper and it supports the case of Nearey’s [20] proposed double weak theory of phoneme perception. For more details and the consonant error rates, I refer to [34]. Fig. 7 shows that the correctness of identification of either C or V in the CV tokens (but not in the VC tokens) was dependent of the correctness of identification of the other phoneme in the same token.

Figure 7: Error rates for V and C identification in CV- and VC-type tokens with respect to the (in)correct identification of the other segment in the same token.



4. PERCEIVING SPEECH-LIKE TRANSITIONS

Psycho-acoustics has provided us with excellent means to study detection thresholds, difference limen (DL), just-noticeable differences (jnd), similarities, matching, etc. In the past, such tests were only done on simple stationary signals such as pure tones. This learned us f.i. that the jnd of a 1000-Hz tone is about 1.5 Hz. Later on also stationary multi-harmonic single-formant-like periodic signals were used, showing us that the jnd for F2 is some 3-5% and that the jnd for formant bandwidth is some 20-40%! One starts to wonder what the jnd might be of formant transitions. Astrid van Wieringen [39] was one of the first to study this systematically for Dutch.

Figure 8: Difference limen as a function of transition duration for short speech-like transitions, from tone glides to single and complex formant glides, both as onset and as offset.

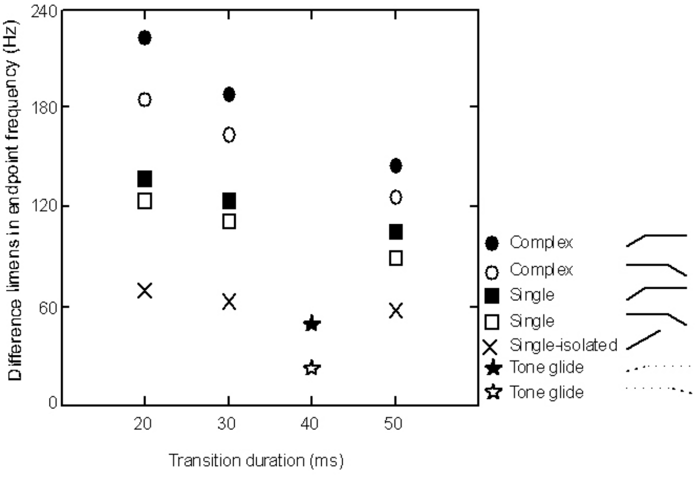


Fig. 8 presents some of our results for same-different judgments of speech-like transition pairs. Several tendencies are clear in this figure. The longer the transition duration becomes, the smaller the DL in endpoint frequency of the transition. The more complex the transition gets, and thus the more natural (ba-da like), the larger the DL and thus the less sensitive! Finally it is clear that sensitivity for final transitions is better than for initial transitions. For more details and other experiments, also with natural stimuli, see [39, 41].

5. SPEECH MODULATION

In sect. 2.3 on consonant reduction, we already hinted on the fact that the intervocalic sound energy difference in VCV segments was lower for spontaneous than for clear speech. This is just one indication of the fact that envelope modulation in (audio frequency bands of) speech is a measure for speech intelligibility. My former TNO-colleagues Houtgast and Steeneken [36, 42] developed the Speech Transmission Index STI to measure and to predict speech intelligibility under conditions of noise, reverberation, communication channels, auditoria, etc. The concept is based on the envelope modulation in octave bands from 125 Hz to 8 kHz. Fourteen modulation frequencies (0.63 to 12.5 Hz) are taken into account. Frequency weighting factors over octave bands are optimized depending on word material, male/female, level-dependent masking, phoneme groups, etc.

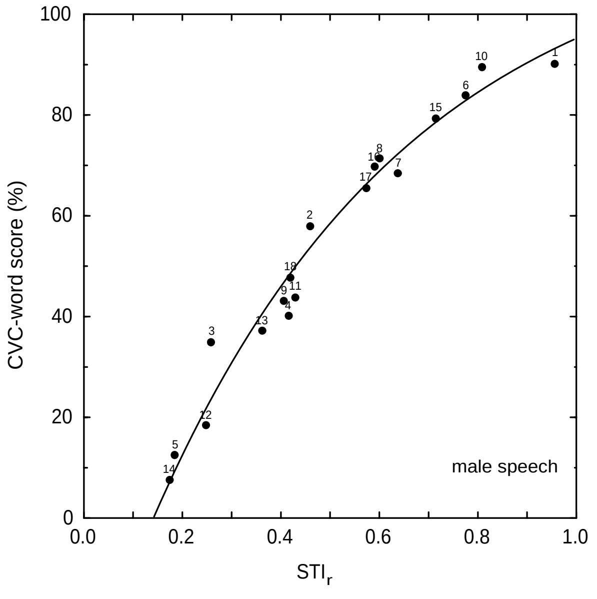


Figure 9: Predicted STI vs. measured CVC score for 18 noise and bandpass-limiting communication-channel conditions.

As an example, Fig. 9 shows the excellent relation (sd of 4.4%) between STI and the intelligibility score with standardized Dutch CVC-lists for 18 different communication channels with conditions of masking noise and bandpass limiting.

But STI also has its limitations, it does not know about linguistic content, it is not good for coded or synthetic speech and it is insensitive for voicing errors or inappropriate pitch. Still, it has proven to be also a very interesting tool for pathological speech evaluation during rehabilitation, as we will see below.

6. DYNAMICS IN DEVIANT SPEECH

In various forms of pathological speech, such as tracheoesophageal (TE) or dysarthric speech, dynamic features of speech are not optimal. In several PhD projects [1, 4, 12] we have studied the voice and articulation characteristics of TE-speech. These TE-speakers speak with air that passes via a voice prosthesis, which is a one-directional valve between trachea and esophagus. This air can put an alternative voice source (neoglottis in the esophagus) into vibration. Although most TE-speakers can communicate rather well despite their alternative way of speaking, a rehabilitation program is frequently advisable. Rehabilitation programs [13], with much attention for voicing, timing, and phoneme, word and sentence pronunciation, are time consuming and costly and are only partly covered by medical insurance. Here, speech technological means could be helpful for feedback and for objectively measuring progress. Various groups [6, 8, 17, 19] have made progress in this direction. Speech recognizers using phonological features are most often applied. The actual recognition score can be used, but it might diagnostically be more interesting to use the feature values themselves. An ongoing project by Clapham and van Son [4, 5] follows this direction for TE-speech.

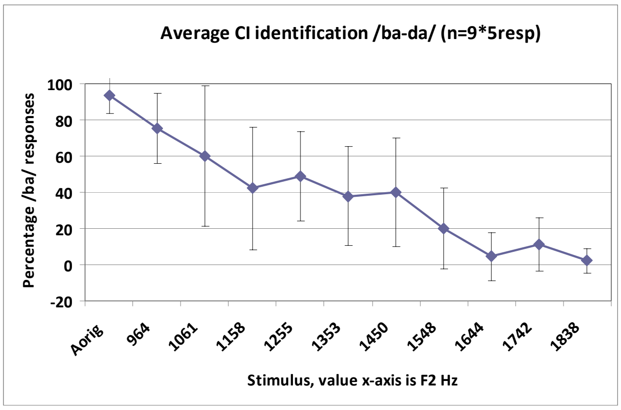
A somewhat different approach is taken in a recent paper by Falk et al. [8], they use a composite measure to predict dysarthric word intelligibility. Factors they take into account are related to atypical vocal source excitation, temporal dynamics, hypernasality, and disordered prosody. With respect to the theme of the present paper, the factor considering the long-term perturbations in the temporal dynamics, is most interesting. The modulation spectrum, as already discussed in sect. 5, is used for this. Various studies have shown that natural speech contains dominant modulation frequencies with a spectral peak at around 4 Hz [7]. Highly intelligible speech generally has a better spread of modulation frequencies above 4 Hz. It is hypothesized that prolonged phonemes, slower speech, as well as an unclear distinction between adjacent phonemes, will cause a shift of the modulation frequency content to modulation frequencies below 4 Hz. This is why Falk et al. proposed to use the *low-to-high modulation energy ratio (LHMR),* being the ratio between modulation energy below and above 4 Hz. The single best measure to predict dysarthric word intelligibility was the 'LP residual kurtosis' with a 0.88 correlation, but this measure describes vocal source excitation atypicality. However, also LHMR was a good predictor for dysarthric word intelligibility. The composite measure, combining six factors, did even better. This opens interesting opportunities for diagnostics and selective feed-back in speech rehabilitation programs.

6.1. Speech dynamics for CI-users

Although cochlear implant users lack much of the fine spectral detail, because of the way speech is processed and presented to the electrodes in the inner ear, many of these CI-users learn to communicate rather well. Shannon et al. [31] and others have indeed shown that with primarily temporal cues in a small number of frequency bands, still good speech recognition for normal listeners is possible.

Still, properly identifying /ba-da/-type synthetic stimuli in isolation appears to be a real challenge for many patients [40]. This is shown in Fig. 10 for a group of 9 post-lingual deaf patients for synthetic stimuli in which F2 onset varied from 964 Hz (ba-like) to 1838 Hz (da-like). The variation among subjects is substantial and none of them had a steep curve as is common for normal-hearing subjects for these stimuli.

Figure 10: Percentage /ba/ responses (plus standard deviation), averaged over 9 post-lingual deaf CI-users, for synthetic stimuli as a function of F2 onset frequency.



7. DYNAMICS IN SPEECH TECHNOLOGY

This is not the most appropriate conference to discuss the role of speech dynamics in speech technology. However, everyone will agree that a proper interpretation of speech dynamics in automatic speech recognition (ASR) and speech synthesis (TTS) is vital in these fields as well.

7.1. Speech dynamics in ASR

In an often cited paper by Mari Ostendorf [22] 'Moving beyond the 'beads-on-a-string model' of speech' she proposes to do better than just using word models as a sequence of phoneme segments, most often in the form of automatically trained triphones. Poor ASR performance is often related to pronunciation variation in the form of substitutions, deletions and insertions, but probably also feature spreading. She proposes to use acoustically-derived subword units plus context dependence to express pronunciation variation. Speaker adaptation can also be beneficial [30, 38].

7.2. Proper dynamics in TTS

In the early days of rule-based formant synthesis it was very hard to model formant transitions properly. The use of diphones seemed to be a way out of this since the transitions were now part of the units themselves. Now, with concatenative unit synthesis, dynamic variation seems to be even less of a problem. However, smoothly joining the units, by optimizing 'join cost' and 'target cost' is not always fully successful.

In a recent paper, Mayo et al. [18] used multi-dimensional scaling (MDS) to decide which acoustic cues were most responsible for judging synthetic speech naturalness. 24 Representative test sentences were generated via the Festival system. The listeners' task was to make a binary decision ('similar' or 'different') about the degree of similarity in naturalness of each pair of (two different) sentences. The 3-dimensional MDS map was then matched against a variety of acoustic characteristics. Listeners appear to be especially sensitive to many different aspects of join quality.

8. CONCLUSIONS

Many topics related to speech dynamics have been discussed. Smooth transitions are an essential aspect of natural speech, and if synthetic speech is not taking good care of that, it is considered unnatural. If there is deviant speech processing, such as for CI-users, a proper /ba-da/ distinction is difficult.

Global dynamic aspects of speech are very nicely represented in the 'modulation transfer function' which is the basis for the speech transmission index STI. The modulation spectrum, or features derived from that, such as the 'low-to-high modulation energy ratio (LHMR)', are good predictors of the word intelligibility of dysarthric speech. Such measures might be very useful for diagnostics and selective feedback in speech rehabilitation programs also for TE speech.

Given the communicative importance of natural formant transitions, it is surprising that the human ear is not very sensitive to isolated representations of those, in terms of difference limen or just noticeable difference. Given also the substantial variability in realization, as a function of stress, style, speaking rate, etc., it is understandable that ASR is still no easy job. The more respect we must have for the high performance of the human detection, perception, recognition and understanding system of natural speech in everyday communication [26, 29].

9. REFERENCES

[1] As, C.J. van (2001), *Tracheoesophageal speech. A multidimensional assessment of voice quality*, PhD thesis Univ. of Amsterdam, 211 p.

[2] Bergem, D. R. van (1995), *Acoustic and lexical vowel reduction*, PhD thesis Univ. of Amsterdam, 194 p.

[3] Chládková, K and Hamann, S (2011), High vowels in Southern British English: /u/-fronting does not result in merger, *Proc.17th ICPhS*, Hong Kong.

[4] Clapham, R.P., Hilgers, F.J.M. & van Son, R.J.J.H. (2011), Automatic phonological feature evaluation: Is the effect of speech therapy seen in recognition scores for voicing and manner?, *Proc. 12th Interspeech*, Florence, Italy.

[5] Clapham, R.P., van Son, R.J.J.H. and Hilgers, F.J.M. (2011), Automatic and human evaluation of SUS stimuli for speech intelligibility evaluation before and after speech therapy, *Proc. 12th Interspeech*, Florence, Italy.

[6] De Bodt, M.S., Hernández-Diaz Huici, M.E. and Van De Heyning, P.H. (2002), Intelligibility as a linear combination of dimensions in dysarthric speech, *Journal of Communication Disorder*s 35(3), 283-292.

[7] Drullman, R., Festen, J.M. and Plomp, R. (1994), Effect of reducing slow temporal modulations on speech perception, *J. Acoust. Soc. Am.* 95(5), 2670-2680.

[8] Falk, T.H., Chan, W.-Y. and Shein, F. (2011 in print), Characterization of atypical voice source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility, *Speech Communication*.

[9] Harst, S. van der (2011), *The Vowel Paradox. A sociophonetic study on Dutch*, PhD thesis Radboud University Nijmegen, 380 p.

[10] Jacobi, I. (2009), *On variation and change in diphthongs and long vowels of spoken Dutch*, PhD thesis Univ. of Amsterdam, 163 p.

[11] Jacobi, I., Pols, L.C.W. and Stroop, J. (2007), Dutch diphthong and long vowel realizations as changing socio-economic markers, *Proc. 16th ICPhS*, Saarbrücken, Germany, 1481-1484.

[12] Jongmans, P. (2008), *The intelligibility of tracheoesophageal speech: An analytic and rehabilitation study*, PhD thesis Univ. of Amsterdam, 272 p.

[13] Jongmans, P., van Rossum, M., van As-Brooks, C.J., Hilgers, F. and Pols, L. (2008), An evidence-based rehabilitation program for tracheoesophageal speakers, *Proc. Invitational Round Table 'Evidence-based Voice and Speech Rehabilitation in Head and Neck Cancer'*, Amsterdam, 41-60.

[14] Koopmans-van Beinum, F.J. (1980), *Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions*, PhD thesis Univ. of Amsterdam, 163 p.

[15] Lindblom, B. (1963), Spectrographic study of vowel reduction, *J. Acoust. Soc. Am.* 35(11), 1773-1781.

[16] Lindblom, B. and Studdert-Kennedy, M. (1967), On the rôle of formant transitions in vowel recognition, *J. Acoust. Soc. Am* . 42(4), 830-843.

[17] Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M. and Nöth, E. (2009), PEAKS – A system for the automatic evaluation of voice and speech disorders, *Speech Communication* 51(5), 425-437.

[18] Mayo, C., Clark, R.A.J. and King, S. (2011), Listeners' weighting of acoustic cues to synthetic naturalness: A multidimensional scaling analysis, *Speech Communication* 53(3), 311-326

[19] Middag, C., Martens, J.-P., Van Nuffelen, G. and De Bodt, M. (2009), Automated intelligibility assessment of pathological speech using phonological features, *EURASIP Journal on Advances in Signal Processing*, 1-9.

[20] Nearey, T.M. (1997), Speech perception as pattern recognition, *J. Acoust. Soc. Am.* 101(6), 3241-3254.

[21] Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.P., Moortgat, M. and Baayen, H. (2002), Experiences from the Spoken Dutch Corpus Project. *Proc. 3rd LREC*, Las Palmas, 340-347.

[22] Ostendorf, M. (1999), Moving beyond the 'beads-on-a-string' model of speech, *Proc. ASRU Workshop*, Keystone, CO, 79-84.

[23] Pols, L.C.W. (1971), Real-time recognition of spoken words, *IEEE Trans. Comp.*, C-20, 972-977.

[24] Pols, L.C.W. (1977) *Spectral analysis and identification of Dutch vowels in monosyllabic words*, PhD thesis Free Univ. of Amsterdam, 152 p.

[25] Pols, L.C.W. (2005), Samenspraak. Acquiring and implementing phonetic knowledge, *IFA Proceedings* 26, 7-23

[26] Pols, L.C.W. (2009), Phonetics: Consistency and variability, In: G. Fant, H. Fujisaki and J. Shen (Eds.), *Frontiers in Phonetics and Speech Science*, The Commercial Press, Beijing, 349-357.

[27] Pols, L.C.W., Lyakso, E., van der Stelt, J.M., Wempe, T.G. and Zajdo, K. (2006), Vowel data of early speech development in several languages, *Proc. Multiling*, Stellenbosch, South Africa.

[28] Pols, L.C.W. & van Son, R.J.J.H. (1993), Acoustics and perception of dynamic vowel segments, *Speech Communication* 13(1-2), 135-147.

[29] Pols, L.C.W. & van Son, R.J.J.H. (2006), Speech dynamics: Acoustic manifestations and perceptual consequences, In: P. Divenyi, S. Greenberg and G. Meyer (Eds.), *Dynamics of Speech Production and Perception*, NATO Science Series 1, Life and Behavioural Sciences – Vol. 374, IOS Press, 71-80.

[30] Pols, L.C.W. and Weenink, D.J.M. (2005), Vowel recognition and (adaptive) speaker normalization, *Proc. 10th SPECOM*, Patras, Greece, Vol. 1, 17-24.

[31] Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. and Ekelid, M. (1995), Speech recognition with primarily temporal cues, *Science* 270, 303-304.

[32] Son, R.J.J.H. van (1993), *Spectro-temporal features of vowel segments*, PhD thesis Univ. of Amsterdam, 195 p.

[33] Son, R.J.J.H. van & Pols, L.C.W. (1990), Formant frequencies of Dutch vowels in a text, read at normal and fast rate, *J. Acoust. Soc. Am.* 88(4), 1683-1693.

[34] Son, R.J.J.H. van & Pols, L.C.W. (1999), Perisegmental speech improves consonant and vowel identification, *Speech Communication* 29(1), 1-22.

[35] Son, R.J.J.H. van & Pols, L.C.W. (1999), An acoustic description of consonant reduction, *Speech Communication* 28(2), 125-140.

[36] Steeneken, H.J.M. (1992), *On measuring and predicting speech intelligibility*, PhD thesis Univ. of Amsterdam, 162 p.

[37] Sussman, H.M., Bessel, N., Dalston, E. and Majors, T. (1999), An investigation of stop place of articulation as a function of syllable position: a locus equation perspective, *J. Acoust. Soc. Amer.* 101, 2826-2838.

[38] Weenink, D.J.M. (2006), *Speaker-adaptive vowel identification*, PhD thesis Univ. of Amsterdam, 236 p.

[39] Wieringen, A. van (1995), *Perceiving dynamic speechlike sounds. Psycho-acoustics and speech perception*, PhD thesis Univ. of Amsterdam, 256 p.

[40] Wieringen, A. van (2011), personal communication

[41] Wieringen, A. van & Pols, L.C.W. (2006), Perception of highly dynamic properties in speech, In: S. Greenberg and W.A. Ainsworth (Eds.), *Listening to speech. An auditory perspective*, Lawrence Erlbaum Associates, Publishers, Mahwah, NJ, 21-38.

[42] Wijngaarden, S.J. (Ed.) (2002), *Past, present and future of the Speech Transmission Index*, TNO Human Factors, Soesterberg, 140 p.