Forum Acusticum 2011 - Aalborg - Denmark

Objective Evaluation of Consonant-Vowel pairs produced by Native Speakers of Danish

Thomas U. Christiansen
Department of Electrical Engineering, Centre for Applied Hearing Research, Ørsteds Plads, Building 352, Technical University of Denmark, DK-2800 Lyngby, Danmark.

Peter Juel Henrichsen
Copenhagen Business School, Department of International Language Studies and Computational Linguistics, Dalgas Have 15, DK - 2000 Frederiksberg, Denmark

Summary
Nonsense syllable speech materials are often used when investigating speech perception in quiet and under adverse conditions. The main advantage of using nonsense syllables over words and sentences is that the acoustic as well as the linguistic context is minimal. This paper presents three anechoic recordings of 13 male and 13 female native talkers of Danish each speaking 65 nonsense syllables repeated three times with the neutral intonation contour for Danish (in total 15210 syllables). The authors compared and ranked groups of three recordings. These three recording had the same talker and had identical phonetic content. The syllables were ranked according to the general "appropriateness" and consistency, i.e., prototypical production of the consonant-vowel (CV) with respect to applicability in speech perceptual studies. The results were compared to results of an automatic method based on acoustic measures. The two novel ideas are 1) to devise an automated method for evaluating "appropriateness" of CVs and 2) to develop a Danish CV-material annotated with an objective measure of "appropriateness" for each recorded CV. The latter would potentially render more CV's appropriate for perceptual studies. Moreover, objective evaluation would make it possible to examine any perceptual effects of variability in CV production (for example how susceptible different renderings by the same talker of CV's are to background noise). To the knowledge of the authors, no such material has yet been published for any language.

PACS no. 43.71.Es, 43.7.Arj

1. Introduction

Perception of spoken language is a complex process involving several processing stages of quite disparate nature. Such processing stages relate to hearing, lexical structure (sometimes called mental lexicon), phonetic, phonemic, morphologic, syntactic and semantic organisation of language. Rather than investigating the process as a whole, many studies have focused on quantitatively characterising the capacity of hearing to identify phonetic segments like consonants and vowels (e.g. [1, 2, 3]).

Such studies require speech materials which eliminate or reduce effects of confounding linguistic factors, such as processing of lexical structure and morphology. Nonsense syllables, i.e. , syllables that are not words, are widely used in these studies. Moreover, syllables have minimal acoustic context because they are the shortest naturally occurring speech sounds. This in turn, minimises any confounding co-articulatory (acoustic) effects.

While several CV speech materials are avialable in English, fewer are avialable for smaller languages like Danish. There are two main reasons for developing similar speech materials across languages, one practical and one theoretical.

1. Speech perceptual studies require that the native language of talker and listener be matched in order to avoid biasing-effects originating from non-native talkers and listeners. This is true even if the speech material does not include lexicalised words (see below). Since talkers of a given language is most easily accessible in the home country, it is practical to have similar speech materials for different languages.
2. While some aspects of speech perception are similar between languages others differ. This is true even for "simple" speech stimuli as used in the present paper, e.g. the voice onset time (VOT) in

the French /p/ is shorter than the English equivalent [4].

Most speech perceptual studies use speech materials with an implicit assumption that all the individual speech tokens (here syllables) are above a certain "threshold of acceptability". This threshold is not given, defined or even mentioned. Moreover, the speech tokens deemed appropriate for a given perceptual study are considered to be equally appropriate.

The aim of the present paper is to devise a method by which the laborious and error prone process of selecting nonsense syllables suitable for use in perceptual studies, is performed automatically, i.e., based on an objective evaluation method. In addition, each speech token will be explicitly labelled according to the degree of objective appropriateness. This rating/ranking does not preclude the eventual division of the material into appropriate and inappropriate, however, the division decision can be adapted to the requirement of the specific perceptual study at hand, thus maximising the number of speech tokens available for any given study. Moreover, it is possible to validate potentially erroneous conclusion, from speech perceptual studies, if they rely on tokens that differ substantially in rating/ranking. To the knowledge of the authors, no such material has yet been published for any language.

2. Sound data materials

2.1. Speech material

The Danish consonants recorded in the present study correpond to the phonemes /ptkbdgfsvmnrlhʃjw/[1] roughly corresponding to the following phones in IPA-notation [5] [pʰtˢkʰɓɖɠfsʊmnʁlhʃjw]. Note that the two approximants /j/ and /w/ were included in the recording as if they were consonants[2] .

Consonants were followed by one of three long vowels /iau/ corresponding to vowel qualities designated by IPA-symbols [iæu]. This first consonant-vowel (CV) syllable was stressed. Some combinations of consonants and vowels coincide with Danish words. In order to dissociate meaning from all syllables a second unstressed /tu/-syllable was added. So the recorded nonsense syllables consisted of four speech sounds a consonant and a vowel followed by /tu/. We refer to these syllables as CVtu.

To keep talkers alert six fillers with unstressed second syllable /ta/ ([tæ] in IPA notation) was incorporated into the material (/ʃata/ /lata/ /wita/ /mita/ /ruta/ /juta/). Eight additional /v/-syllables was included, since we speculate that /v/ is articulated with a higher degree of variability than the other consonants and plan to investigate this speculation elsewhere. For the present purpose, however, this makes it more difficult to use /v/ and therefore we disregards these recordings in the present paper.

The total of the seventeen (consonants) times three (vowels) plus six fillers and eight additional /v/'s were transcribed and randomly distributed across six lists as shown in Table I.

Table I. Lists of nonsense syllables in phoneme notation

List 1	List 2	List 3	List 4	List 5	List 6
pa:tu	pi:tu	pu:tu	ka:tu	ki:tu	ku:tu
ru:tu	nu:tu	mi:tu	ma:tu	na:tu	ni:tu
vi:tu	va:tu	li:tu	vu:tu	mu:tu	la:tu
ʃa:ta	ju:ta	ru:ta	wi:ta	la:ta	mi:ta
ti:tu	ta:tu	da:tu	bu:tu	tu:tu	ba:tu
vu:tu	fi:tu	fu:tu	fa:tu	si:tu	sa:tu
ha:tu	ra:tu	ri:tu	lu:tu	hu:tu	hi:tu
vi:tu	vi:tu	vu:tu	vu:tu	vi:tu	vu:tu
wa:tu	ʃu:tu	wi:tu	ʃi:tu	ja:tu	ga:tu
su:tu	ji:tu	ʃa:tu	vi:tu	wu:tu	ju:tu
bi:tu	du:tu	gi:tu	gu:tu	di:tu	

2.2. Recording procedures

The recordings were carried out in two stages. The aim of the first stage was to produce a CD, which could be used in the second stage. This CD contains sound recordings of nonsense syllables as shown in Table I.

The second stage consisted in recording talkers repeat the content of the CD from the first stage. The recordings from the second stage is the topic of the present paper while the recordings from the first stage is merely used as prompting material.

2.2.1. First stage

In the first stage the authors were recorded speaking each item from Table 1 three times in succession with the neutral sentence intonation contour for Danish (falling). At the beginning of each recording the authors uttered the Dansih phrase "Nu bliver der sagt" (English: "Now this will be said"). The best of the two recordings was used to produce the CD.

The nonsense syllables were put on the CD with six tracks, each of which corresponds to a column in Table

[1] We adopt the common practice of denoting phonemes between // and phones in [ ]

[2] Although the Danish /v/ is closer to an approximant than the English counterpart it is considered to be a consonant in Danish phonology [6]

I such that each track starts with the prompting sentence "Nu bliver der sagt" immediately followed by the first nonsense syllable repeated three times. We refer to these three utterances of the nonsense syllables as a triplet. Subsequent triplets were preceded by 4 seconds of silence. This allows for the talkers to repeat the triplet from the CD.

2.2.2. Second stage

Recordings were carried out in the small anechoic chamber at the Technical University of Denmark [7] using a low noise 1-inch B&K 4179 microphone with a B&K 2660 preamplifier attached to a SoundDevices 722 harddisk recorder. The microphone power supply was a B&K 2807. The microphone was mounted on a stand no less than 1 meter from the mouth of the talker. The talker was seated in a desk chair facing the microphone. The system was calibrated with a B&K 4239 calibrator so that 94 dB SPL 1 kHz calibration tone corresponded to the maximum level of the harddisk recorder.

The prompting material was played back by a Revox B226 Compact CD player over a DynAudioAcoustics BM6 loudspeaker attached to an AT-JR-32dB/10W amplifier at a clearly audible level.

The talkers were instructed to repeat what they heard including carrier sentences, F0 and nonsense syllables. They were instructed to do so at a natural level of vocal effort. The first list was presented in its entirety and followed by a short break. Subsequent lists were recorded either singly or in sequences of two or three. Test subjects were frequently offered water and breaks between lists.

3. Perceptual evaluation

The aim of the perceptual evaluation of the speech recordings was to rank them accroding to appropriateness. The evaluation was based on comparisons within talker and within unique phoneme content. e.g. listeners were asked to rank three recordings each consisting of /pitu pitu pitu/ uttered by a single talker.

Recordings were annotated with either 1, 2 or 3 where 1 is most appropriate and 3 is least appropriate. In cases where the perceived difference in appropriateness was small the letter "i" for indistinguishable was added. The indistinguishable label was always put on at least two recordings if applied. Recordings with extraneous sounds like coughing, stomach sounds or sounds from moving clothes etc was annotated with a "u" for unusable.

The criteria used for the evaluations of the triplets are (in no particular order):

- Evenly falling F0
- Equal length of phonation
- Equal consonant quality
- Equal vowel quality
- Equal length of vowel
- Equal length of pauses

All of these criteria were applied across the three repetitions of the nonsense syllables. Out of the 732 triplets evaluated the authors agreed on only 25% of the rankings. This negative result can be interpreted in two ways. First, to achieve a better agreement between evaluators it is a prerequisite that the criteria including order of importance be defined explicitly. In this interpretation the reported results are merely a result of poorly defined criteria for the subjective evaluation. The second interpretation is that the poor inter-evaluator agreement is symptomatic of the task in that it is inherently difficult to evaluate all of the criteria as requested in a consistent manner across a larger speech material. As a consequence, it is unclear whether the evaluators would be able to reproduce their own evaluations, let alone other evaluators'. With this interpretation it appears that an objective assessment would provide a sounder basis for consistent evaluations.

The subjective evaluation of whether triplets were suitable for use in perceptual experiments (i.e. marked "u") were as follows:

- Evaluator one alone: 5
- Evaluator two alone: 107
- Both evaluators: 50
- Total number of triplets marked with "u" by one of the two evaluators: 162
- Total number of triplets evaluated by each evaluator: 732

Evaluator 2 is clearly more critical than Evaluator 1. However, Evaluator 2 agrees with almost all Evaluator 1's rejections.

In the following we examine whether we can devise an automatic method, which would reject the same triplets as (either of) the human listeners (i.e. the 162 rejected triplets above). Moreover, we explore whether this method can be enhanced to rank triplets.

4. Objective evaluation

The aim of the automatic evaluation procedure presented here is to gauge the technical and, to some extent, linguistic consistency of the recorded material. For this the automatic evaluator uses three parameters derived from the acoustic signal: sound pressure level, F0 (when defined, i.e., for vowels and other sonorants), and harmonicity-to-noise ratio (HNR) [8]. Briefly, the HNR is defined as the ratio between the energy in the periodic parts of the signal and the energy in the aperiodic parts of the signal. This ratio is typically expressed in dB. We used the open-source PRAAT software [9] to compute these parameters. The phonetic constituents of each CVtu triplet are referred to as: C1 V1 t1 u1 <PAUSE1> C2 V2 t2 u2 <PAUSE2> C3 V3 t3 u3.

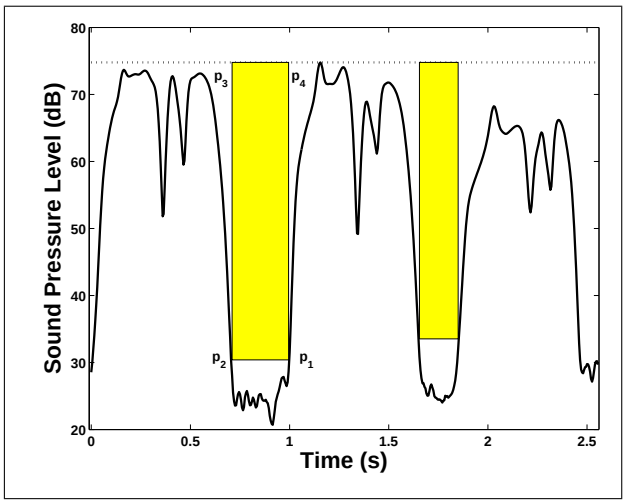

Figure 1. Rectangulation for determination of silent periods. The two silent periods in each triplet are determined by the two largest non-overlapping rectangles that can be drawn between the sound pressure level curve and the global maximum (represented by the upper dotted line). P_1, P_2, P_3 and P_4 are labels of the rectangle corners (see text for details)

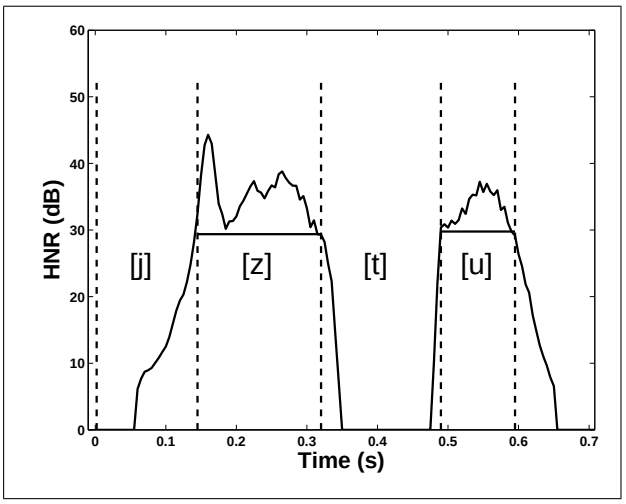

Figure 2. Rectangulation for vowel delimitation (c.f. Figure 1) The vertical lines illustrate the principle of rectangulation (see text for details)

Preparing the evaluator, each triplet has to be segmented into these 14 constituents . The segmentation is performed in two steps. The first step determined the two periods of silence <PAUSE1> and <PAUSE2>. The second step conisisted in delimiting the two vowels in each CVtu, i.e., V1, u1, V2, u2, V3 and u3, which in turn delimits the remaining segments six segments C1, C2, C3, t1, t2 and t3.

The method used for delimiting the periods of silence is called the principle of rectangulation and is illustrated in Figure 1. This principle works by defining a rectangle R with four corners P_1, P_2, P_3 and P_4 in the following way. For each point P_1 on a monotonically increasing part of the sound pressure level curve, point P_2 is defined as the closest point to P_1 on the curve with smaller abscissa and the same ordinate. P_3 has the same abscissa as P_2 and an ordinate corresponding to the maximum ordinate for the entire curve. P_4 has the same ordinate as P_3 and the same abscissa as P_1 (see Figure 1). The silence period is defined as the abscissas of P_1 and P_2 respectively from the rectangle with the largest area. Additional silence periods are defined in the same way except that no overlap with any previously identified silence periods is allowed.

Analogously, the two vowels of each CVtu is delimeted by applying the principle of rectangulation to HNR as shown in Figure 2 with two modifications. First, the P_1 is taken from monotonically **decreasing** parts of the curve rather than monotoniclly increasing parts. Second, instead of the global maximum of the ordinate, P_3 is determined as the lowest ordinates from the curve with abscissa equal to that of either P_1 or P_2 (cf. 2). Though it is also possible to delimit the vowels based on sound pressure level, applying HNR is straightforward, and in turn more accurate, since vowels and voiced consonants (sonorants) produce positive HNR as opposed to the obstruents. In general, vowel delimitation is less robust than delimitation of silence since, in an acoustic perspective, phonation ("vowelhood") is more vaguely defined than silence. The difference between a sonorant consonant like [w] and a relatively weak full vowel like [u] is often a matter of linguistic function rather than acoustic profile[3]. In consequence, the segmentation of C and V for sonorant Cs must be somewhat arbitrary.

The resulting delimitation of the C and t of each CVtu is indirectly given the delimitation of the V and u (cf. Figure2).

Based on the computed delimations of segments, all triplet recordings were analysed using the parameters duration, sound pressure level (dB), F0 (Hz for sonorant elements only), and HNR (dB). The sound pressure level, F0 and HNR shown in Table II were selected from the 5 ms interval which had the highest value in the interval. A sample from the evaluation log is shown in Table II.

Based on the evaluation log, a number of (language-specific) observations can be made: durations, sound pressure levels, and F0s (when defined) are generally decreasing through the triplet. Within the individual CVtu group, this pattern is repeated at a smaller scale: the stressed vowel (V) is generally longer and louder than the unstressed [u].

[3] In many languages, the division of phones into vowels and consonants is much less motivated than for the Germanic languages like English and Danish, including e.g. retroflex consonant-vowels, nasal consonant-vowels etc.

Table II. Sample from evaluation log: Female talker uttering [ma:tu]

Phone (IPA)	Duration (ms)	Intensity (dB)	F0 (Hz)	HNR (dB)
[m]	134	68	N/A	N/A
[æ:]	250	73	297	33.40
[t]	150	73	N/A	N/A
[u]	120	72	181	33.41
[m]	65	65	N/A	N/A
[æ:]	275	74	223	41.71
[t]	145	69	N/A	N/A
[u]	100	72	264	38.23
[m]	45	61	N/A	N/A
[æ:]	250	67	209	35.21
[t]	155	63	N/A	N/A
[u]	65	67	217	22.42

5. Correlation between objective and subjective evaluations

First question concerns the overall quality assessment: Can we predict, based on the objective data, which triplets are defective or otherwise unusable? Addressing this question, we studied the following symptoms of defectiveness:

1. Duration
   - **a)** Relative difference between the duration of <PAUSE1>) and <PAUSE2> exceeds 50%
   - **b)** Relative difference in duration of two in (V1, V2, V3) exceeds 50%
   - **c)** Relative difference in duration of two in (C1, C2, C3) exceeds 50%
2. Sound Pressure Level
   - **a)** Relative difference in intensity of two in (V1, V2, V3) exceeds 50%
   - **b)** Relative difference in intensity of two in (C1, C2, C3) exceeds 50%
3. Fundamental Frequency
   - **a)** F0 of V3 is greater than F0 of V1
   - **b)** F0 undefined for any vowel in (V1, V2, V3, u1 u2 u3) (e.g. 'creaky' vowel)
4. Harmonicity-to-noise ratio
   - **a)** HNR < 10 for any vowel in (V1, V2, V3, u1 u2 u3) (e.g. 'creaky' og otherwise unclear vowel and/or extremely short phonation)

Each of the criteria above counts for one error. Errors are counted for each CVtu triplet. CVtu's with two or more errors are interpreted as "unusable", corresponding to the "u" tag of the manual scoring procedure.

- Total amount of CVtu triplets scored: 732
- Triplets manually tagged as "u" (unusable): 162
- Triplets automatically tagged as "u": 129
- Triplets tagged as "u'" both automatically and manually: 115

The experiment shows a rather good correspondence between the two evaluation procedures: 115 out of 129 (89%)automatically applied rejections conform with human judgments.

6. CONCLUSIONS

Speech materials can be quite demanding to develop due to the load of manual evaluation and annotation work. At the time of writing, only about one quarter of the CVtu material has been evaluated by human listeners (six of 26 talkers).

The correlation of the objective evaluations of the speech material to the parts of the material which has been evaluated manually, is rather good in that 89% of the syllables rejected by human listeners is also rejected by the objective method. This is indeed encouraging and will prove useful in the evaluation of the remainder of the recordings. Moreover, since the proposed method is based on a relatively small amount of data, we speculate that improvements are indeed possible and even likely.

With respect to the ranking of triplets within talkers, the results are less conclusive. The perceptual evaluations were not consistent across evaluators. At present it is difficult to assess the degree to which this is due to in-homogenous evaluations criteria across evaluators or due to the complexity of the task. A prudent assumption would be that both factors play a role. In this case the proposed method is potentially of great value, since it provides a way of evaluating the phonetically relatively simple speech material objectively and consistently in a efficient way. The method used for rejecting triplets can easily be enhanced to propose ranking of triplets within talkers in that rejection criteria are based on continuous parametres which have been summarized in a reject/accept outcome.

Though far from complete, the current material shows - perhaps unsurprisingly - that not all our talkers are succinct and consistent in performance. Identifying and expelling the inadequate talkers in an early automatic evaluation would be extremely helpful. Our next goal is thus to develop the methods put forward here, into a device for unsupervised talker evaluation. Not only will this save a lot of time, screening based on acoustic criteria is less prone to subjective and biased judgments, hence far more reproducible.

Acknowledgement

This project has been funded by the Danish Research Council for Culture and Communication.

References

[1] G. A. Miller and P. E. Nicely. An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, 22(2):338–352, March 1955.

[2] H. Fletcher. *Speech and Hearing in Comunication, The ASA edition*. Acoustical Society of America, 1995.

[3] M. Ardoint, S. Sheft, P. Fleuriot, S. Garnier, and C. Lorenzi. Perception of temporal fine-structure cues in speech with minimal envelope cues for listeners with mild-to-moderate hearing loss. *International Journal of Audiology*, 49(11):823–831, 2010.

[4] L. Lisker and A. Abramson. A cross-language study of voicing in initial stops - acoustical measurements. *Word-journal of the international linguistic association*, 20(3):384–422, 1964.

[5] *The Handbook of the International Phonetic Association*. Cambridge University Press, 1999. ISBN 9780521637510.

[6] N. Gr nnum. Illustrations of the ipa: Danish. *J. Int. Phon. Assoc.*,, vol. 28:99?105, 1998.

[7] F. Ingerslev, O. J. Pedersen, P. K. Møller, and J. Kristensen. New rooms for acoustic measurements at the danish technical university,. *Acustica*, 19:185–199, 1968.

[8] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences, Amsterdam*, volume 17, pages 97–110, 1993.

[9] P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001.