MAKING SENSE OF STRANGE SOUNDS: (MUTUAL) INTELLIGIBILITY OF RELATED LANGUAGE VARIETIES. A REVIEW

VINCENT J. VAN HEUVEN

1. INTRODUCTION

1.1 Two basic questions

In this paper we ask two questions, which superficially seem to ask the same thing but in actual fact do not. First, we ask to what degree two languages (or language varieties) A and B resemble each other. The second question is how well a listener of variety B understands a speaker of variety A.

When we ask to what degree two language varieties resemble one another, or how different they are (which is basically the same question), it should be clear that the answer cannot be expressed in a single number. Languages differ from each other not in just one dimension but in a great many respects. They may differ in their sound inventories, in the details of the sounds in the inventory, in their stress, tone and intonation systems, in their vocabularies, and in the way they build words from morphemes and sentences from words. Last, but not least, they may differ in the meanings they attach to the forms in the language, in so far as the forms in two languages may be related to each other. In order to express the distance between two languages, we need a weighted average of the component distances along each of the dimensions identified (and probably many more). So, linguistic distance is a multidimensional phenomenon and we have no a priori way of weighing the dimensions.[1]

The answer to the question how well listener B understands speaker A can be expressed as a single number. If listener B does not understand speaker A at all, the number would be zero. If listener B gets every detail of speaker A's intentions, the score would be maximal.

The primary goal of human language is to communicate intentions from speaker to listener. When listener B does not know the structural details of

International Journal of Humanities and Arts Computing 2 (1–2) 2008, 39–62
DOI: 10.3366/E1753854809000305

speaker A's language, communication will be less than optimal, and if the difference between speaker A's and listener B's linguistic codes is larger than some critical amount, communication will fail altogether. Intelligibility between languages may serve as the ultimate criterion to decide how structural dimensions should be weighed against each other in the computation of linguistic distance. Suppose, for instance, that differences in word order hardly compromise the communication between A and B, but that even small discrepancies between sound systems cause a complete communication breakdown. Then, phonology should be weighted much more in the computation of linguistic distance than syntax.

So, the two basic questions have a mutual feeding relationship. On the one hand, we would like to be able to predict from differences and similarities between two languages A and B how well listener B will understand speaker A. Here we need a detailed survey of structural similarities and differences along all of the dimensions along which languages may differ, and we need to know how to weigh the dimensions against each other in order to make the prediction. On the other hand, we need to know how well listener B understands speaker A. The intelligibility score is the only criterion against which the relative importance of linguistic dimensions can be gauged. It is the only reasonable criterion if we subscribe to the communicative principle underlying linguistic structure. In this article our initial focus will be on the second question because we want to use intelligibility as a criterion for weighing the different dimensions of linguistic distance. In later sections we will also consider factors that influence intelligibility.

1.2 Defining the problem

The problem that we wish to address is the following. Given two related language varieties A and B, where A and B share a substantial part of their lexicon and linguistic structure, by what mechanism is listener B able to understand speaker A? That is to say, we are interested in the *psycholinguistic mechanism* that enables communication between speakers and listeners of related language varieties – such as dialects of a language of related languages within a language family.

A human language processor, i.e. a listener, may have at his disposal adaptive strategies to cope with deviant speech input. For instance, a Dutch listener B when confronted with English input A may realise that the sound shape /hɑʊs/ refers to the same concept 'house' as the obviously cognate Dutch form /hœys/. Once the Dutch listener has discovered this relationship he may apply this sound transformation to other English forms, such as /lɑʊs, mɑʊs, lɑʊd, snɑʊt/ 'louse, mouse, loud, snout', which all transform regularly to Dutch /lœys, mœys, lœyt, snœyt/. In this case, the listener has discovered a rule that relates the English sound shapes to their equivalents in Dutch. The transformation is not as simple as it seems, however. The phonology of Dutch distinguishes between two rounded low diphthongs /ɑʊ and /œy/, where English has only one. Not only does the Dutch listener of English have to learn that Dutch /œy/ maps onto a different vowel in English, viz. /γʊ/ but also that Dutch /ɑʊ/ as in /ɣɑʊt/ 'gold' or /hɑʊt/ 'hold' maps onto the English sound combination /əʊl/, and so on. To keep the problem within manageable bounds, therefore, I will exclude such learning strategies from our problem. I will assume that the listener's linguistic knowledge is static and that no rules are being developed to cope with the deviant input speech. In other words, I explicitly limit the problem of understanding deviant speech to first confrontation, assuming that listener B has no previous experience with the kind of aberrations that are characteristic of language A.

To simplify matters further, I will assume a laboratory setting for the testing of intelligibility of deviant speech. The input speech will be sound only, presented out of context. No visual or situational cues will be present in the stimulus.

Languages may also differ in their syntax. Differences in word order may determine the meaning of sentences. If the default word order in the language is Subject-Verb-Object (SVO) then the sentence *X kills Y* implies that Y dies. Such sentences will be incorrectly understood by listeners of an Object-Verb-Subject (OVS) language: they will believe instead that X dies as a result of the killing action performed by the subject Y. It appears that such gross typological differences are rare within groups of closely related languages. Again, to simplify the problem, therefore, I will assume that there are no differences in word order between the language of speaker A and listener B. Or rather, whatever differences in word order may exist between the two languages, they do not compromise intelligibility.

1.3 Approaching the problem

When we listen to someone who speaks in a related language that we have not heard before in our life, speech understanding is compromised to a greater or lesser extent. Situations in which speech input is non-optimal abound in everyday life. The speaker may be handicapped by some language or speech pathology (e.g. stuttering, cleft palate speech, alaryngeal speech, e.g. after surgical removal of the larynx and vocal cords). Special kinds of pathologies are accent, whether foreign or native, and computer speech.[2] Alternatively, the speaker may be perfectly normal but the communication channel may be polluted by noise (ambient noise, competing speech input, harmonic distortion, echoes and reverberation, selective amplification and filtering), which may be continuous or intermittent (perceptual adaptation to intermittent noise is harder).

The amazing fact is that the native listener is generally quite successful in getting the speaker's intentions even if the input speech is highly defective and even if the communication channel is noisy. Human spoken language has evolved such that it is extremely robust and works under the most adverse circumstances. The science of phonetics, more than any other branch of knowledge, studies the process of speech perception. A full-fledged theory of human speech perception should allow us to understand the robustness of speech communication and predict how the listener would reconstruct the speaker's message even if the input speech is defective or when the communication channel is noisy. I refer to relevant chapters in the *Handbook of Speech Perception* (Remez and Pisoni, 2005) for sketches of such theories.

I therefore embrace the null hypothesis that understanding a speaker of a related variety of one's native language does not involve any special mechanisms. Rather, the listener simply marshals up the mechanisms that he routinely brings to bear in the processing of speech input under suboptimal listening conditions. I suggest, in other words, that insights into the normal speech perception mechanism should be sufficient to provide answers to our basic question: how well would a listener B understand speaker A if A and B are related but non-identical languages or language varieties. Note that within the science of phonetics I include, somewhat imperialistically, two specialisations that address the perception of defective input speech. These are (i) the phonetics of foreign language learning and (ii) speech technology, specifically the quality assessment of speech synthesis. There is a large body of research on both (i) and (ii) that we may fruitfully turn to for ideas on speech intelligibility, perceptual assimilation of strange sounds, and word recognition.

2. SPEECH INTELLIGIBILITY

In this section I will argue that the most important and central aspect of speech understanding is the recognition of the words, i.e. of the smallest units of language in which a more or less fixed meaning is coupled with a more or less fixed sequence of sounds. We will then briefly review techniques that have been developed to measure speech intelligibility and express intelligibility scores in terms of the percentage of words correctly recognised.

2.1 Word recognition is key

We will define intelligibility in quite practical terms as the percentage of linguistic units correctly recognised by the listener in the order in which they were spoken. Intelligibility can be tested at several levels of the linguistic hierarchy, be it at the level of meaningless units (sounds or phonemes), or at the level of meaningful units such as morphemes and words.

It has become standard practice in speech intelligibility measurement to test the recognition of linguistic units at several linguistic levels. Typically, intelligibility tests are part of a test battery that addresses sounds, words and sentences separately (see van Bezooijen and van Heuven, 1997 and references given there). When we want to apply speech intelligibility tests to the problem of establishing the success of communication between speaker and hearer of related language varieties, we are not so much interested in the success with which the listener identifies individual sounds. Rather, we are interested in the percentage of words that the listener gets right. Therefore, measuring the success of phoneme identification is only useful in so far as this measure helps us to predict the success of word recognition. The underlying assumption here is that word recognition is the key to speech understanding. As long as the listener correctly recognises words, he will be able to piece the speaker's message together.

2.2 Functional testing versus opinion testing

In the literature on quality assessment of speech synthesis a division is often made between functional intelligibility testing and opinion testing (e.g. van Bezooijen and van Heuven, 1997). Functional intelligibility tests measure the real thing. They measure to what extent a listener actually recognises linguistic units (words) in spoken stimuli. A traditional functional test is dictation; here listeners simply write down what (they believe) the speaker said. Dictation draws heavily on the listeners' memory. In intelligibility testing it is not realistic to repeat the spoken utterance, since speakers in a real-life situation normally say things only once. In order to reduce memory load, sentences can be exploded, i.e. read in short phrase-like chunks with pauses in between to write down the response, or parts of the message may be printed on the listener's answer sheet such that he has to recognise selected (blanked-out) words only. Typically, the score of a functional intelligibility test is a percentage that expresses what proportion of the linguistic units present in the stimulus materials were correctly recognized by the listener.

When listeners have recognised a word, that word will remain active in the listeners' memory for a long time (up to several hours or even a whole day, see e.g. Morton, 1969). The next time the listeners hear the same word, they will recognise it with very little effort. This so-called priming phenomenon results in ceiling effects. This is a problem if the same word has to be recognised by the same listener in different versions, for instance when spoken in the listener's native language B and in a related language A. In order to avoid priming effects, word recognition experiments block the different versions of stimulus words over different listeners so that each listener hears only one version of each stimulus word. Blocking of versions over listeners is not a problem when the number of versions is limited. In some studies on intelligibility of related language varieties, however, the number of versions is as much as fifteen (e.g. Gooskens and Heeringa, 2004 for Norwegian dialects, or Tang and van Heuven 2007, 2009 for Chinese dialects). In such more complicated experiments, blocking is done through Latin square designs in which each listener hears one-fifteenth part of the stimulus material in each of 15 different varieties, and yet hears materials in each of the 15 varieties in equal proportions, and never hears the same word twice (not even in a different variety). The blocking of versions of groups of listeners makes functional intelligibility testing a laborious undertaking. For this reason functional intelligibility testing is shunned when the number of language varieties under study is large.

It was discovered (or at least claimed) in work on quality assessment of talking machines that so-called opinion testing is an adequate shortcut to functional intelligibility testing. In opinion testing, listeners are asked how well *they think* they would understand a speech sample presented to them. The same sample can be presented to the same listener in several different versions, for instance synthesised by several competing brands of reading machines and by a human control speaker (Pisoni, Greene and Nusbaum, 1985; van Bezooijen and van Heuven, 1997). The listener is familiarised with the contents of the speech sample before it is presented so that recognition does not play a role in the process. All the listener has to do is to imagine that he has not heard the sample before and to estimate how much of its contents he thinks he would grasp. The response is an intelligibility judgment, between 0 'I think I would not get a single word of what this speaker says' and 10 for 'I would understand this speaker perfectly, I would not miss a single word.' It has been shown that the mean score averaged over a group of listeners/judges (so-called Mean Opinion Score or MOS) very strongly differentiates between speech of differing quality (high concurrent validity with functional intelligibility scores).

Tang and van Heuven (2009) computed the correlation between functional and opinion tests of intelligibility among 15 Chinese dialects. They found correlation coefficients around $r = .8$. This is a high degree of correlation but it also shows that opinion test do not account for all the variability in the functional test scores: some 35 per cent of the variance in the functional test scores goes unaccounted for. On the basis of this finding it seems advisable to attempt functional testing if at all possible; only if the number of language pairs targeted is very large, is opinion testing an option as a non-ideal but manageable alternative. And even in such large-scale comparisons it would always be advisable to cross-validate the opinion test results with functional counterparts for a subset of language pairs sampled from the larger ensemble.

2.3 Avoiding ceiling effects

When the language varieties of the speakers only differ in very subtle ways from that of the listener, it may be difficult to differentiate between close and not so very close varieties. In order to avoid such ceiling effects it may be useful to make the listener's task more difficult. What is generally done in such situations is that information in the stimulus is reduced by some form of signal degradation. There are many ways to degrade the input speech. It can be achieved by filtering (removing amplitude from the signal in specific frequency bands), by signal compression, by adding various kinds of noise to the signal or by replacing selected fragments of the signal by silence (or noise).

Filtering the speech signal is done when we listen to someone over an ordinary telephone. Here frequencies below 300 Hz and above 3300 Hz are removed from the signal. Normally, communication between native speakers and native listeners remains perfectly feasible with this impoverished kind of signal. When either the speaker or the listener is non-native, communication tends to become problematic.

Signal compression such as Linear Predictive Coding (LPC) is the basis of GMS telephony. It reduces input speech to a relatively small set of numbers that describe successive speech samples of, say, 10 ms. At the receiver end the speech is regenerated but with considerable loss of quality. The severity of the data reduction can be varied in small steps, which makes this a very useful research tool in intelligibility studies.

Adding noise to the communication channel is an effective method of making perfectly intelligible speech difficult to understand. Many types of noise have been tested for their effectiveness as a masker of speech. White noise affects all frequencies from low to high indiscriminately. This makes it a relatively ineffective masker, since speech has its energy concentrated at low frequencies. A more effective masker would be pink noise (which emphasises low frequencies) but the most effective way to mask speech is by adding more speech to it, i.e. competing voices. Lately, so-called speech noise or babble noise has become a very popular masker. This is basically speech recorded from many speakers added together. The masking noise can have a fixed intensity, for instance equal to the mean peak intensity of the vowels in the utterance. Alternatively, the noise may be intensity modulated such that when the intensity of the speech signal goes up by a particular number of decibels, so does the intensity of the masking noise. Communication between native speakers and native listeners withstands a lot of masking noise. The masking noise may be up to 12 dB stronger than the speech signal and the listener may still get the gist of the message. When either the speaker or the listener is non-native, however, communication fails at less extreme signal-to-noise ratios (van Wijngaarden, 2001).

In the preceding paragraphs we have considered the measurement of the dependent variable in intelligibility research, i.e. the quality of word recognition in sentential context. In the following sections we will address the question how the quality of word recognition can be predicted from a comparison of closely related languages at the level of smaller units, such as phonemes and allophones.

3. PERCEPTUAL ASSIMILATION OF STRANGE SOUNDS

3.1 Ask the listener

The way we perceive sounds is shaped by our linguistic experience. Native listeners of English sort incoming speech sounds into categories that are specific to English; Chinese listeners have learnt from childhood onwards to sort sounds in terms of the categories that are most appropriate for Chinese. At the centres of these native language categories prototypes are set up, which act like magnets. Tokens of speech sounds that differ from the prototype are perceptually drawn closer to it (the nearer they physically are to the prototype, the stronger the magnet effect), so that the listener is never aware of the (small) mismatch between the token and the prototype (Kuhl and Iverson, 1995). At the boundary between adjacent categories in perceptual space, however, even small differences can be adequately heard, so that sound discrimination at category boundaries is sharper than within categories.

When we hear sounds spoken in a language variety that differs from our native language, the incoming sounds will deviate to a lesser or greater extent from the prototypes we are used to. Nevertheless we categorise the large majority of the incoming sounds to the prototypes that we have learnt. Only when the discrepancy between an incoming sound and any existing prototype is very large, will the listener refuse to categorise the incoming sound. Best, McRoberts and Goodell (2001) have set up a typology of what they call assimilation patterns that may be observed when a listener is first confronted with sounds that deviate from the prototypes in the native language. Basically a non-native phone may be assimilated to a native category in one of three ways:

(i) *C (categorised)*: it may be categorized as an exemplar of a native phone. It may vary from a very good (prototypical) exemplar to a poor one (on a 1–7 goodness scale).

(ii) *U (uncategorised)*: the token may be at the boundary between two (or more) native categories such that the listener cannot decide which category to assimilate the token to. The sound falls within the native phonological space but in between existing categories.

(iii) *N (non-assimilable)*: the token is not assimilated into the native phonological space at all. It is heard as a non-speech sound instead. This may happen, for instance, when an English listener is first confronted with African click sounds. Here the listener often thinks the speaker clapped hands while speaking.

When studying the perception of sounds in a related language variety, category N will be extremely rare. It seems impossible, by today's standards, to predict whether a non-native sound will be categorised, and if so how, just by comparing sound recordings or physiological measurements of such tokens. Phonetic theory has just not come far enough. As a practical way out, the assimilation behaviour should be tested through experimentation, as in the example below.

In the field of second-language learning, the learner's native language is called the source language, and the language to be learnt is called the target language.

In a sound assimilation experiment, the learner is asked to categorise foreign (target) sounds in terms of his native (source) language with forced choice, and to rate each token for goodness (or 'typicality'), for instance on a scale from 0 (very poor token) to 10 (excellent token). Table 1 presents the results of such an assimilation experiment (Sun and van Heuven, 2007) in which Mandarin listeners were asked to identify the 19 vowels of British English in terms of 14 Mandarin (surface) vowel categories.

The results show, for instance, that any (half) open English monophthong is assimilated to Mandarin /ɑ/, although some are considered a very poor token (e.g. English /ɛ/ is rated as a token of Mandarin /ɑ/ at 2.8 on the 10-point goodness scale)

If we are to predict how well listener B will understand speaker A in a related language, we would first have to know how the sounds in speaker A's variety map onto the inventory of listener B, and how easy it would be for listener B to assimilate a particular sound to the category of his choice. Experiments such as the one exemplified here, would be a necessary first step. Such experiments have not been done in the context of predicting intelligibility of closely related languages.

3.2 Prediction of sound categorisation through learning algorithms

An interesting and promising development, and also an alternative to asking native listeners directly how they perceive strange sounds, is offered by learning algorithms. Suppose we have collected a large number of tokens, by many speakers, of all the sounds in the inventory of a language, for instance, all the monophthongs of English as spoken by American adults. One may then measure relevant acoustic properties of these vowel tokens, such as the first and second formant frequencies F1 and F2 – which would adequately represent vowel height and backness, respectively. The distances between the vowel tokens in the acoustic space can be scaled so as to be perceptually more realistic through Bark transformation (e.g. Traunmüller, 1990). Next, differences between speakers (and between sexes) can be substantially reduced through some simple normalisation procedure (most successful one is the Lobanov transformation, which is simply a z-normalisation within speakers, Lobanov, 1971). We may then submit these transformed and normalised data to an automatic classification procedure such as Linear Discriminant Analysis (LDA; for details on the above procedures, and references, see Wang and van Heuven, 2006). By comparing category membership and the (transformed) acoustic properties of the vowel tokens, the LDA will automatically set up category boundaries in the vowel space such that vowel tokens are optimally sorted (i.e. with the least number of classification errors) into the native categories.

Table 1. Mean percent identification and goodness rating (in parentheses) of English vowel stimuli in terms of Mandarin vowel categories. Boldfaced values indicate the modal identification response. Only identification scores > 10 per cent are included.

		Chinese Vowel Responses (percent identification and rating 1–10)													
		ɨ	i	*ɨ*	y	eɪ	æ	ə	u	ɤ	oʊ	ɔ	ɑ	ɑʊ	ɪə
English Vowel Stimuli	i:	23				**66 (5.9)**									
	ɪ					**73 (6.4)**		11							
	eɪ					**91 (7.4)**									
	ɛ						11			13			**33 (2.8)**		16
	æ:												**66 (5.4)**		16
	ə									**86 (6.9)**					
	u:								**91 (6.9)**						
	ʊ										**63 (7.3)**			13	
	oʊ										**86 (7.5)**				
	ɔ:												**88 (7.7)**	11	
	ɔ												**89 (7.1)**		
	ʌ												**70 (7.1)**	25	
	ɑ:												**64 (6.1)**	20	
	ɑɪ					13	11						**52 (3.4)**		
	ɑʊ												11	**81 (7.1)**	
	ɔɪ								28			**53 (6.6)**			
	ɪə	11				**61 (5.5)**									14
	ɛə					**48 (4.8)**									17
	ʊə								31			**61 (7.4)**			

Once the LDA is trained on a given set of native speech sounds, we may apply the same set of decision rules to a new dataset, which may be another set of vowel tokens produced by native speakers of the same language as the training data (in which case performance will be very good to excellent). We may also use the decision rules to categorise a dataset with vowel tokens that deviate from the training data. This may be a set of vowel tokens produced by foreign-language learners but it may also be a set of vowels of a different (in this case related) language. The LDA will then tell us how a native listener of the target language would classify each input vowel token. In this way, the LDA is a model of the native listener of the target language. Such a model can be used to predict how listeners of source language B would assimilate the vowels of target language A to their native language categories (Strange et al. 1998, 2004). The same methodology should also work for the assimilation of consonant sounds, provided, of course, that the targeted acoustic dimensions are appropriate for consonant classification.

I have not yet seen this methodology applied to the problem of predicting the perception of a closely related language. Note, however, that although the method described here would probably yield the desired result, it is not driven by theoretical insight; it merely uses the mediating device of a empirically derived computational model. The method does not allow us to directly compare linguistic-phonetic descriptions of the vowel systems of the languages concerned and predict how listeners of one language would categorise the vowels of the other language.

Let us suppose that we now know how the sounds of language A are mapped onto the inventory of a closely related language B, so that we know which vowels and consonants in listener B's language are activated to what extent by the successive incoming sounds produced by the speaker of language A. How would listener B be able to recognise words in the defective input? This is what we will consider in the next section.

4. FROM SOUNDS TO WORD RECOGNITION

4.1 Model of human word recognition

We know from psychophysics that short-term memory keeps a faithful representation of the auditory input no longer than 250 ms (e.g. Crowder and Morton, 1969; Massaro, 1972). After a quarter of a second, the details of the auditory input have evaporated from memory. In a language such as English most words last longer than 250 ms. Therefore, a major problem in spoken word recognition is how the human listener is able to recognise a word even though the acoustic information that defines it is never available for inspection in its entirety. In this respect spoken word recognition presents a challenge that is absent in visual word recognition, where the reader may always refocus on earlier text input.

In order to account for spoken word recognition a range of models have been proposed. Here I will be eclectic and describe in quite general terms what a reasonable model of human word recognition might look like.

It is widely accepted that the human brain is a massive parallel processor. For every word we know, there is a specialised group of brain cells, also called 'word recognition unit' or 'logogen' (Morton, 1969) that has learnt to respond only to information that is characteristic of that particular word. If we know, say, 50,000 different words, then we have 50,000 logogens. When we listen to speech, the auditory information is fed to all 50,000 logogens in parallel. When the incoming sound matches the internal specification of the logogen, its activation is increased; when there is no match (or an outright clash between what is actually heard and what should have been heard), the logogen's activation remains stationary (or is reduced). The better the incoming sound matches the internal specification of the logogen, the greater its contribution to the overall activation of the logogen.[3]

However, incoming speech sounds do not activate words directly. At a lower level in the system there are recognition units for sound categories (phonemes or similar). The phone units are bi-directionally connected with the logogens. When the input acoustics activate the phoneme /k/, all words with a /k/ in their specification will increase their activity. When, for instance, the word *cat* is being said, any word with a /k/ in it is activated. As the logogen for *cat* is activated, so are (through back-propagation) all the phonemes that are internally specified in the logogen for *cat*, such as the /æ/ and the /t/. When the subsequent sound input indeed contains /æ/ and /t/, these phonemes will be active on two counts: by bottom-up activation through sensory input and by top-down activation through back-propagation. Phonemes in words that are not being activated by sensory input receive negative signals ('inhibition') from more successful candidates, so that very soon after the onset of a word only one feasible candidate remains, which is then recognised (winner takes all). Moreover, activation of a word leads to activation of all other words that are semantically (and syntactically) related to it. When a word is deactivated the activation of all related words is also reduced. When a word is actually recognised, it remains active for a long time, and so are all words that are neurally connected to it. This is how semantic and syntactic dependencies are accounted for. This view of human word recognition draws heavily on ideas behind the TRACE model (McClelland and Elman, 1986, see also additional references in note 3).

There are several more effects that have been found to affect word recognition. These effects will also play a role when listeners do not get input in their native language but when the input speech is distorted due to the fact that the speaker has a foreign accent or speaks a closely related language.

4.2 *Frequency effects*

Words that we have heard often before tend to be recognised sooner (from less sensory input) than infrequent words. This frequency effect is accounted for by the fact that the activation of a word that was actually recognised, remains high for a long time, and never fully returns to its previous resting level. Highly frequent words, therefore, have acquired a permanent headstart in the recognition process. As a result, when the input speech in the related language is ambiguous or otherwise unclear, thereby activating multiple recognition candidates in the mind of the listener, high-frequency recognition candidates will be favoured.

These, and other, models of auditory word recognition neatly account for the phenomenon that a listener may recognise a long word without having to keep the entire sound shape of the word in auditory memory. Incoming sound is short-lived. All it does is activate phonemes to a greater or lesser degree, and then it dies. Acoustic information is thereby recoded into a more abstract neural representation with a longer life cycle.

4.3 *Superiority of the word beginning?*

Older models of word recognition (whether auditory or visual) attached special importance to the beginning of words. For instance, the Cohort model (Marslen-Wilson and Welsh, 1978; Nooteboom, 1981) claimed that a word could never be recognised if the sounds in the word onset (defined as the first 200 ms of the word) could not be heard. In later experiments, however, it was shown that the word onset is not indispensable, and that, in fact, auditory information in any part of the word contributes equally in principle to the recognition process (Nooteboom and van der Vlugt, 1988) – as is implied by the neural network view presented in Section 4.1.

Earlier sounds in a word enter the auditory system before the later sounds. It is advantageous for the word recognition system to reduce the number of competing candidates as rigorously as possible. Keeping many alternatives open requires extra processing capacity, which is a commodity. There are clear indications that the languages in the world tend to concentrate contrastive information in the beginning of words. For instance, in any language I know, the number of different sounds that may occur at the beginning of a word is larger than the inventory of sounds that may occur at the end of a word. The advantage of this organisational principle is that words can be recognised sooner (i.e. from a shorter onset portion) than in the case of a more even distribution of contrastive elements over the length of the words.

Ideally, words are recognised before their acoustic end is reached. This is typically the case in longer, polysyllabic words. In the word *elephant*, for instance, after the fourth phoneme (i.e. when the sounds [ɛləf] have been heard) no other words remain in the lexicon than *elephant* (and its derivations). In the Cohort model, the lexical uniqueness point (UP, the point from the word onset where it is uniquely distinguished from all competitors in the lexicon) plays an important role. It is at the UP that the listener gets access to the lexical entry, and retrieves all information on the word that is stored in the lexicon (including its meaning, syntactic properties and sound shape). From the UP onwards, the word form is predictable. The listener will check whether indeed the next sounds are as expected, and as a bonus, the listener will know where the next word begins.

4.4 Neighbourhood density

A more sophisticated account of lexical competition during word recognition is offered by the Neighbourhood Activation Model (NAM, Luce and Pisoni, 1998). A practical way of defining a word's neighbourhood is by listing all words that deviate from the target by just one sound. Thus the (British) English word *cat* has a total of 30 neighbours:[4]

bat, pat, mat, fat, vat, that, gnat, sat, chat, rat, hat	(11)
kit, Kate, coat, caught, cot, cart, court, curt, coot, cut, kite	(11)
cap, cab, cam, can, cash, Cass, Cal, catch	(8)

Generally, short words live in densely populated neighbourhoods. Long words live in sparsely populated neighbourhoods. An everyday word such as *computer* has no neighbours at all. Generally, words with many neighbours will be more difficult to recognise than words in small neighbourhoods. This is a matter of lexical redundancy.

Especially when the input sounds are non-prototypical, the human listener cannot definitely rule out competitors. On account of this, short words with many competitors in a dense neighbourhood will be more difficult to recognise. These predictions were born out by the results reported by Luce and Pisoni (1998) in a study in which they carefully controlled token frequencies, neighbourhood density and word length.

4.5 Vowels versus consonants

To conclude this section, let us consider the potentially different contributions of vowels versus consonants to the word recognition process. On the one hand, vowels are louder than consonants; they have more carrying power and can therefore be better heard in adverse circumstances. From a structural, linguistic view, vowels are the heads of syllables.

In spite of the structural and acoustic dominance of vowels, it seems that the contribution of vowels to word recognition is less important than that of consonants. Van Ooijen (1994:110–117; 1996) asked listeners to correct non-words to words in a so-called word reconstruction task. Here the non-words differed from the nearest word in one vowel and one consonant. Replacing either the vowel or the consonant was enough to change the non-word back to a word, as shown in the examples below (three out of a total of 60):

Non-word	Words after V-change	Words after C-change
irmy	army	early
nottice	notice	novice
tisk	task tusk	risk disk

Subjects who were instructed to change vowels only, failed to reconstruct the word in 28 percent of the responses and restored the non-word to the nearest word by inadvertently changing a consonant in another 7 percent of the cases. Subjects who were told only to change consonants failed to reconstruct the word in 42 percent or changed a vowel instead in 15 percent of the cases. When, in a third condition, subjects were left free to choose whether they wished to change either a vowel or a consonant, they opted for each solution in equal proportion. Crucially, however, when they opted for a vowel substitution, reaction time was much faster (1595 ms) than when they resorted to consonant substitution (1922 ms).

It is not entirely clear why vowels contribute less to the identity of words than consonants. It is true that languages typically have more consonant than vowel phonemes. So, from an information-theoretic point of view it should be easier to restore the vowels than to restore the consonants simply because the number of alternatives to choose from is smaller in the case of vowels. Next, in most languages there are more consonants in the shape of words than vowels. Even though CV is the optimally simple and universally preferred syllable type, most languages have more complex syllable types as well. The number of vowels per syllable will always be one, no more, no less. The number of consonants will be at least one, but often more. This skew would also lend more importance to consonants in word recognition. It may also be the case that all vowels resemble each other more than consonants resemble other consonants. Vowels typically only differ in their formants coding height (F1) and backness/rounding (F2). Variation in duration and nasality is secondary (and the same two features are also available for consonants). Consonants differ in many more dimensions, and the acoustic differences along the various dimensions seem to be more contrastive.

Given the evidence presented above, then, it would be reasonable to expect deviations in vowels to be less damaging when listening to speech in a related language variety than deviations in the consonants. Gooskens, Heeringa and Beijering (this volume) are the first to examine the relative weight of vowels versus consonants in the context of intelligibility of related languages. In their correlational study they found that the intelligibility of 17 Scandinavian (Danish, Norwegian, Swedish) dialects for Standard Danish listeners, as determined by a functional translation test, could be predicted better from deviations in the consonants ($r = -.74$) than in the vowels ($r = -.29$). It would make sense, therefore, to incorporate the different contribution of vowels versus consonants in future models of intelligibility of related languages.

5. ROLE OF PROSODY

5.1 *Defining (word) prosody*

Prosody is the ensemble of all properties of the speech signal that cannot be accounted for by the properties of the constituent phonemes in their early-to-late order (van Heuven and Sluijter, 1996 and references therein). An example of prosody at the word level is stress. Stress is defined here as the abstract linguistic property of a word that tells us which syllable in the word is stronger than any other. In a language with stress, every (content) word has a stress position.[5] The sounds in a stressed syllable are pronounced with greater effort, which results in (i) longer duration, (ii) more extreme articulatory positions (spectral expansion of vowels), (iii) greater loudness (higher intensity and flatter spectral tilt) and (iv) more resistance to coarticulation. When a word is communicatively important in the discourse (depending on the intentions of the speaker) the stressed syllable in the word is additionally marked by a conspicuous change in vocal pitch (a rise, fall, or both).

Some languages have so-called fixed stress; the position of the stress is fixed for the entire vocabulary by a single rule. In Finnish (and related languages) the stress is always on the first syllable. In Polish, the stress is always on the prefinal syllable. In languages with fixed stress, hearing a stress tells the listener where one word ends and where the next word begins. This demarcative function may be important in the perception of continuous speech, as a way to reduce the problem of finding the word boundaries. I am not familiar with any research on perceptual problems caused by incorrect stress in languages with fixed, demarcative stress.

Other languages may have variable, or contrastive, stress. Here the position of the stress differs from one word to the next. Either the stress position can be derived by a set of rules (weight-sensitive stress systems) or has to be learnt by heart for each word in the vocabulary as a lexical property. In such languages identical segment strings may yet be distinct words solely because they differ in the position of the stress. An example would be the English minimal stress pair *trusty* ('trustworthy', initial stress) versus *trustee* (board member of a foundation, final stress). The number of minimal stress pairs in Germanic languages is very limited. Therefore, it seems unlikely that the primary function of stress in such languages is to differentiate between words (Cutler, 1986). Rather, it would appear that differences in stress position allow the listener to subdivide the vocabulary into a small number of rhythmic types, within which words can be recognised more efficiently because of the reduced lexical search space.

As far as we know, the majority of the languages in world have stress. Other languages have lexical tone.[6] In a prototypical tone language any syllable in a word may be pronounced with a different melody, for instance at a high tone (H) or at a low tone (L). In such a tone language there would be four types of word melody on two-syllable words: HH, HL, LH and HH. It is not the case that prominence (or greater perceived strength) is associated with either the H or the L tone; this is the crucial formal difference between stress and tone. The primary function of tone would be to help differentiate between words in the lexicon. Tone languages such as Mandarin, a language with four lexical tones, contain many minimal word pairs, triplets and even quartets, that only differ in the tone pattern. An often cited example is the Mandarin syllable /ma/, which means 'mother' with high level tone (HH), 'hemp' with mid-rising tone (MH), 'horse' with low dipping tone (MLH) and 'scold' with high falling tone (HL). One would expect word recognition, even in connected speech, to depend considerably on tonal information; the role of tonal information should increase and be more or less indispensable when speech is heard in severe noise. There is very little research on the role of lexical tone in speech recognition, and virtually none at all when it comes to understanding speech in a closely related language. I will present some (preliminary) data below. In the next two sections I will first review some work on the role of stress in word recognition and then of tone.

5.2 Stress and word recognition

It has often been remarked that the contribution of stress to the process of word recognition should be a modest one. Orthographies reflect effects of stress only in exceptional cases. In the writing systems of European languages, the position of the stress is not indicated in the spelling (with the exception of Spanish, which writes accent marks on syllables with stress in irregular position). Word tones are not written in the orthographies of Norwegian, Swedish, Serbo-Croatian and Welsh. The basic idea is that the words in languages can be recognised from their segmental make up, and that word prosody is largely redundant (especially in sentence contexts).

Table 2. Percent correctly named words (left) and naming latency of correct responses (ms). Words with a melodic accent synthesised on the lexically stressed syllable are listed along the main diagonal of the matrix (boldface).

Lexical stress	Stress synthesised on syll #			Stress synthesized on syll #		
on syll. #	1	2	3	1	2	3
1	**66**	44	56	**1500**	1800	1650
2	34	**81**	31	1630	**1510**	1640
3	34	25	**63**	1700	1690	**1390**

My take on the role of stress (and of prosody in general) is that it is extremely robust against noise and distortion. Because it is a slowly varying property of the speech code, it will normally not be needed in the recognition of words when listening to speech in one's native language. However, when communication suffers from noise, prosody fulfils the role of a safety catch. Listening to speech in a closely related language is basically listening to speech in noise. So, in these circumstances I would predict that stress is important to word recognition; especially incorrect stress, i.e. stressed realised in unexpected positions, will be highly detrimental to word recognition. Such effects should be even stronger when the language does not have stress but word tone (see below).

If it is true that stress becomes more important as the quality of the input speech degrades, we predict that word recognition will suffer if stress is on the wrong syllable in low quality speech. This was clearly shown in van Heuven (1985). Correct recognition of words synthesised from low-quality diphones was severely reduced (and delayed by 120 to as much as 310 ms) if medial or final stress was shifted to an incorrect position in Dutch words. However, shifting an initial stress to a later position was less detrimental in terms of percent correct naming but still yielded severe delays (see Table 2).

On the basis of such results I would predict that unexpected stress positions play an important negative role in understanding speech in a closely related variety. Given that the sounds in the related language do not match the prototypes of the listener's system, word prosody will assume a more prominent role. Now, if the stress were marked in the wrong position, chances of the listener accessing the right portion of his lexicon are very small, and failure of the word recognition process will be the result.[7]

5.3 Lexical tone and word recognition

Gooskens and Heeringa (2004) studied judged distance between 15 Norwegian dialects. In Norwegian, stressed syllables may have one of two different tones (unstressed syllables have no tone), which makes it a restricted tone system.

Similarly, Tang and van Heuven (2007) investigated judged distance and judged intelligibility between 15 (Mandarin and non-Mandarin) Chinese dialects – with four to nine lexical tones, depending on the particular dialect. To get some grip on the contribution of tonal information to distance and intelligibility, speech materials in both studies were presented both with full tonal information and in monotonised versions (using PSOLA analysis and resynthesis, a technique that allows the researcher to change the melody of a speech utterance but leaves the segmental quality unaffected).[8] The results of both studies showed that removing tonal information from the speech utterances did not clearly influence the judgments by the listeners – except that they were somewhat less outspoken.

Yang and Castro (this volume) computed tonal distance between dialects of tone languages (Bai and Zhuang) spoken in the South of China, close to the Vietnamese border. They then regressed tonal distance (computed in several different ways) against functional intelligibility measurements and found that segmental and tonal distance correlated roughly equally strongly with intelligibility (both around r = .7). irrespective of the method used. Curiously enough, Tang (2009), who correlated similar tonal distance measures with both functional and judged intelligibility measures for all pairs of 15 Mandarin and non-Mandarin Chinese dialects obtained no r-values better than .4.

In order to get some idea of the relative importance of tonal information for word recognition in tone languages, an experimental set-up is required in which segmental and tonal information is manipulated independently. Such experiments are difficult to find in the literature. Zhang, Qi, Song and Liu (1981) report recognition scores for several versions of Mandarin materials. Recognition of tones was close to ceiling no matter what kind of filtering had been applied to the signals (whether low pass or high pass) while correct identification of segments (vowels, consonants) was severely affected. This shows that tone, like other prosodic features, is an extremely robust property in speech communication. When melodic properties were removed from the stimuli (using resynthesis with noise excitation or excited by a monotonised sawtooth wave), word recognition scores dropped to 24 and 16 percent, respectively; while sentence intelligibility was at 24 and 33 percent, respectively. When the sawtooth excitation was given its original melody, word and sentence scores rose to 50 and 73 percent correct; adding noise excitation (during obstruents) to the frequency-modulated sawtooth source yielded word and sentence scores of 60 and 90 percent correct.

A more direct study on the relative importance of segmental versus tonal information for the intelligibility of a tone language is reported by Zhu (2009). He established the intelligibility of the 25 Mandarin SPIN test sentences (male voice) used by Tang and van Heuven (2009).[9] Sentences were presented with high-quality segments, with moderate loss of quality (low-pass filtered at 1 kHz) and with practically all spectral information removed (low-pass filtered at

Table 3. Intelligibility (per cent correct recognition of sentence-final word) broken down by melodic version (presence versus absence of pitch information) and by segmental information (excellent, reduced, none). Data from Zhu (2009).

Melodic	Segmental quality			
version	High	Moderate	Poor	Mean
Original	97	83	23	69
Monotonised	98	47	10	52
Mean	98	68	17	61

300 Hz). Each of these three versions were presented with full melodic information as well as monotonised (in a fully blocked design). Intelligibility scores were as shown in Table 3.

These results show that information on tones is fully redundant when segmental quality is high. However, when segmental quality is compromised, tone information makes a large contribution to word recognition and sentence intelligibility. The effect is especially important when segmental quality is moderate. Here the presence of tone information keeps intelligibility at a high level; when the pitch information is eliminated, scores drop below the intelligibility threshold (commonly set at 50 percent word error rate).

Note that in the studies reviewed here, there is always some residual information in the signal that carries information on the identity of the word tones. We know that the tones of Mandarin are also cued by differences in duration and by differences in intensity contour. Follow-up experiments are needed here in which these secondary acoustic properties are also controlled in the stimulus materials.

I should also point out that the results reported above on the importance of word tone and of word stress cannot be compared directly. In the stress experiment, stress was either on the correct or in some wrong position, it was never absent.[10] In the tone experiment, the tones were (nearly) absent but never wrong or misleading. Additional research will be needed in order to come to a more balanced view of the relative importance of stress and tone (as two typologically competing manifestations of word prosody) for speech intelligibility.

6. CONCLUSION

The upshot of the review presented in the sections above is that we are still a long way off from being able to predict success in speech understanding (or word recognition in continuous speech, as a more modest intermediate goal) from a comparison of the two languages engaged in semi-communication. At the same time, however, I have tried to show that the problem is not insoluble. Given some realistic simplifications and a substantial research effort to apply known techniques that have proven their value in other contexts, accurate predictions of mutual intelligibility should be feasible.

As a short-term research agenda, I would recommend in-depth, detailed studies of the effects at the lower levels of the linguistic hierarchy on the recognition of words (isolated and in short sentences). We need to establish how the vowels, consonants and word-prosodic categories (stress, tone) are perceived by the listener of a related language. Once we know what perceptual confusions arise due to the deviant phonetic properties of the related input language, can we attempt to predict the effects at the higher linguistic levels (understanding of sentences and paragraphs). And only if we know the precise effects of the deviant input at the phonetic level, will it be possible to predict intelligibility of a related language by comparing source and target languages at the symbolic levels (i.e. by comparing segmental and tonal transcriptions of words and sentences).

REFERENCES

R. van Bezooijen and V. J. van Heuven (1997), ‘Assessment of speech synthesis’, in D. Gibbon, R. Moore and R. Winksi, eds., *Handbook of standards and resources for spoken language systems* (Berlin/New York), 481–653.

C. T. Best, G. W. McRoberts and E. Goodell (2001), ‘Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener’s native phonological system’, *Journal of the Acoustical Society of America*, 109, 775–794.

P. Boersma (2001), ‘Praat, a system for doing phonetics by computer’, *Glot International*, 5, 341–345.

P. Boersma and D. Weenink (1996), *Praat, a system for doing phonetics by computer*. Report nr. 136, Institute of Phonetic Sciences, University of Amsterdam (Amsterdam).

Y. Chen, M. Robb, H. Gilbert and J. Lerman (2003), ‘Vowel production by Mandarin speakers of English’, *Clinical Linguistics & Phonetics*, 15, 427–440.

B. Comrie, M. S. Dryer, M. Haspelmath and D. Gil, eds., (2005), *World Atlas of Language Structures*. Oxford: Oxford University Press.

R. G. Crowder and J. Morton (1969), ‘Precategorical acoustic storage (PAS)’, *Perception & Psychophysics*, 5, 365–373.

A. Cutler (1986), ‘Forbear is a homophone: Lexical stress does not constrain lexical access’, *Language and Speech*, 29, 201–220.

C. Gooskens, W. Heeringa and K. Beijering (this volume), ‘Phonetic and lexical predictors of intelligibility’.

C. Gooskens and W. Heeringa (2004), ‘Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data’, *Language Variation and Change*, 16, 189–207.

W. Heeringa (2004), ‘Measuring dialect pronunciation differences using Levenshtein distance’ (Ph-D thesis, University of Groningen).

V. J. van Heuven (1984), ‘Segmentele versus prosodische invloeden van klemtoon op de herkenning van gesproken woorden’ [Segmental versus prosodic influences of stress on the recognition of spoken words], *Verslagen van de Nederlandse Vereniging voor Fonetische Wetenschappen*, 159/162, 22–38.

V. J. van Heuven (1985), ‘Perception of stress pattern and word recognition: recognition of Dutch words with incorrect stress position’, *Journal of the Acoustical Society of America*, 78, S21.

V. J. van Heuven and A. M. C. Sluijter (1996), ‘Notes on the phonetics of word prosody’, in R. Goedemans, H. van der Hulst and E. Visch, eds., *Stress patterns of the world, Part 1: Background*, HIL Publications (volume 2), Holland Institute of Generative Linguistics (The Hague), 233–269.

R. van Hout and H. Münsterman (1981), 'Linguïstische afstand, dialect en attitude' [Linguistic distance, dialect and attitude], *Gramma*, 5, 101–123.

D. N. Kalikow, K. N. Stevens and L. L. Elliott (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.

P. K. Kuhl and P. Iverson (1995), 'Linguistic experience and the "perceptual magnet effect"', in W. Strange, ed., *Speech perception and linguistic experience: Issues in cross-language research* (Timonium, MD), 121–154.

B. M. Lobanov (1971), 'Classification of Russian vowels spoken by different speakers', *Journal of the Acoustical Society of America*, 49, 606–608.

P. A. Luce and D. B. Pisoni (1998), 'Recognizing spoken words: The neighborhood activation model', *Ear and Hearing*, 19, 1–36.

W. D. Marslen-Wilson and A. Welsh (1978), 'Processing interactions and lexical access during word recognition in continuous speech', *Cognitive Psychology*, 10, 29–63.

D. W. Massaro (1972), 'Preperceptual images, processing time and perceptual units in auditory perception', *Psychological Review*, 79, 124–145.

J. L. McClelland and J. L. Elman (1986), 'The TRACE model of speech perception', *Cognitive Psychology*, 18, 1–86.

J. Morton (1969), 'Interaction of information in word recognition', *Psychological Review*, 76, 165–178.

E. Moulines and E. Verhelst (1995), 'Time-domain and frequency-domain techniques for prosodic modification of speech', in W. B. Kleijn and K. K. Paliwal, eds., *Speech coding and synthesis* (Amsterdam), 519–555.

S. G. Nooteboom (1981), 'Lexical retrieval from fragments of spoken words: beginnings versus endings', *Journal of Phonetics*, 9, 407–424.

S. G. Nooteboom and M. J. van der Vlugt (1988), 'A search for a word-beginning superiority effect', *Journal of the Acoustical Society of America*, 84, 2018–2032.

D. Norris (1994), 'Shortlist: A connectionist model of continuous speech recognition', *Cognition*, 52, 189–234.

D. Norris, J. M. McQueen and A. Cutler (2000), 'Merging information in speech recognition: Feedback is never necessary', *Behavioral and Brain Sciences*, 23, 299–370.

B. A. van Ooijen (1994), 'The processing of vowels and consonants' (doctoral dissertation, Leiden University).

B. van. Ooijen (1996), 'Vowel mutability and lexical selection in English: Evidence from a word reconstruction task', *Memory & Cognition*, 24, 573–583.

S. Peperkamp and E. Dupoux (2002), 'A typological study of stress deafness', in C. Gussenhoven and N. Warner, eds., *Papers from the Seventh Laboratory Phonology Conference* (Berlin), 203–240.

D. Pisoni, B. Greene and H. Nusbaum (1985), 'Perception of synthetic speech generated by rule', *Proceedings of the IEEE*, 73, 1665–1676.

B. Remijsen and V. J. van Heuven (2005), 'Stress, tone and discourse prominence in the Curaçao dialect of Papiamentu', *Phonology*, 22, 205–235.

B. Remijsen and V. J. van Heuven (2006), 'Introduction: between stress and tone', *Phonology*, 23, 121–123.

W. Strange, R. Akahane-Yamada, R. Kubo, S. A. Trent, K. Nishi and J. J. Jenkins (1998), 'Perceptual assimilation of American English vowels by Japanese listeners', *Journal of Phonetics*, 26, 311–344.

W. Strange, S.-O. Bohn, S. A. Trent and K. Nishi (2004), 'Acoustic and perceptual similarity of North German and American English vowels', *Journal of the Acoustical Society of America*, 115, 1791–1807.

L. Sun and V. J. van Heuven (2007), 'Perceptual assimilation of English vowels by Chinese listeners. Can native-language interference be predicted?', in B. Los and M. van Koppen, eds., *Linguistics in the Netherlands 2007* (Amsterdam), 150–161.

C. Tang and V. J. van Heuven (2007), 'Mutual intelligibility and similarity of Chinese dialects. Predicting judgments from objective measures', in B. Los and M. van Koppen, eds., *Linguistics in the Netherlands 2007* (Amsterdam), 223–234.

C. Tang and V. J. van Heuven (2009), 'Mutual intelligibility of Chinese dialects experimentally tested', *Lingua*, 119, 709–732.

H. Traunmüller (1990), 'Analytical expressions for the tonotopic sensory scale', *Journal of the Acoustical Society of America*, 88, 97–100.

H. Wang and V. J. van Heuven (2006), 'Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers', in J. M. Van de Weijer and B. Los, eds., *Linguistics in the Netherlands 2006* (Amsterdam/Philadelphia), 237–248.

S. J. van Wijngaarden (2001), 'Intelligibility of native and non-native Dutch speech', *Speech Communication*, 35, 103–113.

C. Yang and A. Castro (this volume), 'Representing tone in Levenshtein distance'.

E. van Zanten and R. W. N. Goedemans (2007), 'A functional typology of Austronesian and Papuan stress systems', in V. J. van Heuven and E. van Zanten, eds., *Prosody in Indonesian languages* LOT Occasional Series 9 (Utrecht), 63–88.

L. Zhu (2009), 'The relative contribution of segmental and tonal information to the intelligibility of Standard Mandarin' (MA thesis, Dept. English, Shenzhen University).

END NOTES

[1] In some studies a one-dimensional distance value was obtained by having listeners judge the overall distance or strangeness of some language (variety) relative to their own (van Hout and Münsterman, 1981; Gooskens and Heeringa, 2004; Tang and van Heuven, 2007). This measure correlates almost perfectly with judged intelligibility (Tang and van Heuven, 2007), so that it seems that intuitions about linguistic distance are primarily based on intelligibility. However, in the studies mentioned, the varieties were always related to the language of the judges. It would be crucial to check whether listeners also have clear and reliable intuitions on linguistic distance if they do not understand the stimulus languages at all. As far as I have been able to ascertain, such research has not been done.

[2] That foreign accent is a speech pathology is implied by Chen et al. (2003), who published a study of Chinese accent in English in the journal *Clinical Linguistics & Phonetics*.

[3] The logogen model can be seen as an early model that involves the concept of neural networks. In more recent developments of such theories of word recognition, such as Trace (McClelland and Elman, 1986) and Shortlist A (Norris 1994), and computational implementations of the latter in Merge (Norris, McQueen and Cutler, 2000) the term logogen is no longer used but the concept of a word (or stem morpheme) as a configuration of specialised neurons still plays a central role.

[4] Here we will ignore neighbors that could be generated by deletion or addition of a sound, although established practice requires that we include these in the neighborhood.

[5] Monosyllabic function words may have unstressable vowels (lexical schwa).

[6] The World Atlas of Linguistic Structures (WALS, Comrie, Dryer, Haspelmath and Gil, 2005) lists 220 tone languages versus 307 no-tone languages (chapter 13); at the same time it lists 502 stress languages, divided in chapter 14 between 282 with fixed stress (281 in chapter 15) versus 220 with no-fixed stress (219 in chapter 15). Van Zanten and Goedemans (2007:64) estimate that languages with stress-based word prosody, tone-based systems and languages without word prosody occur in 80, 16 and 4 per cent of the world's languages, respectively. Clearly, languages without word prosody are rare; moreover, languages that independently exploit both stress and tone seem to be anomalous and may develop only as a result of contact between a stress language and a non-related tone language (Remijsen and van Heuven, 2005, 2006). It would seem, therefore, that word prosody of the world's languages is either stress-based or tone-based.

[7] These predictions could be made for all other languages with variable (distinctive) stress systems. I do not know what to predict in the case of incorrect stress in languages with a fixed stress system. It has been shown that French listeners, for example, are ‘stress deaf’ (Peperkamp and Dupoux, 2002), since French with its fixed final stress never uses stress to distinguish one word from another. However, French listeners could use stress as a word separator. Whether they do, and what happens when French words are incorrectly stressed, has not been researched in any detail.

[8] PSOLA: Pitch-Synchronous Overlap and Add is an analysis-resynthesis technique in the time domain. For a description see e.g. Moulines and Verhelst, 1995). The technique is widely available through Praat speech processing software (Boersma 2001, Boersma and Weenink, 1996).

[9] SPIN test stands for Speech in Noise test. This functional intelligibility test was developed for use in audiology (establishing the extent of patients’ deafness) by Kalikow, Stevens and Elliot (1977).

[10] In van Heuven (1984) I included Dutch materials with no acoustic marking of word stress at all – by synthesizing words from diphones exclusively excerpted from strongly accented source syllables and omitting all temporal, dynamic and melodic stress marking from the synthesis. Word intelligibility appeared unaffected by this manipulation. It would seem therefore that only stress in incorrect position should be penalized. I similar vein, I would predict that simply removing tone information from Mandarin stimuli (as in the experiments reviewed) is not nearly as detrimental to intelligibility as is pronouncing the words with (phonetically correct) tones of the wrong type.