Development of the A§E test battery for assessment of pitch perception in speech

Willemijn Heeren[1], Martine Coene[2], Bart Vaerenberg[2,3], Andrei Avram[4], Anna Cardinaletti[5], Luca del Bo[6], Alexandru Pascu[4], Francesca Volpato[5], Paul J Govaerts[2,3]

[1]Leiden University, The Netherlands, [2]University of Antwerp, Belgium, [3]The Eargroup, Antwerp-Deurne, Belgium, [4]Bucharest University, Romania, [5]Ca' Foscari University, Venice, Italy, [6]DelBo Tecnologia Per L'Ascolto, Milan, Italy

Objectives: The auditory speech sounds evaluation 2009 test battery for assessment of speech pitch perception is presented. It was designed to (a) assess perception of pitch in linguistic contexts without the confounds of secondary acoustic cues, (b) be usable with listeners from different language backgrounds, and (c) be suitable for use in a clinical setting. The need for this test battery arises from increased awareness of the importance of prosody in clinical practice, and the development of methods for improving pitch perception in listeners with profound hearing losses.
Methods: Identification and discrimination tasks based on linguistic contexts were developed to establish listeners' just noticeable differences (JNDs) for pitch changes. Stimuli were pseudosentences and pseudowords based on speech from a female speaker, overlain with stylized pitch contours. Target pitch excursions were varied from the 200 Hz baseline to a maximum of 349 Hz. Ninety normal-hearing listeners participated in test validation that assessed goals (a)–(c), established test–retest reliability, and gathered normative data.
Results: The JNDs on non-linguistic, control tasks were lower than on linguistic ones, showing that non-linguistic tasks may overestimate pitch perception in speech. Listeners from different language backgrounds scored comparably on most linguistic tasks, and test–retest differences were non-significant. Test usability as evidenced by task duration and subject experience seemed satisfactory for clinical use.

Keywords: Pitch perception in speech, Intonation perception, Test development

Introduction

As part of the speech signal, pitch contributes to syntactic and semantic disambiguation (e.g. Kuo *et al*., 2008), to discourse structure, by, for instance, marking new versus given information (e.g. Swerts *et al*., 1994; Savino, 2004), and to clause typing, by marking a phrase as a statement or a question (e.g. Van Heuven and Haan, 2000). It furthermore helps to track speakers in competing speech (e.g. Brokx and Nooteboom, 1982; Assmann, 1999), and provides information on speaker characteristics such as dialect, gender, and emotion (e.g. Vroomen and Collier, 1993; Bachorowski and Owren, 1999). Also, early in life, prosody – including pitch – may help infants to start identifying word boundaries in continuous speech (Jusczyk, 1997).

Several types of hearing-impaired listeners have reduced pitch perception. Cochlear implant (CI) users, for example, reach high levels of speech intelligibility for sentences in a quiet background, but pitch perception is reported to be suboptimal with current devices. It has repeatedly been shown that adult CI users are significantly worse at musical perception of pitch and melody recognition than normal-hearing adults (e.g. Gfeller *et al*., 2002; Kong *et al*., 2004; Laneau *et al*., 2006; Sucher and McDermott, 2007). Also, in speech perception, CI users have difficulties perceiving intonation (e.g. Green *et al*., 2004; Meister *et al*., 2007) and lexical tones (e.g. Barry *et al*., 2002; Ciocca *et al*., 2002), especially when the speaker's pitch is relatively high, such as for women and children (Green *et al*., 2004; Chatterjee and Peng, 2008).

While there is an increased awareness of the importance of prosody perception in clinical settings and new methods to improve pitch and music perception in listeners with profound hearing losses are being developed (e.g. electric-acoustic stimulation (EAS)), a need to measure (improved) perception of speech pitch in clinical contexts is emerging. Most clinical

Correspondence to: Paul J Govaerts, The Eargroup, Herentalsebaan 75, Antwerp-Deurne B-2100, Belgium. Email: dr.govaerts@eargroup.net

DOI 10.1179/1754762811Y.0000000035

tests were designed to measure segmental perception (e.g. Kalikow *et al*., 1977; Plomp and Mimpen, 1979), but only few prosody perception tests are available. For English, the minimal auditory capabilities test battery includes subtests that measure patients' prosodic perception (Owens *et al*., 1981). A more recent development was undertaken for German by Meister *et al*. (2007), who developed six tests to assess prosody perception in CI users. For different varieties of English as well as a number of other languages there are versions of the PEPS-C test for testing prosody in children (Peppé and McCann, 2003; Peppé *et al*., 2010). These tests can be used to measure the perception of *prosodic* information in speech, but not the perception of *pitch per se*.

As we want to be able to measure how well listeners can exploit pitch information in speech, a new test battery for measuring speech pitch perception was developed. It is an extension of the auditory speech sounds evaluation (A§E) test (Govaerts *et al*., 2006). The main goals of the new test battery are: (a) to assess perception of pitch information in linguistically relevant contexts, (b) to be usable with listeners from different language backgrounds, and (c) to be sufficiently easy and short for use in clinical practice. The tests presented here differ from those developed earlier in two main respects. First, the stimulus materials in the new tests only vary in pitch, and do not contain co-varying, secondary cues. Second, the new tests were designed such that they can be used with listeners from a number of different language backgrounds, making them more widely applicable than existing ones.

In the rest of this paper the design and development of the test battery are first presented, followed by a validation based on a check of the three aforementioned goals using normal-hearing listeners. In future applications of tasks from the test battery the normal-hearing listeners' results can be used as normative data. In the discussion the implications of this validation for further development of the test battery as well as first results from hearing-impaired listeners are presented.

Methods

The goal of the prosodic tests is to assess listeners' perception of pitch in linguistic contexts. This aim was pursued by developing tests that estimate listeners' just noticeable differences (JNDs) for pitch changes in speech stimuli modeled after linguistically relevant situations.

The tests were designed to be usable with listeners from three different language backgrounds, targeting both Romance languages (Italian, Romanian) and a Germanic one (Dutch). The prosodic tests were based on two linguistic functions that can be conveyed by pitch movements and that occur in each of these languages: (a) clause typing, i.e. marking a phrase as a statement or a question by a pitch movement on the utterance's final syllable, and (b) lexical stress, i.e. the differentiation between word meanings of sound sequences containing the same segmental order, but with prominence on different syllables.

(a)	IT	Il tavolo è sporco./?	'the table is dirty ./?'
	NL	De staking is voorbij ./?	'the strike is over ./?'
	RO	Casa arde ./?	'the house is on fire ./?'
(b)	IT	′principi – prin′cipi	'princes – principles'
	NL	′voorkomen – voor′komen	'happen – prevent'
	RO	′imobil – imo′bil	'building – immobile'

Through intonation only many languages can indicate the difference between statements and questions. A statement is associated with a low boundary tone, and a question is associated with a high one (e.g. Pierrehumbert, 1980). Question/statement identification is less accurate in CI users. A study using natural stimuli yielded 80% correct responses from patients as opposed to near-perfect scores for normal-hearing listeners (Meister *et al*., 2007). Somewhat lower scores, 70–75% correct, were obtained by Green *et al*. (2005). When stimuli for question/statement classification were taken from a continuum along which pitch direction changed from falling (statement) to rising (question), CI users showed shallower psychometric functions than normal-hearing controls (Chatterjee and Peng, 2008).

Correct perception of lexical stress may be crucial for semantic disambiguation, and is also thought to facilitate the recognition of words (Cutler, 2007). A lexically stressed syllable, i.e. the most prominent syllable in a word, is not necessarily marked by an F0 movement, but in its canonical form, or when being introduced in a sentence, F0 marking is generally present on the lexically stressed syllable. When comparing an F0 movement to other cues that may indicate the location of lexical stress, such as duration and intensity, it has been shown that in English F0 is able to override the others (Fry, 1958). When comparing the trade-off between cues in the languages under study, Dutch and Romanian seem to follow this general trend (Avram, 1970; Van Katwijk, 1974). In Italian duration has been indicated as the most important cue (Bertinetto, 1980), but it has also been argued that this is the case especially when combined with F0 (Alfano, 2006). Spitzer *et al*. (2009) found that CI users seem to exploit stress cues for segmentation of the speech stream, and also that access to F0 information helped EAS listeners in their task.

Test design

The tests developed from these linguistic contexts will henceforth be referred to as the *sentence intonation*

(SI) test, addressing clause typing, and the *word stress pattern test*, addressing lexical stress.

The SI test

A same–different discrimination paradigm was used in which the listener hears two consecutive sentences, one of which has a final rise (A). The other sentence can be either exactly the same (A) or different (B), that is, without a final rise. The listener's task is to indicate whether the sentences were the same (AA) or not (AB or BA).

Each sentence was modeled by a sequence of four-to-six syllables, thus varying the position of the target syllable. Over each syllable sequence a pitch contour was overlain with, in all cases, a fixed pitch accent on the second syllable, and a variable-sized final rise on the last syllable (see Fig. 1A). The second syllable carried the fixed pitch accent to have at least one pitch accent per sentence in addition to the boundary tone (e.g. Pierrehumbert, 1980). This pitch accent was set to a pointed hat (H*L) with a maximum excursion of 40 Hz (3.15 semitone (ST)) from the 200 Hz baseline (female speaker). This excursion size falls within the range of pitch accent excursions found across a number of languages (e.g. Campione and Véronis, 1998); a minimum of 1.5–3 STs is needed to convey linguistic meaning (Gussenhoven and Rietveld, 1985). The final rise was varied in size from a flat ending that remained at 200 Hz to a rise of 149 Hz above the baseline, i.e. 349 Hz. The resolution of the steps was 1/12 ST until 208 Hz, 1/6 ST until 230 Hz, and 1/3 ST over 230 Hz. This resulted in 41 stimulus levels.

With the goal of establishing JNDs for pitch excursions in linguistically relevant contexts an adaptive one up–one down procedure was adopted (Levitt, 1971) that estimated the 50%-point on a participant's psychometric function. Both stochastic processes and internal controls were used to determine the exact number of reversals needed for good threshold estimation per listener, which was preferred over the use of a fixed number of reversals. The procedure started at a relatively large ΔF of 41 Hz, and either decreased ΔF after discrimination of the two intervals or increased when the participant failed to discriminate the stimuli.

The word stress pattern test

Each word was modeled by a three-syllable sequence. A four-category identification task was used; the listener indicated which of the three syllables of the nonsense word carried a pitch accent, or that there was no noticeable accent at all. Fig. 1B shows the model of pitch accents on the word in the case that the second syllable is accented. The possible sizes of the accent were taken from the same series as used for the sentence test. The same adaptive staircase procedure was used for threshold estimation.

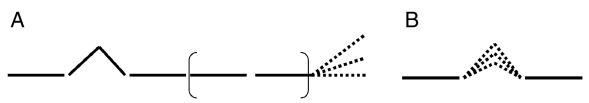

Figure 1 Models of the sentence pitch contour (A), and the word pitch contour (B). The latter illustrates possible pitch movements on the second syllable only.

Stimulus materials

The speech sounds for the stimuli were determined by comparing phoneme inventories and syllable forms across the three languages. Statistics on syllable type were gathered by analyzing over 10 000 syllables per language using the different translations of the Lisbon Treaty (URL: http://eur-lex.europa.eu/, last visited 16/09/2010). This resulted in the choice of consonant – vowel (CV) as syllable type, which occurred in 34% of the Dutch syllables to 54% of the Italian ones. At the segmental level, many CVs contained combinations of phonemes present in each of the three languages. The added requirements of using voiced, sonorant speech sounds (to allow stimuli to carry pitch continuously) that are furthermore robust toward between-language and within-language variation resulted in the selection of six syllables: /mi, ma, mu, ni, na, and nu/.

Sentence and word forms were based on the models from Fig. 1. For the sentences, 3 lengths (4, 5, 6 syllables) × 4 forms per length were made, and for the words there were 10 different three-syllable forms. Syllable occurrence was balanced out (Appendix 1). All speech editing and analysis were done using the program Praat (Boersma and Weenink, 2008).

To generate all three-to-six-syllable CV sequences, a grammar consisting of diphones and triphones was constructed. This allows for the maintenance of natural formant and intensity transitions. Units were chosen as long as possible, thus reducing the number of locations where irregularities in the audio may arise. There were three types of units: onsets, mid-syllables, and offsets. Onset diphones consisted of either [m] or [n] preceded by silence (*#m-, #n-*). Mid-syllables consisted of triphones beginning in [m] or [n], followed by a full vowel, and ending in [m] or [n] (*-mVm-, -mVn-, -nVm-, -nVn-, where* $V = \{/i, a, u/\}$). Offsets were also triphones, but ended in silence (*-mV#, -nV#, where* $V = \{/i, a, u/\}$).

A word list containing all phones was made, and recorded with a female speaker (Dutch native, trained phonetician). She read the word list using relatively flat intonation. The recordings were made directly onto the computer (44.1 kHz, 16 bits) using a Sennheiser MKH 416T directional condenser microphone. The di- and triphones were cut from the recordings in the middle of the consonants, and at zero crossings, such that wave forms started with a

movement toward the positive maximum, and ended in a rise from the negative minimum. Concatenation would then result in a smooth continuation of the wave form.

Duration and intensity were normalized. For duration normalization, the original phoneme durations in each of the di- and triphones were first measured. Next, target durations were set to the average durations of the phonemes in the mid-syllable triphones. A comparison of the original and target durations is given in Appendix 2. Durations were manipulated by cutting or adding periods of the speech signal. If the difference between original and target durations was small, manipulation was done in the middle of the speech sound, but if the difference was larger, manipulation was spread throughout the phoneme. Note that the length of 151 millisecond for final vowels was measured where the vowel's intensity was not more than 6 dB under the stimulus' average intensity of 84 dB. The duration of offset triphones was set to 270 millisecond each to arrive at equal inter-stimulus intervals during testing. All duration variation of the normalized phones lay within one period, about 5 millisecond, from its target duration. The phones were stored in separate wave files.

To exclude effects of syllable intensity on perception, the phones' intensities were normalized per position. The mid-triphones were scaled to an average intensity of 84.0 dB, offset triphones were scaled to a lower mean intensity, 82.4 dB, to not boost intensity in the first part of that triphone as the second part would consist of a reduction to silence. For similar reasons, the onset diphone was scaled to a mean intensity of 80.0 dB.

Word and sentence forms were made using this phone set, and downsampled to 16 kHz. Next, pitch contours with a 200 Hz baseline were computed for the concatenated audio files. Through PSOLA re-synthesis each stylized contour was substituted for the file's original pitch contour. The pitch accents show a peak at 50 millisecond after vowel onset. The final rises were aligned with the end of the voicing in the final syllable, and had a duration of 120 millisecond after 't Hart *et al*. (1990, p.73).

Each word or sentence was saved to disk, resulting in 504 sentence stimuli (3 sentence lengths × 4 forms per length × 42 pitch size variants, including 0 Hz) and 1240 word stimuli (10 word forms × 3 pitch locations × 41 pitch size variants + 1 default, i.e. flat, contour per word form). An independent check of the materials' acoustic contents showed that stimuli varied in pitch, but not in duration or intensity.

A set of low-pass-filtered stimuli was also generated. This was done under the assumption that the critical information in the stimuli is available in the lower frequencies. Each word and sentence stimulus was low-pass filtered (The Filter() function implemented in MATLAB was used 300 Hz cut-off frequency, 90 dB attenuation in magnitude over a 50 Hz transition width) and high-pass-filtered white noise was added (250 Hz cut-off frequency, 85 dB gain in magnitude over a 50 Hz transition width).

Test validation and normative data collection

The validation assessed the main goals of the test battery: (a) measure perception of pitch in linguistically relevant contexts, (b) be usable with listeners from different language backgrounds, and (c) be sufficiently easy and short for use in clinical practice. This was evaluated with normal-hearing listeners in audiology centers in Belgium, Italy, and Romania.

First, the tests were designed to assess perception of pitch information, where crucial information is contained in the low frequencies, that is under 300 Hz. This entails that listeners are expected to show comparable behavior on low-pass-filtered versions of the speech materials. Low-pass filtering only maintains the frequencies in which the fundamental frequency (F_0) is contained, while suppressing the higher harmonics. The results of both speech tests were compared in parallel tests with results on low-pass-filtered stimuli.

One of the assumptions underlying the development of this test battery is that pitch perception tests using synthetic complex sounds may not be fully representative to assess the perception of pitch in speech contexts. The idea is that speech stimuli may be processed differently by the human listener than non-speech stimuli. This assumption predicts a performance difference between the speech and non-speech tasks. To make this comparison, three synthetic tone complex discrimination tasks were added to the test battery (details are given in the next section). In addition, the correlation coefficients between test outcomes were determined to assess the question to what extent scores on one (type of) test can be predicted from scores on another.

Second, the speech tests were designed for use with listeners from different language backgrounds. On the non-speech tests listeners are expected to perform comparably, irrespective of language background. For the speech tests, small differences in group performance may be found, as it is probably not the case that pitch is weighed similarly in each of the languages, even though the linguistic phenomena on which the tests were built exist in each of the three languages. Potential differences are not expected to be very large, though.

Third, to assess the usability of the tests, task durations were measured and listener feedback was gathered through posttest questionnaires. Additionally, test–retest reliability was assessed by retesting one-third of the listeners. The results obtained with

normal-hearing listeners may serve as normative data for future use.

Validation method

Ninety normal-hearing listeners participated, 30 from each language background (Dutch, Italian, and Romanian). Participants gave written informed consent. Normal hearing was screened through tonal audiometry (hearing loss <20 dB on 0.125–8 kHz). Participants were between 18 and 53 years old (evenly distributed over gender). Twenty-nine listeners, equally divided over language backgrounds, and per language background equally divided over the genders, returned for a re-test.

The test battery contained seven tasks: the two speech tests, word stress pattern (WSP) test and SI test, a low-pass-filtered version of each of these tests, and three-tone complex discrimination tasks: harmonic complexes, harmonic intonation, and disharmonic intonation.

The first tone complex discrimination task, harmonic complexes (HCs), estimates the JND for discrimination of level tones. Harmonic and disharmonic intonation (HI and DI) estimate JNDs for discrimination of tone changes by presenting harmonic or inharmonic pitch glides. All stimuli in the non-speech tests were 600 millisecond in duration and had an F0 of 200 Hz (i.e. the speaker's F0). The intensity of the harmonics decreased compared to F0 (−6 dB at 400 Hz, −12 dB at 600 Hz, and −18 dB at 800 Hz). White noise was added to each non-speech stimulus complex (signal-to-noise ratio +10.9 dB) to make them sound more natural and easy to listen to. The glides were modeled after the intonation movements in the SI task, showing the same change rate as the speech stimuli, that is a 120 millisecond linear sweep that started 270 millisecond before the end of the stimulus. In the HI task the three harmonics co-varied with F_0, but in the DI task only F_0 changed, whereas the higher harmonics remained unchanged. These two variants can be compared with the unfiltered and the filtered versions of the speech tasks.

In all discrimination tasks a 500 millisecond inter-stimulus interval was used, and stimulus intensity was varied in a roving manner (±2 dB). To prevent effects of learning, test orders were counterbalanced across listeners. For test–retest reliability, a subset of listeners completed the test battery twice with an interval of minimally 1 week. Tests were presented in the same order during the two test sessions.

Procedure

Participants were tested individually, seated in a sound-treated booth facing a loudspeaker. The tester remained outside the booth. Test items were played at 70 dB HL. For the WSP tasks the participant was instructed to indicate on which syllable a pitch accent was perceived or to indicate that no accent was perceived at all. For all discrimination tasks the participant was instructed to indicate if the two stimuli were the same or not.

Each of the seven tests started with a training module to familiarize the participant with the procedure and the stimuli. During training some of the sounds or sound pairs from the actual test were presented, and ΔF levels were either set by the tester or through an automatic training mode. The maximum training time per test was 10 minutes. During the test phase, participants in general received no feedback on the correctness of their responses. However, in the case of a false positive an alarm buzz was played to discourage listeners from reporting non-existent differences, and the tester reminded the participant to only indicate the presence of a rise or pitch accent when it was reliably detected.

The adaptive algorithm continued to present stimuli until the threshold was reached, and then automatically ended the test. When the maximum of 100 trials was reached before a JND was computed, the test was also ended. Short pauses were given between tests. On completion of the test battery, participants filled up a questionnaire, expressing their experiences by judging statements that they evaluated along a 5-point Likert scale from *fully disagree* to *fully agree*. In total, 88 questionnaires were gathered (30 NL, 29 IT, 29 RO).

Analysis

Per task and per listener, a JND was obtained in hertz. In 15 out of 816 cases (i.e. 2%) the JND was set to the maximum value of 149 Hz (no JND found within 100 trials). Fourteen (2%) scheduled tests were not run, as testers forgot to run a task (13 cases), or the listener chose to discontinue (1 case).

One-sample Kolmogorov–Smirnov tests showed that results were not normally distributed, $2.3 < Z < 7$, $P < 0.001$. Therefore, the median was taken as a representative of central tendency, and the research questions were assessed using non-parametric statistics. To obtain a measure for test–retest reliability, signed differences were computed for each test–retest pair. The significance level alpha was set to 0.05, and multiple comparisons were Bonferroni-corrected.

Results

An overview of the normative JNDs per test and per language background is shown in Fig. 2. Table 1 summarizes the first through third quartiles.

Speech versus non-speech tasks

Average JNDs for speech versus non-speech tests were subjected to Wilcoxon signed ranks tests for related

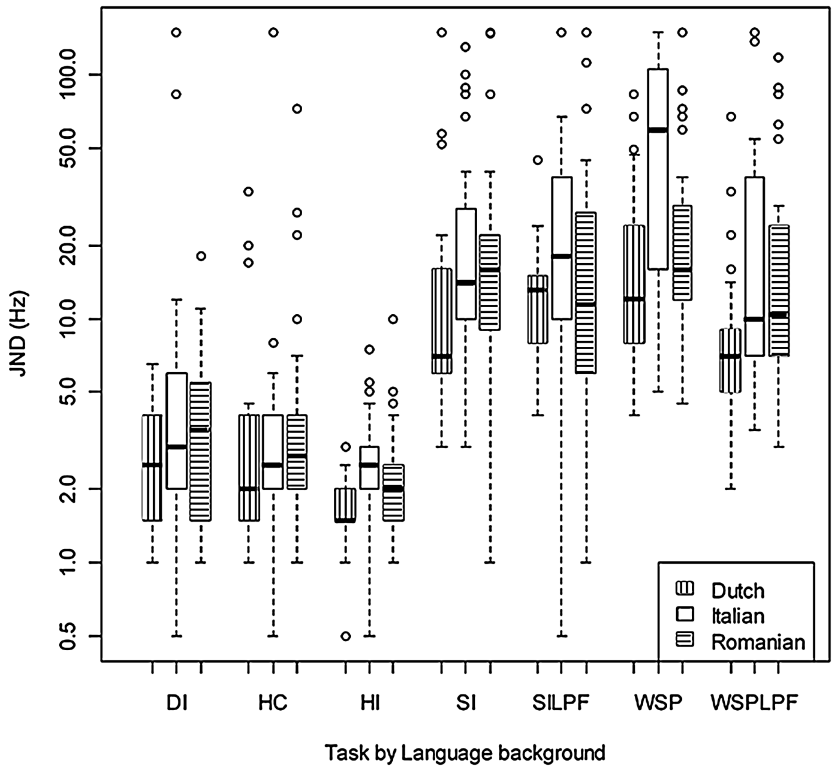

Figure 2 JNDs in hertz per test and per language background (DI, disharmonic intonation; HC, harmonic complexes; HI, harmonic intonation; SI, sentence intonation; SILPF, sentence intonation low-pass filtered; WSP, word stress pattern; WSPLPF, word stress pattern low-pass filtered).

Table 1 P25, P50, and P75 values in hertz for test and retest, and per language background

Task	Dutch			Italian			Romanian		
	P25	P50	P75	P25	P50	P75	P25	P50	P75
Harmonic complexes	1.5	2.0	4.0	1.8	2.5	4.5	2.0	2.8	4.5
Retest	1.4	1.8	3.5	1.3	2.0	3.8	0.9	1.3	2.3
Harmonic intonation	1.5	1.5	2.0	2.0	2.5	3.3	1.5	2.0	2.8
Retest	1.0	1.3	1.6	0.8	1.5	2.3	1.4	1.5	2.5
Disharmonic intonation	1.5	2.5	4.0	2.0	3.0	6.0	1.5	3.5	5.9
Retest	1.4	1.8	2.5	0.8	2.0	3.5	1.0	1.8	3.3
Sentence intonation	6.0	7.0	16.8	9.5	14.0	30.5	8.8	16.0	22.0
Retest	3.4	4.5	12.3	3.5	12.0	22.5	4.8	7.0	11.3
Sentence intonation LPF	7.9	13.0	15.3	10.0	18.0	42.5	6.0	11.5	27.0
Retest	5.8	6.3	7.3	4.5	10.8	17.5	4.0	7.5	11.8
Words stress pattern	8.0	12.0	25.8	16.0	59.5	111.3	11.5	16.0	31.3
Retest	5.9	7.0	23.5	10.0	38.0	82.0	10.5	17.0	25.5
Word stress pattern LPF	5.0	7.0	9.3	7.0	10.0	41.3	6.8	10.5	25.3
Retest	4.0	4.8	7.5	4.8	9.0	92.3	4.4	7.0	11.0

Test data were gathered from 90 listeners, retest data from 29 out of 90.

samples. Listeners got lower JNDs for the non-speech (2.5 Hz) than the speech tests (16.9 Hz), $Z = -8.1$, $P < 0.001$. Per language background, the same pattern of results was found: Dutch, $Z = -4.6$, $P < 0.001$; Italian, $Z = -4.6$, $P < 0.001$; Romanian, $Z = -4.8$, $P < 0.001$.

Cross-linguistic comparison

Listener performance per test was compared between the different language backgrounds. Kruskal–Wallis non-parametric analysis of variances showed significant differences between language backgrounds for the harmonic SI test, $\chi^2 = 13.7$, df = 2, $P = 0.001$, and the WSP test, $\chi^2 = 13.3$, df = 2, $P = 0.001$. On the HI task, higher median JNDs were found for Italian listeners, 2.5 Hz, as compared to the Dutch, 1.5 Hz ($Z = -3.9$, $P < 0.001$). On the WSP task, higher JNDs were also found for Italians, 59.5 Hz, as compared to both the Dutch listeners, 12 Hz ($Z = -3.4$, $P = 0.001$), and the Romanians, 16 Hz ($Z = -2.7$, $P = 0.007$).

Tests with filtered versus unfiltered materials

Wilcoxon signed ranks tests for the type of prosodic contrast and the effect of filtering show that JNDs did not differ between the versions of the SI test

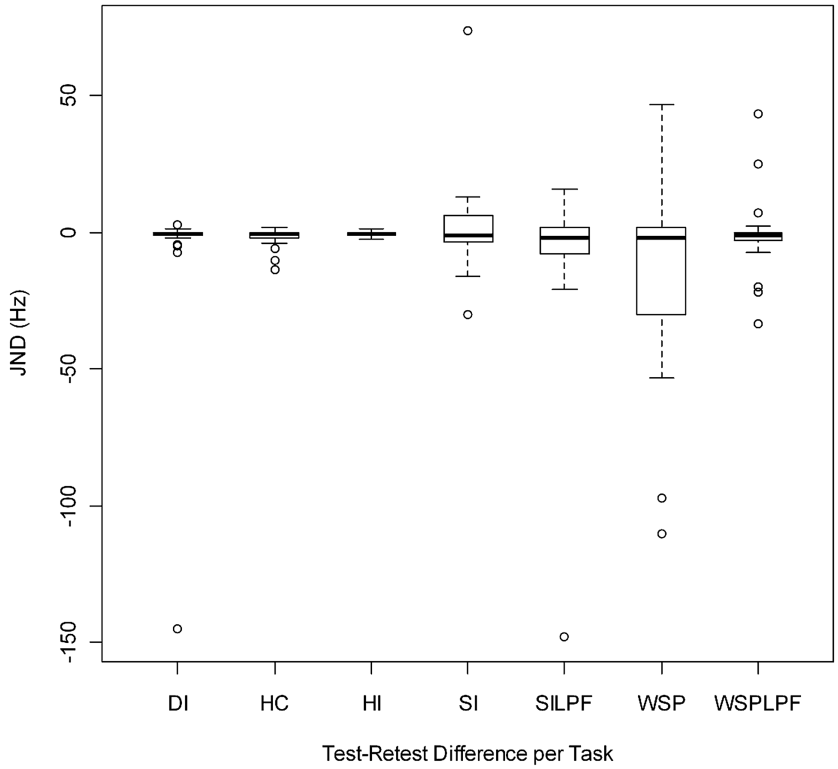

Figure 3 JND test–retest differences, shown per task.

($Z = -0.3$, n.s.). A difference was found, however, between two versions of the WSP test ($Z = -4.4$, $P < 0.001$): the filtered task yielded lower JNDs than the unfiltered one. Analyses on the WSP tasks per language background showed lower JNDs on the filtered than the unfiltered test for both the Dutch ($Z = -3.2$, $P = 0.001$) and the Italian listeners ($Z = -2.8$, $P = 0.005$).

Test–retest reliability

Fig. 3 shows the signed test–retest differences per task. Wilcoxon signed ranks tests revealed that these differences only significantly diverged from zero for the HC test ($Z = -3.1$, $P = 0.002$). For the non-speech tasks, the inter-quartile range (IQR) was from −2.0 to +0.25. For the speech tasks this IQR was larger, $Z = -4.2$, $P < 0.001$, but most noticeably for the WSP task, that is −32.9 to +1.9, whereas across the other three tasks the range was from −8.0 to +6.5.

Correlations between test scores

To assess the question of whether there was a difference in difficulty between the speech tests the Friedman test was run, the non-parametric equivalent of a repeated measures analysis. A significant effect of test was found ($P < 0.001$), and *post hoc* Wilcoxon signed ranks analyses showed that this difference is explained by a significantly higher JND for the WSP test in comparison with all other tasks ($P <= 0.004$). In general, the WSP test was more difficult than the SI test. However, within individual language backgrounds this effect was only found for the Italians ($Z = -3.0$, $P = 0.002$).

Table 2 Correlation matrix showing the correlations between tests

	DI	HC	HI	SI	SILPF	WSP	WSPLPF
DI	–	0.294	0.479	0.385	0.421	0.261	0.344
HC	0.294	–	0.313	0.556	0.467	0.476	0.541
HI	0.479	0.313	–	0.457	0.403	0.340	0.365
SI	0.385	0.556	0.457	–	0.489	0.442	0.471
SILPF	0.421	0.467	0.403	0.489	–	0.294	0.422
WSP	0.261	0.476	0.340	0.442	0.294	–	0.418
WSPLPF	0.344	0.541	0.365	0.471	0.422	0.418	–

To assess the question to what extent scores on one test can be predicted from scores on another, correlations between test results were computed using Spearman's rho (Table 2). All correlations were positive: a low JND on one task patterns with lower JNDs on other tasks, and a high JND on given task patterns with higher JNDs on other tasks. All correlations were significant, but DI-WSP.

Task durations and questionnaires

Median test durations per language background for the WSP tasks were between 2 and 3 minutes with IQRs varying from 1 to 2 minutes. For the SI tasks median durations were 3–5 minutes (with 1–6-minute IQRs). For the non-speech tasks, median durations were 2 minutes in all cases and IQRs varied from 0 to 2.8 minutes. The SI tasks were longest, which is explained by their long stimuli (four to six syllable pseudosentences) in comparison with those of other tests. Across tests the minimum duration was 1 minute and the maximum ranged from 6 (WSP) to 14 minutes (HC).

Questionnaire responses are summarized in Appendix 3. When comparing between listeners from different language backgrounds, and correcting the significance level for the number of comparisons per topic (instructions, test experience, false alarms, stimuli), five significant differences were found. Whereas listeners generally rated the training as *not confusing*, the Italians were somewhat less extreme in their rating than the Romanian listeners (1.5 versus 1; $Z = -3.1$, $P = 0.002$). Romanian and Dutch listeners rated the tests of moderate ease, but the Italians found them easy more often than the Dutch (4 versus 3, $Z = -3.2$, $P = 0.001$). The false alarm sound was found more startling by the Dutch than by the Italians (4 versus 2; $Z = -2.9$, $P = 0.004$). To the Dutch the stimuli sounded more like words and sentences than to Italian listeners (4 versus 2; $Z = -3.2$, $P = 0.002$), and stimulus naturalness was rated somewhat higher by the Dutch than by the Italians (both medians were 2; $Z = -3.5$, $P < 0.001$).

Discussion

The main goals of the validation were (a) to test whether the test battery assesses perception of pitch information in linguistically relevant contexts, (b) to assess the test battery's use with listeners from different language backgrounds, and (c) to check that it is sufficiently easy and short for use in clinical practice. In addition, test–retest reliability was investigated.

The JNDs were higher on speech tasks than non-speech tasks. This supports the idea that separate, linguistically based tests are justified for the assessment of perception of pitch information in speech. Thresholds for the perception of intonation (pitch glides) in tone complexes seem to overestimate listener performance on intonation perception in speech. This is supported by the medium correlations that were obtained between tasks. JNDs for non-speech tasks were comparable to those reported in the literature (Green, 1976, p. 262). The performance difference between the two test types may have been caused by several factors. The speech stimuli differed from the non-speech ones in both the type of content (tone complexes versus multiple syllables), and their length (600 versus 886–1638 millisecond). These two dimensions are related, as longer stimuli have more ecological validity for speech perception than short ones; Speakers' utterances are generally longer than 600 millisecond. Also, discrimination of longer stimuli puts higher demands on auditory short-term memory, which may have influenced performance, e.g. Pisoni (1973). Moreover, in the perception of speech different dimensions are integrated, e.g. segmental and supra-segmental information. These compete for attention, even when only one dimension is relevant to the task (e.g. Carrell *et al.*, 1981; Repp and Lin, 1990). Perception of pitch changes in speech stimuli may therefore inherently pose more of a challenge to listeners.

To investigate whether the speech tests measured perception of information contained in the lower frequencies both a filtered and a non-filtered version of the speech tests were presented. The prediction was that JNDs on non-filtered tests should not be lower, thanks to availability of other cues. JNDs on the parallel tests were comparable for the SI task, but not for the WSP task. For both Italian and Dutch listeners, the unfiltered version of the WSP task was more difficult, though by different degrees. Note that the direction of the difference did not go counter to our prediction: listeners were not better on the non-filtered task. For Dutch and Italian listeners, the benefits of having the harmonics present in the unfiltered speech stimuli seemed to be outweighed by other aspects of the signal. The perceptual integration explanation mentioned earlier may account for this result. Hearing out the pitch in non-filtered speech stimuli may be more difficult as it is embedded in ongoing, but irrelevant segmental changes in the acoustic signal, such as the formant structure. Results supporting such an explanation were reported in Klatt (1973), who found small increases in JND for pitch when the pitch movement was presented in the syllabic context with formant changes (/ya/) instead of in a steady vowel (/ɛ/). More recently, a comparable explanation was forwarded in Green *et al.* (2004), who found that perception of temporal pitch cues worsened when changes in formant structure introduced spectral variation. Moreover, bias effects occurring in speech perception, such as the intrinsic pitch of vowels and language-dependent stress position biases, may have made the unfiltered tasks more difficult for listeners. However, the difference between filtered and non-filtered stimuli was only found for the WSP task, and only for a subset of the listeners.

Another question was whether the JNDs found for the various tests were comparable between language backgrounds. For most tests no differences were found, but there were two significant effects, one on a non-speech task and one on a speech task. The effect found for Harmonic Intonation is unlikely to be explained by the listeners' language backgrounds; actual differences lie in the range of 1 Hz only and may therefore be ignored when listeners' real world speech communication is considered. The Italian median score for the WSP task stood out (59.5 Hz), whereas the Dutch and Romanians got much lower and comparable JNDs (12/16 Hz). Interestingly, in the low-pass-filtered version of this test no difference between language backgrounds was found.

The performance difference in the unfiltered WSP task may be a result of language background differences

between listeners. There may, however, also be a secondary explanation for the differences in test performance; in this validation setup language background coincided with test location and with tester. The differences may therefore partially have resulted from small procedural differences between the testing sites, for instance, in the amount of training provided (even though a standardized protocol was used). On the one hand, a difference in training amount would be expected to affect not just one, but all tests, and this was not the case. On the other hand, such a difference might only show up in the most difficult task, which the WSP test may be considered to be: it is an identification task instead of a discrimination task, in which speech stimuli are presented, instead of tone complexes.

To further investigate the contribution of test location a follow-up study was run. One tester collected test–retest WSP data from 26 listeners of the three language backgrounds (10 NL, 8 IT, 8 RO). The design and procedure were the same as in the main experiment; test/retest data were collected with an interval of minimally 1 week. In all instances, the automatic training mode of the A§E software was used. Wilcoxon-signed ranks tests showed no significant test–retest differences: the median JND was 13.5 Hz for the first and 10.5 Hz for the second moment of testing, and Kruskal–Wallis tests showed no differences between the language backgrounds. These findings support the hypothesis that part of the differences found for the WSP task may be explained by tester/training variation, and are effectively reduced by use of the standardized, automatic training mode.

We cannot rule out, however, that differences in language background and therefore linguistic experience contributed to the variation in the WSP results. Italian listeners may need more training on this particular task when it is less natural for them than for the other listeners. Additional evidence for this view came from the test–retest data where Italians showed the largest difference for the WSP task (21.5 Hz versus 5 and 1 Hz for Dutch and Romanians, respectively). For Dutch listeners, F0 is predicted to be the primary cue to prominence, followed by duration (Van Katwijk, 1974). For Italians, however, the prediction is the other way around (Bertinetto, 1980). As duration was normalized in the tasks, the absence of this cue might have affected the Italian listeners differently than the Dutch. Still, effects were small enough for standard training to eliminate between-language differences. As for the SI task, no differences between the language backgrounds were found, which is consistent with the prediction that F0 is the most important cue for clause typing in the three languages. The test battery presented here allows insight into how well F0 is processed in language-like contexts, and how this relates to the perception of F0 in tone complexes. In real speech, cues other than F0 may be present, and the amount to which these are used by listeners may differ between languages.

Test–retest reliability showed small, but generally non-significant, improvements from the first to the second moment of testing. This confirms the reliability of the tests. The small differences can be explained by a learning trend: the first test session may have familiarized participants with the procedure and the materials. Test–retest differences measured on the speech tasks were found to be somewhat larger than on non-speech tasks. The larger test–retest differences found for speech tasks as opposed to the non-speech ones may indicate that stimulus resolution was too high. The *largest* step size used in this test prototype was 1/3 ST (~4–5 Hz). When looking at the listener's task in actual spoken communication, relevant pitch movements lie in the range of 2–8 ST, ~24–117 Hz for this speaker (e.g. Gussenhoven and Rietveld, 1985; Campione and Véronis, 1998). This would mean that we are attempting to measure listeners' discrimination or identification of pitch differences that are meaningless from a linguistic perspective. We therefore take these results as ground for reducing the test resolution, which is explained in the next subsection.

Correlations between tests were all positive and most were significant. This is not surprising as all tests were designed to measure perception of pitch or pitch changes. Only few correlations, however, got over 0.5, that is medium correlation. These results therefore do not strongly suggest that tasks are interchangeable. As for the usability of the tests both participants' impressions and test durations were considered. The questionnaires showed that participants were fairly positive about the instructions and the tests. Test difficulty and duration were judged average, and the false alarm buzz seemed to be effective as it was perceived as somewhat startling, though listeners did not feel that it affected their performance. The judgments on the stimulus materials seemed to indicate that the listeners from the different language backgrounds perceived the pseudo-speech as language, but not as their native language. Average test durations with normal-hearing listeners were several minutes per task, which seems to be acceptable for transfer to a clinical context. For the DI and HI tests it has been shown that test durations in clinical populations are 2.5 minutes on average (Vaerenberg *et al.*, 2011).

Adjusting the resolution

Test–retest results suggested that although the tests were generally reliable, the absolute differences obtained for the speech tasks were large enough to question the current fine-grained measurement of JNDs. As differences of a few hertz between test and retest cannot be considered relevant in terms of speech perception, the step sizes used in the test battery were increased. Such

increases might reduce both within-listener variation and test durations.

A new step size (i.e. resolution) was proposed on the basis of the test–retest differences obtained. We assumed that the differences shown in Fig. 3 can be interpreted as listener-internal variation. This variation was therefore modeled by the standard deviation of the median test–retest difference, that is, IQR/1.35 (IQR stands for inter-quartile range). As a good step size would not measure such random variation, the new resolution for the speech tasks was determined by first computing new step sizes in hertz, and subsequently rounding them to the nearest fraction on a ST scale. This was done by adding 2 times IQR/1.35 to the median test–retest difference for non-speech tests, and 1 time IQR/1.35 for the speech ones. From these step sizes in hertz and a pre-determined default step size of 1/12 ST, the new minimum resolution was set to 1/6 ST for the non-speech and 1/2 ST for the speech stimuli (Fig. 4).

Testing hearing-impaired listeners

The healthy cochlea provides information about complex acoustic signals by means of spectral and temporal coding. Temporal coding, and specifically the temporal fine structure (TFS), is thought to contribute to speech pitch perception (e.g. Xu and Pfingst, 2003; Moore, 2008), and spectrally, harmonics contribute to pitch perception. In our low-pass-filtered tasks, only the first harmonic (F0) and low-frequency TFS are available to listeners. Assuming that in complex tones the lower harmonics above the fundamental are actually more important for pitch perception than the fundamental frequency itself, i.e. residue pitch (e.g. Stagray *et al.*, 1992), the test battery proposed here seems to support the assessment of the availability of low-frequency TFS information for speech pitch perception rather than spectral pitch.

Following the literature, there are at least two populations that experience difficulties with the perception of TFS cues: menièriform listeners, and hearing-impaired individuals with sensorineural hearing loss. Menière's disease is a disorder of the cochlea that affects balance and hearing. It is characterized by a hearing loss which is primarily located in the lower-frequency region (125–1000 Hz). It has been claimed that Menière's disease is associated with abnormal firing in the auditory nerve and that this results in a decreased ability to use TFS cues (Chung *et al.*, 2004). As such, Menière-patients are expected to show decreased performance on tasks targeting these cues.

Hearing-impaired individuals with a sensorineural loss may represent another population of listeners unable to infer pitch from TFS cues (Buss *et al.*, 2004; Moore, 2008). More particularly, it has been hypothesized that poor speech intelligibility in listeners with sensorineural hearing loss may be because of their reduced ability to use TFS information. Lorenzi *et al.* (2006) measured identification scores for unprocessed and TFS speech in normal hearing and hearing-impaired listeners, and found that, whereas normal-hearing listeners obtain good scores with both types of speech, hearing-impaired listeners performed well with unprocessed speech, but performed very poorly with speech containing only TFS cues. Results of hearing impaired listeners on the HI and DI tasks from our test battery showed that listeners with low-frequency loss and CI users, but not hearing impaired listeners with high-frequency loss, had significantly higher JNDs on both tests than the norm data reported here, and higher JNDs on the disharmonic than the harmonic task (Vaerenberg *et al.*, 2011).

The questions arise how this relates to current rehabilitation strategies for hearing-impaired individuals and how the newly developed test battery for prosodic perception can contribute. Nowadays, advances in

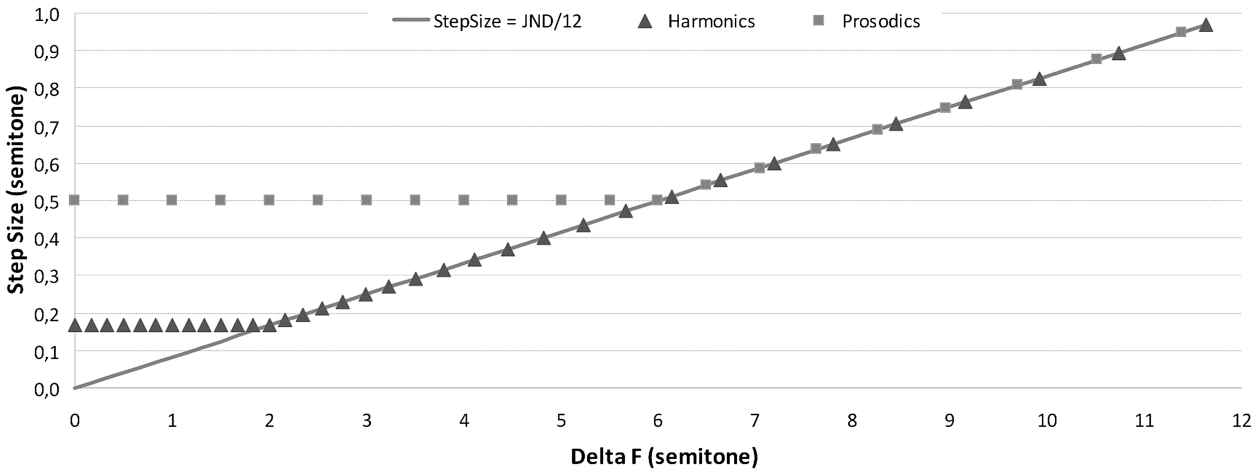

Figure 4 Model for new step sizes for the harmonic, i.e. non-linguistic, test (triangles), and the prosodic, i.e. linguistic, tests (squares).

hearing devices enable intervention by means of a conventional hearing aid or a CI, depending among others, on the type, degree and configuration of the patient's hearing loss. The development of a CI is based on the idea that in most deaf patients, in spite of the damaged cochlea, enough auditory nerve fibers are left for direct stimulation. Unable to code TFS, CIs provide very restricted information about pitch. For CI users, acoustic stimulation of residual low-frequency hearing is expected to provide the TFS cues that are necessary for pitch perception while at the same time electric stimulation of high-frequency sounds conveys spectral information that is not encoded in classical hearing aids (Gantz and Turner, 2003).

As argued above, both the DI task and the filtered speech tasks may be taken to represent listening conditions for which the thresholds are largely dependent on the availability of TFS cues. It is expected that only listeners who are able to make use of these cues will obtain low JNDs. As a consequence, relatively high JNDs on the DI and filtered SI and WSP tasks as compared to lower JNDs on the other tasks of the test battery (HC, HI, and unfiltered SI/WSP) can demonstrate the listener's inability to infer pitch information from TFS cues.

Although the optimal fitting of hybrid EAS devices is still under investigation, it is generally accepted that the restoration of TFS information in the low-frequency region thanks to EAS will have beneficial effects (Gfeller *et al*., 2006). For instance, the combination of a CI with a 10 mm electrode array (instead of 20–30 mm arrays) and a hearing aid showed better performance on speech perception in noise and melody recognition than a traditional CI (Gantz *et al*., 2005). Perception of speech, in quiet and in noise, was found to be generally better with EAS than with electric or acoustic stimulation alone (Dorman *et al*., 2008). The latter study furthermore showed that melody recognition was better with EAS than with electrical stimulation alone, whereas voice discrimination did not differ between conditions.

The results of a pilot study using the test battery presented here show that performance on tasks that only provide pitch cues below 300 Hz (DI and filtered SI) is worse in subjects using a CI processor with electrical stimulation alone than in those wearing a CI with EAS processor. These results extend the results on the non-speech tasks obtained by Vaerenberg *et al*. (2011). Crucially, the pilot showed a median improvement in JND of 24 Hz in six CI users for DI when retested under an EAS condition as opposed to electrical stimulation alone. As for the filtered SI task, two out of six listeners showed JNDs within the normal range (45 and 24 Hz) when retested in the EAS condition. In Schauwers *et al*. (in preparation) the intonation perception skills of different hearing-impaired populations in the speech tasks are presented, comparing outcomes of hearing aid users with those of CI users with electrical stimulation alone and with EAS (speech processor, Neurelec France).

In conclusion, the design and validation of a test battery aiming to assess speech pitch perception were presented. Its main contributions are three-fold. First, tasks from the test battery may be used to assess the perception of pitch in speech-like stimuli instead of tone complexes that may be considered less representative of communicatively realistic situations. Second, the stimulus materials vary in pitch only, and do not contain co-varying, secondary cues as opposed to other prosodic tests. Third, the new tests can be used with listeners from a number of different language backgrounds, making them more widely applicable than existing ones.

The validation suggested that non-speech tests using tone complexes may overestimate listener performance when it comes to pitch perception in speech. This justifies the use of linguistically based tests for the assessment of perception of pitch in speech. We have furthermore established normative data from normal-hearing listeners, and shown that these listeners, despite different language backgrounds, score comparably on most tasks. The relatively short task durations and the questionnaire results seem to make the tests suitable for use in clinical practice. For part of the tasks it has been shown that they aid in the diagnosis of impairments in low-frequency perception, and the pilot suggests that combinations of particular tasks can be used to measure improvements in perception through new hearing rehabilitation strategies, such as EAS.

Acknowledgements

This research was supported by EU-FP7-SME-222291 Dual Pro *Dual electric-acoustic speech processor with linguistic assessment tools for deaf individuals with residual low-frequency hearing*. We would like to thank Agnes Légèr and Christian Lorenzi for the independent check of the stimulus materials' acoustic contents, Anne van der Kant for help in testing listeners, Vincent van Heuven and Johan Rooryck for valuable discussion, and two anonymous reviewers for helpful comments on an earlier version of this manuscript. Vincent Péan of Neurelec France provided the scripts for generating low-pass filtered stimuli.

Appendix 1

Word and sentence forms are given in Table 3.

Appendix 2

Phone and phoneme durations are given in milliseconds. In the case of initial syllables, phone 1 is silence, and in the case of final syllables, phone 3 is silence (see Table 4).

Table 3 Word and sentence forms

Four-syllable sentences	Five-syllable sentences	Six-syllable sentences	Three-syllable pseudowords	
ma-nu-ma-ni	mu-ni-ma-na-nu	mi-ni-mu-ma-nu-na	ma-mi-nu	na-mu-mi
mi-nu-ni-ma	ma-mu-ni-na-mu	mu-ma-na-ni-mu-mi	ma-ni-mu	ma-nu-ni
nu-ma-na-mi	ni-mu-ma-nu-na	na-nu-ni-mu-na-ma	mi-na-mu	ni-mu-na
na-mi-ma-nu	nu-na-mu-na-mi	ni-nu-mu-mi-na-nu	mu-na-ni	nu-ma-ni
			mu-ni-ma	nu-mi-ma

Table 4 Phone and phoneme durations are given in milliseconds

	Original duration				Normalized duration				
	Phoneme 1	Phoneme 2	Phoneme 3	Total	Phoneme 1	Phoneme 2	Phoneme 3	Total	Final
#m-	166	57	–	223	50	59	–	109	108
#n-	196	61	–	257	50	59	–	109	109
-mam-	50	189	50	289	53	151	57	261	265
-mim-	46	127	57	230	53	151	57	261	263
-mum-	51	146	61	258	53	151	57	261	258
-man-	48	215	38	301	53	151	45	249	250
-min-	62	108	42	212	53	151	45	249	249
-mun-	59	129	47	235	53	151	45	249	251
-nan-	44	200	51	295	49	151	45	245	248
-nin-	56	108	54	218	49	151	45	245	244
-nun-	42	153	40	235	49	151	45	245	246
-nam-	53	186	62	301	49	151	57	257	258
-nim-	48	115	70	233	49	151	57	257	254
-num-	53	135	41	229	49	151	57	257	256
-ma#	71	435	92	598	53	151	50	254	270
-mi#	76	391	115	582	53	151	50	254	270
-mu#	76	334	70	480	53	151	50	254	270
-na#	63	364	78	505	49	151	50	250	270
-ni#	72	364	64	500	49	151	50	250	270
-nu#	77	393	80	550	49	151	50	250	270

Table 5 Median and quartile questionnaire responses

	NL			IT			RO		
	p25	p50	p75	p25	p50	p75	p25	p50	p75
Instructions									
Training was helpful	4	5	5	4	4	5	4	5	5
Task not understood	1	1	1	1	1	2	1	1	1.5
Training was confusing	1	1	1	1	1.5	2	1	1	1
Instructions were clear	4	5	5	4	5	5	5	5	5
Test experience									
Test was fun to do	4	4	4.25	3	4	4	3	4	5
Test was too difficult	2	2	2.25	2	2	3	2	2	3
Test was easy	2	3	3	2.5	4	4	2	3	4
Test was too long	2	2	3	2	2	4	2	2	4
I felt insecure	1	2	3	2	2	2	2	3	4
I guessed a lot	1.75	2	2	1	2	2	1	1	2
It was difficult to remain concentrated	2	2	3.25	2	4	4	2	2	2
I used a listening strategy	1.75	2	4	2	2	2	1	1.5	3
False alarms									
Alarm sound was startling	2.5	4	4	2	2	4	2.5	4	4
Alarm sound improved performance	3	3	4	2	2	3	2	3	3.5
Alarm sound made me insecure	2	2	4	2	2	3.5	1.5	2	3.5
Stimulus materials									
Stimuli were like words/sentences	2.75	4	4	2	2	4	1	3	4
Stimuli sounded natural	2	3	4	1.5	2	2	1.5	2	3
Stimuli sounded native	1.75	2	4	1	2	2	1	2	2

Appendix 3

Median and quartile questionnaire responses (1 = not agree; 5 = fully agree) per language background are given in Table 5.

References

't Hart J., Collier R., Cohen A. 1990. *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press.

Alfano I. 2006. La percezione dell'accento lessicale: un test sull'italiano a confronto con lo spagnolo. Proceedings of 2nd AISV. p. 632–656.

Assmann P.F. 1999. Fundamental frequency and the intelligibility of competing voices. Proceedings of the International Congress of Phonetic Sciences. p. 179–182.

Avram A. 1970. Sur le role de la fréquence dans la perception de l'accent en roumain. Proceedings of the Sixth International Congress of Phonetic Sciences Prague 1967. p. 137–139.

Bachorowski J.A., Owren M. 1999. Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *Journal of the Acoustical Society of America*, 102: 1054–1063.

Barry J.G., Blamey P.J., Martin L.F.A., Lee K.Y.S., Tang T., Ming Y.Y., Van Hasselt C.A. 2002. Tone discrimination in Cantonese-speaking children using a cochlear implant. *Clinical Linguistics & Phonetics*, 16: 79–99.

Bertinetto P.M. 1980. The perception of stress by Italian speakers. *Journal of Phonetics*, 8: 385–395.

Boersma P., Weenink D. 2008. Praat: doing phonetics by computer, http://www.praat.org.

Brokx J.P.L., Nooteboom S.G. 1982. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, 10: 23–36.

Buss E., Hall III J.W., Grose J.H. 2004. Temporal fine-structure cues to speech and pure tone modulation in observers with sensorineural hearing loss. *Ear and Hearing*, 25(3): 242–250.

Campione E., Véronis J. 1998. A statistical study of pitch target points in five languages. Proceedings of ICSLP, 1391–1394.

Carrell T.D., Smith L.B., Pisoni D. 1981. Some perceptual dependencies in speeded classification of vowel color and pitch. *Perception and Psychophysics*, 29: 1–10.

Chatterjee M., Peng S. 2008. Processing F0 with cochlear implants: modulation frequency discrimination and speech intonation recognition. *Hearing Research*, 235: 143–156.

Chung B.J., Hall III J.W., Buss E., Grose J.H., Pillsbury H.C. 2004. Menière's disease: effects of glycerol on tasks involving temporal processing. *Audiology and Neurotology*, 9: 115–124.

Ciocca V., Francis A.L., Aisha R., Wong L. 2002. The perception of Cantonese lexical tones by early-deafened cochlear implantees. *Journal of the Acoustical Society of America*, 111: 2250–2256.

Cutler A. 2007. Lexical stress. In: Pisoni D., Remez R. (eds), *The Handbook of Speech Perception*. Blackwell Publishing. p. 264–289.

Dorman M.F., Gifford R.H., Spahr A.J., McKarns S.A. 2008. The benefits of combining acoustic and electric stimulation for the recognition of speech, voice and melodies. *Audiology and Neurotology*, 13: 105–112.

Fry D.B. 1958. Experiments in the perception of stress. *Language and Speech*, 1: 126–152.

Gantz B.J., Turner C.W. 2003. Combining acoustic and electrical hearing. *The Laryngoscope*, 113: 1726–1730.

Gantz B.J., Turner C.W., Gfeller K.E., Lowder M.W. 2005. Preservation of hearing in cochlear implant surgery: advantages of combined electrical and acoustical speech processing. *The Laryngoscope*, 115: 796–802.

Gfeller K., Olszewski C., Turner C., Gantz B., Oleson J. 2006. Music perception with cochlear implants and residual hearing. *Audiology and Neurotology*, 11: 12–15.

Gfeller K., Turner C., Mehr M., Woodworth G., Fearn R., Knutson J., Witt S., Stordahl J. 2002. Recognition of familiar melodies by adult cochlear implant recipients and normal-hearing adults. *Cochlear Implants International*, 3: 29–53.

Govaerts P.J., Daemers K., Yperman M., De Beukelaer C., De Saegher G., De Ceulaer G. 2006. Auditory speech sounds evaluation (A§E®): a new test to assess detection, discrimination and identification in hearing impairment. *Cochlear Implants International*, 7: 97–106.

Green D.M. 1976. *An Introduction to Hearing*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Green T., Faulkner A., Rosen S. 2004. Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants. *Journal of the Acoustical Society of America*, 116: 2298–2310.

Green T., Faulkner A., Rosen S., Macharey O. 2005. Enhancement of temporal periodicity cues in cochlear implants: effects on prosodic perception and vowel identification. *Journal of the Acoustical Society of America*, 118(1): 375–385.

Gussenhoven C., Rietveld A.C.M. 1985. On the relation between pitch excursion size and pitch prominence. *Journal of Phonetics*, 13: 299–308.

Jusczyk P.W. 1997. *The Discovery of Spoken Language*. Cambridge, MA: The MIT Press.

Kalikow D.N., Stevens K.N., Elliott L.L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61: 1337–1351.

Klatt D.H. 1973. Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *Journal of the Acoustical Society of America*, 53(1): 8–16.

Kong Y.-Y., Cruz R., Ackland Jones J., Zeng F.G. 2004. Music perception with temporal cues in acoustic and electrical hearing. *Ear and Hearing*, 25: 173–185.

Kuo Y.-C., Rosen S., Faulkner A. 2008. Acoustic cues to tonal contrasts in Mandarin: implications for cochlear implants. *Journal of the Acoustical Society of America*, 123: 2815–2824.

Laneau J., Wouters J., Moonen M. 2006. Improved music perception with explicit pitch coding in cochlear implants. *Audiology & Neurotology*, 11: 38–52.

Levitt H. 1971. Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, 49: 467–477.

Lorenzi C., Gilbert G., Carn H., Garnier S., Moore B.J.C. 2006. Speech perception problems of the hearing impaires reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences*, 103: 18866–18869.

Meister H., Tepeli D., Wagner P., Hess W., Walger M., von Wedel H., Lang-Roth R. 2007. Experimente zur Perzeption prosodischer Merkmale mit Kochleaimplantaten. *HNO*, 55: 264–270.

Moore B.J.C. 2008. The role of temporal fine structure processing in pitch perception, masking and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology*, 9: 399–406.

Owens E., Kessler D.K., Schubert E.D. 1981. The minimal auditory capabilities (MAC) battery. *Hearing Aid Journal*, 34: 9–34.

Peppé S., Martinez-Castilla P., Coene M., Hesling I., Moen I., Gibbon F. 2010. Assessing prosodic skills in five European languages: cross-linguistic differences in typical and atypical populations. *International Journal of Speech-Language Pathology*, 12: 1–7.

Peppé S., McCann J. 2003. Assessing intonation and prosody in children with atypical language development: the PEPS-C test and the revised version. *Clinical Linguistics & Phonetics*, 17: 345–354.

Pierrehumbert J.B. 1980. *The Phonology and Phonetics of English Intonation*. Massachusetts: Institute of Technology.

Pisoni D. 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, 13: 253–260.

Plomp R., Mimpen A.M. 1979. Speech-reception threshold for sentence as a function of age an noise level. *Journal of the Acoustical Society of America*, 66: 1333–1342.

Repp B.H., Lin H.B. 1990. Integration of segmental and tonal information in speech perception: a cross-linguistic study (A). *Journal of the Acoustical Society of America*, 87: S46.

Savino M. 2004. Intonational cues to discourse structure in a regional variety of Italian. In: Gilles P., Peters J. (eds). *Regional Variation in Intonation*. Tübingen: Niemeyer. p. 145–160.

Schauwers K., Coene M., Heeren W., del Bo L., Pascu A., Vaerenberg B., Govaerts P.J. in preparation. Pitch Perception in Hearing-Impaired Adults with Aided and Unaided Hearing Loss.

Spitzer S., Liss J., Spahr T., Dorman M., Lansford K. 2009. The use of fundamental frequency for lexical segmentation in listeners

with cochlear implants. *Journal of the Acoustical Society of America*, 125: EL236–EL241.

Stagray J.R., Downs D., Sommers R.K. 1992. Contributions of the fundamental, resolved harmonics, and unresolved harmonics in tone-phoneme identification. *Journal of Speech and Hearing Research*, 35: 1406–1409.

Sucher C.M., McDermott H.J. 2007. Pitch ranking of complex tones by normally hearing subjects and cochlear implant users. *Hearing Research*, 230: 80–87.

Swerts M., Collier R., Terken J. 1994. Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication*, 15: 79–90.

Vaerenberg B., Pascu A., Del Bo M., Schauwers K., De Ceulaer G., Daemers K., Coene M., Govaerts P. 2011. Clinical assessment of pitch perception. *Otology & Neurotology*, 32(5): 736–741.

Van Heuven V.J., Haan J. 2000. Phonetic correlates of statement versus question intonation in Dutch. In: Botinis A. (ed). *Intonation: Analysis, Modelling and Technology*. Dordrecht/Boston/London: Kluwer. p. 119–144.

Van Katwijk A. 1974. *Accentuation in Dutch*. Amsterdam/Assen: Van Gorcum.

Vroomen J., Collier R., Mozziconacci S. 1993. Duration and intonation in emotional speech. Proceedings of Eurospeech. p. 577–580.

Xu L., Pfingst B. 2003. Relative importance of temporal envelope and fine structure in lexical-tone perception. *Journal of the Acoustical Society of America*, 114: 3024–3027.