Making sense of strange sounds: (mutual) intelligibility of related language varieties. A tutorial

Vincent J. van Heuven
Leiden University Centre for Linguistics/Phonetics Laboratory

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 how much 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 how much 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 should be zero. If the listener B gets every detail of speaker A’s intentions, the score should 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 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

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 & Münsterman 1981; Gooskens & Heeringa 2004; Tang & van Heuven 2007). This measure correlates almost perfectly with judged intelligibility (Tang & 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.
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.

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 or 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 realize that the sound shape /hAUs/ refers to the same concept ‘house’ as the obviously cognate Dutch form /h{ys/. Once the Dutch listener has discovered this relationship he may apply this sound transformation to other English forms, such as /lAUs, mAUs, lAUd, snAUt/, ‘louse, mouse, loud, snout’, which all transform regularly to Dutch /l{ys, m{ys, l{yd, sn{yt/. That is to say, the listener has discovered a rule that relates the English sound shapes to their equivalents in Dutch. This would be a simple example of what it takes to crack the code, i.e. bridging the gap between English and Dutch. To keep the problem within manageable bounds, I propose that we 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. Moreover, 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. 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 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 specializations 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.

2. Speech intelligibility

2.1 Word recognition is key

We will define intelligibility in quite practical terms as the percentage of linguistic units correctly recognized 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

---

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.*
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. 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 recognizes 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 Heuven & van Bezooijen 1995). Functional intelligibility tests measure the real thing. They measure to what extent a listener actually recognizes linguistic units (words) in spoken stimuli. A traditional functional test is dictation; here the listener simply writes down what (he believes) the speaker said. Dictation draws heavily on the listener’s 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 recognize 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 a listener has recognized a word, that word will remain active in the listener’s memory for a long time (up to several hours or even a whole day). The next time the listener hears the same word, he will recognize 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 recognized 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 such that a 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 & Heeringa 2004 for Norwegian dialects, or Tang & van Heuven 2007, 2008 for Chinese dialects). In such more complicated experiments, blocking is done through Latin square designs such that 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 synthesized by several competing brands of reading machines and by a human control speaker (Pisoni, Greene & Nusbaum 1985; van Heuven & van Bezooijen 1995). The listener is familiarized 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 & van Heuven (2008) 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% of the variance in the functional test scores goes unaccounted for.

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 emphasizes 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).

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 are away 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 & 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 will categorize 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 categorize the incoming sound. Best and co-workers (Best, McRoberts & Goodell 2001) have set up a typology of what she calls 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) \( C \) (categorized): 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) \( U \) (uncategorized): 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 accidentally claps his 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 categorized, 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 is tested through experimentation.

In the field of second-language learning, the learner’s native language is called the source language, the language to be learnt is called the target language. In a sound assimilation experiment, the learner is asked to categorize 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 & van Heuven 2007) in which Mandarin listeners were asked to identify the 19 vowels of British English terms of 14 Mandarin vowel categories.
The results show, for instance, that any open or half-open English monophthong is assimilated to Mandarin /a/, although some are considered a very poor token (e.g. English /e/ is rated as a token of Mandarin /a/ 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.

3.2 Prediction of sound categorization 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. We 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 normalization 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 normalized data to an automatic
classification procedure such as Linear Discriminant Analysis (LDA; for details on the above procedures, and references, see e.g. Wang & 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 (i.e. with the least number of classification errors) sorted into the native categories.

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 categorize 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 our 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 acoustic dimensions are targeted that are appropriate for consonant classification.

I have not 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. The method does not allow us to directly compare the vowel systems of the languages concerned and predict how listeners of one language would categorize 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 recognize 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. 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 recognize words 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 specialized 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.

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 recognized (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 recognized, 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.

4.2. Frequency effects

Words that we have heard often before tend to be recognized 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 recognized, 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.

These, and other, models of auditory word recognition neatly account for the phenomenon that a listener may recognize a long word without having to keep the entire sound shape of the word in auditory memory. Incoming sound is short-lived.

---

3 This view of human word recognition draws heavily on ideas behind the TRACE model (Elman & McLelland 1985).
All is 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 & Welsh 1978; Nooteboom 1981) claimed that a word can never be recognized 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 & van der Vlugt 1988) – as is implied by the neural network view presented in the preceding paragraph.

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 organizational principle is that words can be recognized 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 recognized 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 [Elf] 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 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 Neighborhood density

An important notion in this connection is the neighborhood density (Luce & Pisoni 1998). A practical way of defining a word’s neighborhood is by listing all words that deviate from the target by just one sound. Thus the (British) English word cat has a total of 29 neighbors: 4

---

4 Here we will ignore neighbors that could be generated by deletion or addition of a sound.
Generally, short words live in densely populated neighborhoods. Long words live in sparsely populated neighborhoods. An everyday word such as computer has no neighbors at all. Generally, words with many neighbors will be more difficult to recognize than words in small neighborhoods. 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 neighborhood will be more difficult to recognize.

4.5 Vowels versus consonants

To conclude this section, let us consider the potentially different contribution 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) 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):

<table>
<thead>
<tr>
<th>Non-word</th>
<th>Words after V-change</th>
<th>Words after C-change</th>
</tr>
</thead>
<tbody>
<tr>
<td>irmy</td>
<td>army</td>
<td>early</td>
</tr>
<tr>
<td>notice</td>
<td>notice</td>
<td>novice</td>
</tr>
<tr>
<td>tisk</td>
<td>task tusk</td>
<td>risk disk</td>
</tr>
</tbody>
</table>

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 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).

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. However, no model of intelligibility of related languages incorporates this variable at this time.

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. 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. 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)

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.

---

5 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.

6 Monosyllabic function words may have unstressable vowels (lexical schwa).
versus *trustee* (board member of a foundation, final stress). The number of minimal stress pairs in Germanic languages is very limited. Therefore, it seems very unlikely that the 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 recognized more efficiently because of the reduced lexical search space.

As far as we know, less than half of the languages in world have stress. Other languages have lexical tone. 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 difference between stress and tone.

It has often been remarked that the contribution of word prosody to the process of word recognition should be a modest one. Orthographies reflect effects of word prosody 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 recognized from their segmental make up, and that word prosody is largely redundant (especially in sentence contexts).

My take on the role of prosody 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. However, when communication suffers from noise, prosody fulfills 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, and especially incorrect stress, i.e. stressed realized in unexpected positions, are highly detrimental to word recognition.

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 synthesized from low-quality diphones was severely reduced (and delayed by about 150 ms) if medial or final stress was shifted to initial position in Dutch words. However, shifting an initial stress to a later position was less detrimental (table 2).
Table 2. Percent correctly named words (left) and naming latency of correct responses (ms). Words with a melodic accent synthesized on the lexically stressed syllable are listed along the main diagonal of the matrix (boldface).

<table>
<thead>
<tr>
<th>Lexical stress on syll. #</th>
<th>stress synthesized on syll #</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66%</td>
<td>44%</td>
<td>56%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>34%</td>
<td>81%</td>
<td>31%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>34%</td>
<td>25%</td>
<td>63%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>stress synthesized on syll #</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500</td>
<td>1800</td>
<td>1650</td>
<td></td>
</tr>
<tr>
<td>1630</td>
<td>1510</td>
<td>1640</td>
<td></td>
</tr>
<tr>
<td>1700</td>
<td>1690</td>
<td>1390</td>
<td></td>
</tr>
</tbody>
</table>

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 should be 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

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 unsoluble. 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.

References


---

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 & 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.


