LITERATURE OVERVIEW ON PERCEPTUAL AND PHYSICAL NORMALIZATION OF SPEAKER VARIATION

David J.M. Weenink

1. INTRODUCTION

Although the physical differences between linguistically the same utterances produced by speakers with different vocal tract lengths are great, listeners don't seem to have great difficulty in perceiving these utterances as the same.
Apparently the listener is able to extract the relevant linguistic features from the speech signal via some normalization procedure, but little is known about the perceptual strategies used to normalize for these different talkers.
In this paper a survey will be given of perceptual aspects of normalization, proposed physical normalization procedures for vowels and normalization in automatic speech recognition systems. This survey is a first introduction for this author in this research area and is part of a ZWO project on speaker normalization.

2. PERCEPTUAL ASPECTS OF NORMALIZATION

The intrinsic variability of vowels is supported by perceptual and physical data from e.g. Peterson and Barney (1952) for American English and Klein et al. (1973) and Koopmans - van Beinum (1980) for Dutch. It has led many investigators to propose that the listener calibrates (normalizes) each talkers' vowel space on the basis of some reference derived from preceding utterances (Joos, 1948). One aspect of this problem is the question how much information and what kind of information a listener needs in order to adapt to a (new) speaker. The most cited study in this respect is probably the

Ladefoged and Broadbent (1957) one. In this study six versions of the precursor sentence "Please say what this word is" were synthesized with different formant structures. Four test words of the form bVt were also synthesized. Subjects were offered one of the six precursor sentences followed by a test word and were asked to identify the test word. The subjects were undoubtedly influenced in their identification of the test word by the auditory context in which it occurred. The authors' conclusion is that "...the linguistic information conveyed by a given vowel is largely dependent on the relations between the frequencies of the formants of other vowels occurring in the same auditory context" (p. 102).

Dechovitz (1977) confirmed the above statement by using natural speech. Each of a set of bVt test words, spoken rapidly within a sentence carrier, "Please say ... for me", by an adult male, was presented for recognition within (1) a carrier sentence from the male, (2) an identical carrier sentence from a nine year old boy with a substantially different vocal tract length and (3) excised from sentence context. The two speakers achieved the same pitch levels and speaking rate, with the result that the mixed talker sentences were perceived as if uttered by one speaker. Errors in recognition were most substantial for test words embedded within the child's carrier. The results are in line with the Ladefoged and Broadbent study, although in my opinion the error rates are rather high for case (2) and (3), probably due to bad listening conditions. Dechovitz too concludes "... the identity of a vowel may be computed over some stretch of speech longer than the syllable in which it lies". (p. 213).

The lenght of the excerpt needed to gain knowledge about the identity of the vowel has been the course of much discussion and controversies. Strange et al. (1976) have directed considerable attention toward the role of consonantal and speaker context on the identification of naturally spoken vowels.

In the following we will use the term mixed speaker condition for a speaker-randomized condition and a blocked speaker condition for a condition in which the speaker is fixed throughout a full utterance set.

Strange et al. predicted more confusion errors in the mixed speaker condition than in the blocked speaker condition for isolated vowels. If consonantal environment aids in the identification of vowels a further prediction would be that vowels in CVC words are better recognized than in isolated position. Naive listeners identified nine American English vowels in isolation and in pVp words. Their results confirm the above stated: for isolated vowels 43% errors in the mixed speaker condition, 31% errors for the blocked condition; for a pVp word 17% errors in the mixed and 9.5% in the blocked speaker condition. The effect of consonantal context is considerably greater than that of speaker context. Consonantal context is critical for vowel identification, while speaker context is of marginal importance: "... these results offer strong evidence that dynamic acoustic information distributed over the temporal course of the syllable is utilized regularly by the listener to identify vowels" (p. 213).

In a following experiment Verbrugge et al. (1976) investigated if precursor vowels would aid in the identification. They conclude: "Experiments with known subsets of a talker's vowels did not significantly reduce errors on subsequent test tokens: following the point vowels /i/, /ɑ/, /u/, errors averaged 12% on vowels in hVd words and 15% in pVp words; following three central vowels /ɪ/, /æ/, /ʌ/ errors averaged 15% in pVp words. Precursors mainly influenced listeners response biases, rather than facilitating true improvement in vowel identifiability" (p. 198).

There has been some criticism on the work of Strange et al. and on that of Verbrugge et al.

Van Balen (1977) stated that "... the stimulus material they used was deficient, [that] their interpretation of results is questionable (especially with regard to problems of normalization) and [that] they unjustly brush aside generalization problems". (p. 43).

In his own experiment on natural read-out speech (1976) he had presented to four groups of listeners thirty speech excerpts under four different conditions. Group 1 listened for one and a half minute to the voice of the excerpt speaker, group 2 did not hear anything in advance, group 3 listened to another voice and group 4

to music. Group 1 scored best and the listeners of groups 2, 3 and 4 needed at least seven words of a length of one to three syllables to get the same scoring as the subjects of group 1. His conclusion (p. 43) that: "... some stretch of speech is required to get accustomed to the voice of the speaker" is in line with the view of Ladefoged and Broadbent.

Macchi (1980) tested naive listeners' identification of eleven American English vowels spoken in isolation and in consonantal context with an experimental design comparable to that used by Strange et al. She used high quality listening conditions, the speakers and the listeners were closely matched for regional dialect and for response alternative she had the listeners identify the isolated vowels and tVt syllables by rhyming them with English words. The results in her study: 2% misidentification in the blocked speaker condition and 8% misidentification in the mixed speaker condition. No evidence of difference in identifiability between the isolated vowels and vowels in consonantal context was found. The reasons for this difference in findings are according to Macchi due to experimantal methods, the language under study, the vowel set, the set of response alternatives, the degree of dialect mismatch between speakers and listeners, the phonetic training of the listeners, the nature of the stimuli and the quality of the listening conditions.

The study by Assmann et al. (1982) is another one which contrary to Strange et al. reports very high recognition scores for isolated vowels either in mixed or in blocked speaker condition. Even gated vowels, i.e. the segmented stationary parts of vowels, in the mixed speaker condition were well identified, with only 14% errors. The outcome of their study:

(1) The identification performance is clearly related to acoustic characteristics.
(2) Differences between full and gated vowels are related to dynamic information.

I tend to believe the most recent literature (Assmann et al., 1982; Macchi, 1980), which shows that isolated vowels can be recognized very well. I would agree with Macchi that the differences in findings e.g. the high error rates in the Strange et al. study are probably

due to the effects she names.

The number of studies on normalization of vowels is tenfold that of normalization of consonants. Some of the latter are cited here.

Schwartz' (1968) study indicates that listeners can identify the sex of the speaker from the isolated production of /s/ and /ʃ/, but not from /f/ and /θ/. Spectral analysis of /s/ and /ʃ/ showed that the female spectra tended to be higher in frequency than the male spectra.

Rand (1971) observed that listeners' identification of synthetic voiced stops shifted with the vowel space of the context in which they were embedded. When heard in the context of a vowel from a smaller vocal tract the formant transition boundary values between /b/ and /d/, /d/ and /g/ shifted upward.

May (1976) studied the two synthetic fricatives /s/ and /s̆/, which stand for alveolar and alveo-palatal /s/ respectively, in the context of two different /æ/'s: one from a large vocal tract and the other from a small one. Because the noise centerfrequency is the main perceptual cue for distinguishing the /s/ and /s̆/ this noise centerfrequency is varied. Results showed an upward boundary shift for the small vocal tract stimuli as compared with the large vocal tract stimuli. This suggests the /s/-/s̆/ cue is dependent on vocal tract size.

3. PROPOSED NORMALIZATION PROCEDURES FOR VOWELS

Nordstrom and Lindblom (1975) have suggested a uniform scaling which amounts to multiplying the formants to be corrected by a factor $k = L_a/L_r$ where L_a is the vocal tract length associated with the subjects' average F3 of open vowels (vowels with F1 greater than 600 Hz) and L_r is the vocal tract length of the reference 'male'. In the F1-F2 - plane this results in a shift of the formants of women and children into the direction of the origin. These measures are derived from an emperical mean curve of F3 versus simulated vocal tract length in model experiments.

Fant (1975) proposes to include the end-correction of the vocal tract

which amounts to 1.0 cm in the lenghts, according to

$$F3_{ref}/F3_{av} = (L_a-1) / (L_r-1) = (1+k/100)$$

with k the scale factor in percent. The non-uniform extension of the Nordstrom and Lindblom procedure is to make the correction factor a function of both formant number and vowel category.
Van Dijk (1980, 1984) normalizes the formant values of each speaker according to

$$I_{ij} = (F_{ij}-F_{rj}) / F_{rj}$$

with F_{ij} formant j for vowel i.
The reference formant F_{rj} is the j-th formant of a straight tube of lenght L, $F_{rj} = (2j-1)c/4L$. The length is estimated from minimizing the total distance of all formants of all vowels of the vowel system with respect to length in the I-plane.
In Wakita's paper (1977) an attempt is made to utilize LPC techniques to bring the articulatory parameters of vocal tract shape and length into the acoustic domain in order to eliminate inter-speaker differences in acoustic parameters.
The vocal tract length and area function are first estimated from the acoustic speech waveform. Then the area function is normalized to an acoustic tube of the same shape and reference length. The normalized formant frequencies are defined as the resonance frequencies of this tube. Because of the fact that an infinite number of shapes of different length are realizable for a given set of formant frequencies and bandwidths some criterion has to be chosen to calculate the actual length. That length is chosen for which the tube comes closest to a tube of uniform shape.
The distribution of normalized and nonnormalized formant frequencies for nine stationary American vowels were investigated with 14 male and 12 female speakers. Fairly compact distributions of the vowels in the F1, F2, F3 space were obtained.
Disner (1980) makes a comparison of the vowel normalization procedures of Gerstman, Lobanov, Neary and Harshman.

Gerstman's (1968) procedure fixes the maximum and minimum F1 and F2 values in each speaker's vowel system at fixed arbitrary levels and all other F1 and F2 values are then scaled within these ranges. Lobanov's (1971) procedure standardizes the mean and standard deviation for each speaker's vowels with the equation

$$F'_i = (F_i - \bar{F}_i)/\sigma_i,$$

where F_i is a given formant, $\bar{F}_i$ the average value of F_i across all vowels and σ_i the standard deviation of F_i about its mean for all vowels.

The normalization procedure of Neary (1977) can be expressed as follows:

$$G'_{ijk} = G_{ijk} - \bar{G}_k \quad \text{in which}$$

$$G_{ijk} = \ln (F_{ijk}) \quad \text{and}$$

$$\bar{G}_k = \left(\sum_{i=1}^{N} \sum_{j=1}^{M} G_{ijk} \right)/NM$$

with i=1, ... N formants and j=1, ... M vowels.

So G represents the logtransformed frequency of formant i and $\bar{G}_k$ is the average of the log-transformed frequency of formants 1 to N over all vowels of speaker k. This procedure was already used by Pols et al. (1973).

In the Parafac procedure of Harshman (1970), a three mode factor analysis model, the observed formant values are represented as follows:

$$d_{ijk} = \sum_{l=1}^{L} f_{il} v_{jl} s_{kl} + e$$

where d_{ijk} is the observed value for formant i for vowel j as spoken by subject k; f is the loading of formant i on factor 1; v is the loading of vowel j on factor 1; s is the loading of speaker k on factor 1; e is an error term and L is the number of factors.

For the normalization the mean of all vowels for subject k is subtracted out so that d_{ijk} represents a deviation of a speaker from his mean.

The above four procedures are then tested on six different Germanic languages. The comparison between the procedures is on two points. First how they reduce the scatter in each single language and second how they preserve linguistic 'facts' such as e.g. an observation that Dutch /ɛ/ is more open than the English vowel in 'bed'; intermediate between 'set' and 'sat'; or "All German vowels are tenser than their English counterparts". On the basis of this comparison she concludes that normalization techniques which make use of the mean and/or standard deviation of the vowel system can be very effective in reducing the variance between speakers within a single language or dialect, they are however inappropriate for comparing the normalized vowels of one language with the independently normalized vowels of another language. A Parafac-based normalization procedure is the best of the four procedures for purposes of cross-language or cross-dialect comparison.

Labov (1979) too notes that a normalization procedure which shows the greatest clustering is not necessarily the best. Optimal normalization should eliminate only those acoustic differences which are due to differences in vocal tract lengths.

Bladon et al. (1988) use a listener oriented approach to the normalization problem. They start, as they say, from the old Potter and Steinberg (1950) idea that "a certain pattern of stimulation along the basilar membrane may be identified as a given sound, regardless of position along the membrane" and their model draws on current knowledge about human auditory analysis. Speaker vowel-spectra are transformed to a scale with ordinate in sones and abcissa in barks. Female/male versions of vowels on this scale are very similar except for a displacement of about 1 bark.

Another F0-dependent normalization procedure is described in Bladon (1982). According to Holmes (1983) it can be simplified if one could model an assumed higher level of auditory processing that estimates the true formant frequencies, especially F1, even when the harmonics are widely spaced as is the case for highpitched voices.

4. NORMALIZATION IN AUTOMATIC SPEECH RECOGNITION SYSTEMS

In his paper on speaker normalization Jaschul (1979) distinguishes two general approaches for making an automatic speech recognition system suitable for use with several different speakers. The first possibility is that the reference set is built up from utterances of those speakers who will test the system, or in order to make it speaker-independent from utterances of a wide number of speakers. The reference set is then fixed or may be continuously updated. For a discussion of such a system see Klatt (1979). The second possibility, the one Jaschul has tested, first on eight German vowels (1979) and later on including thirteen consonants (1982), is to transform the test speakers' utterances in such a way that they become more similar to the reference set patterns. The reference set in this case may be the set of utterances of one speaker or an artificial set. The transformation can be put in the form of a general matrix equation. From a mean square error optimalization the coefficients of the matrix can be calculated. The transformation is phoneme dependent, because speaker-specificness already enters at the phoneme level. High improvements on the recognition score are reported using this transformation.
In his system for isolated word recognition Furui (1980) reduces the amount of training necessary for building up a reference set for new speakers who will test the system. The complete reference set for the new speaker is obtained by applying a transformation to the test speakers' training utterances. The transformation rules are, independently from the test speakers, obtained via a multiple regression analysis on training speakers' data and are speaker-independent.

5. DISCUSSION

Up to this moment little is known about the way in which human beings adapt to different talkers. As a first stage in getting more knowledge about this proces of adaptation an investigation will be

carried out to see how well and how fast a listener adapts to different talkers under experimentally varying circumstances. This will be done by a systematic evaluation of the identifiability of vowels and consonants in various experimental circumstances. The vowels and consonants are spoken in isolation or taken from short words. The amount of adaptation to the voice of the speaker will be manipulated by varying the amount of context and all this can be done in a blocked or in a mixed speaker condition. This could learn us whether there is any difference in normalization between vowels and consonants.

For the analysis of the stimuli we will use some of the methods described above: (1) the formant normalization procedures based on vocal tract length estimation of Wakita. (2) Data transformations described by Macchi and Van Dijk. (3) Spectral transformations of Jaschul and the quasi-auditory spectra of Bladon and Lindblom. Some of the transformations for vowels will be used on (some) of the consonants too. Whether or not the developed transformations make sense will be tested in subsequent listening experiments.

REFERENCES

Assmann, P.F., Neary, T.M. & Hogan, J.T. (1982). Vowel identification: Orthographic, perceptual and acoustic aspects, J. Acoust. Soc. Am., 71, 975-989.

Balen, C.W. van (1976). The influence of normalization on the intelligibility of excerpts, PRIPU 1, 16-23.

Balen, C.W. van (1977). Different views on problems of normalization, PRIPU, 2, 32-46.

Bladon, R.A.W. (1982). Problems of normalizing the spectral effects of variations in the fundamental, Proc. Inst. Acoust. Autumn Conference A5.1-A5.5.

Bladon, R.A.W., Henton, C.G. & Pickering, J.B. (1983). Testing an auditory theory of speaker normalization. Abstracts of the Tenth International Congress of Phonetic Sciences. Foris Publications, Dordrecht, 421.

Dechovitz, D. (1977). Information conveyed by vowels: a confirmation, Haskins SR-51/52, 213-219.

Dijk, J.S.C. van (1980). How to normalize your own vowel system, IFA Proceedings 6, 67-72.

Dijk, J.S.C. van (1984). Conservation of vowel contrast in various speech conditions. This Proceedings of the Institute of Phonetic Sciences Amsterdam, 19-31.

Disner, S. (1980). Evaluation of vowel normalization procedures, J. Acoust. Soc. Am., 67, 253-261.

Fant, G. (1975). Non-uniform vowel normalization, STL-QPRS 2-3, 1-19.

Furui, S. (1980). A training procedure for isolated word recognition systems, IEEE Trans., ASSP-28, 129-136.

Holmes, J. (1984). On normalization, will be published in Proceedings of Invariance Symposium, Cambridge, October 1983.

Jaschul, J. (1979). An approach to speaker normalization for automatic speech recognition, Proc. IEEE-ICASSP79, Washington D.C., 235-238.

Jaschul, J. (1982). Speaker adaptation by a linear transformation with optimised parameters, Proc. IEEE-ICASSP82, Paris, 1657-1660.

Klatt, D.H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical acces, J. of Phon., 7, 279-312.

Klein, W., Plomp, R. & Pols, L.C.W. (1970). Vowel spectra, vowel spaces and vowel identification, J. Acoust. Soc. Am., 48, 999-1009.

Koopmans-van Beinum, F.J. (1980). Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions, Ph. D. Thesis University of Amsterdam, Academische Pers B.V. Amsterdam.

Labov, W. (1979). Sociolinguistic approach to the problem of normalization, Paper at 9th Int. Congress of Phonetic Sciences, Copenhagen.

Ladefoged, P. & Broadbent, D.E. (1957). Information conveyed by vowels, J. Acoust. Soc. Am., 29, 98-104.

Macchi, M.J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonant context, J. Acoust. Soc. Am., 68, 1636-1642.

May, J. (1976). Vocal tract normalization for /s/ and /š/. Haskins SR-48, 67-73.

Neary, T.M. (1977). Phonetic feature systems for vowels, Ph. D. Thesis, University of Connecticut. Reproduced by Indiana University Linguistics Club, 1978.

Nierop, D.J.P.J. van, Pols, L.C.W. & Plomp, R. (1973). Frequency analysis of Dutch vowels from 25 female speakers, Acustica 29, 110-118.

Nordstrom, P.E. & Lindblom, B.J. (1975). A normalization procedure for vowel formant data, paper 212 at 8th Int. Congress of Phonetic Sciences, Leeds.

Peterson, G.E. & Barney, H.L. (1952). Control methods used in a study of the vowels, J. Acoust. Soc. Am., 24, 175-184.

Pols, L.C.W., Tromp, H.R.C. & Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers, J. Acoust. Soc. Am., 53, 1093-1101.

Potter, R.K. & Steinberg, J.C. (1950). Towards the specification of speech, J. Acoust. Soc. Am., 22, 807-820.

Rand, T.C. (1971). Vocal tract normalization in the perception of stop consonants, Haskins, SR-25/26, 141-146.

Schwartz, M.F. (1968). Identification of speaker sex from isolated voiceless fricatives, J. Acoust. Soc. Am., 43, 1178-1179.

Strange, W., Verbrugge, R., Shankweiler, D. & Edman, T. (1976). Consonant context specifies vowel identity, J. Acoust. Soc. Am., 60, 213-224.

Verbrugge, R.R., Strange, W., Shankweiler, D. & Edman, T.R., (1976). What information enables a listener to map a talker's vowel space? J. Acoust. Soc. Am., 60, 198-212.

Wakita, H. (1977). Normalization of vowels by vocal tract length and its application to vowel identification, IEEE Trans. ASSP-25, 183-192.