The listening talker: A review of human and algorithmic context-induced modifications of speech

The listening talker: a review of human and algorithmic context-induced modifications of speech

Martin Cooke[a,b,*], Simon King[c], Maeva Garnier[d], Vincent Aubanel[a,b]

[a] *Ikerbasque (Basque Science Foundation)*
[b] *Language and Speech Laboratory, Facultad de Letras, Universidad del País Vasco, Vitoria, Spain*
[c] *Centre for Speech Technology Research, University of Edinburgh, UK*
[d] *Department of Speech and Cognition, GIPSA-lab (UMR 5216: CNRS, INPG, University Stendhal, UJF), Grenoble, France*

Abstract

Speech output technology is finding widespread application, including in scenarios where intelligibility might be compromised – at least for some listeners – by adverse conditions. Unlike most current algorithms, talkers continually adapt their speech patterns as a response to the immediate context of spoken communication, where the type of interlocutor and the environment are the dominant situational factors influencing speech production. Observations of talker behaviour can motivate the design of more robust speech output algorithms. Starting with a listener-oriented categorisation of possible goals for speech modification, this review article summarises the extensive set of behavioural findings related to human speech modification, identifies which factors appear to be beneficial, and goes on to examine previous computational attempts to improve intelligibility in noise. The review concludes by tabulating 46 speech modifications, many of which have yet to be perceptually or algorithmically-evaluated. Consequently, the review provides a roadmap for future work in improving the robustness of speech output.

Keywords: Speech production, modification algorithms

1. Introduction

It is a common experience to miss information-bearing fragments of speech relayed over public address systems due to the presence of background noise, or to be surprised by an inappropriately-timed interjection from a speaking navigation system while engaged in other tasks which make additional cognitive demands. Speech output, whether live, recorded or synthetic, is used increasingly in conditions where correct reception of the message is not guaranteed. To ensure that the intended message is correctly understood, developers of speech output technology have little choice at present than to resort to rather crude measures such as increasing speech volume or repeating the message, both of which can create an uncomfortable listening environment. Indeed, for those unfortunate enough to work in an

*Corresponding author
Email addresses: m.cooke@ikerbasque.org (Martin Cooke), simon.king@ed.ac.uk (Simon King), maeva.garnier@gipsa-lab.grenoble-inp.fr (Maeva Garnier), v.aubanel@laslab.org (Vincent Aubanel)

environment of frequent, loud announcements, output speech is a form of noise which can contribute to ill health [317].

Are there alternative approaches to maintaining speech intelligibility in challenging conditions? One source of potential techniques comes from the observation of human talkers in similar contexts. It has long been known that human talkers are capable of adjusting their own speech delivery in response to context. Here, the context might be noise, which leads to speech production changes collectively described as Lombard speech, or it might be the listener or audience, leading variously to the adoption of clear speech, foreigner-directed speech or infant-directed speech amongst other listener-oriented styles. In general, talkers appear to make continuous modifications to message generation to suit the needs of the situation, targetting an appropriate place on a continuum from casual (hypo) to clear (hyper) speech [177]. The modifications to speech that characterise each speech style are moderately-well understood at the level of changes to acoustic parameters (e.g., f_0 and segment duration) and acoustic-phonetic mappings (e.g., vowel space). Precisely if and how observed changes in a talker's speech patterns contribute to intelligibility is far less clear.

Complementary approaches to speech modification are motivated by models of the auditory system, signal enhancement or linguistic entropy. A simple technique is to reallocate speech energy to those frequency regions predicted to contribute most to speech intelligibility, perhaps exploiting estimates of the masker spectrum to determine where to boost speech energy to improve audibility. Speech modification algorithms are also influenced by techniques from cognate domains, e.g., dynamic range compression from audio broadcasting. Likewise, information-theoretic concerns such as increasing message redundancy can inspire modification approaches.

This review article examines both human and algorithmic modifications to speech. Starting with a top-level taxonomy of possible goals of modification as seen from the point of view of a listener (section 2), those listening contexts which induce modifications in human speech production are identified in section 3, and the resulting changes in acoustic, phonetic and higher-level parameters are detailed in section 4. The perceptual effect of human speech modifications is examined in section 5. section 6 reviews speech modification algorithms to date. Finally, in section 7 we draw together both behavioural and algorithmic studies into a unified compilation of speech modifications and suggest new directions for research in intelligibility-enhancing speech modification.

2. A listener-oriented taxonomy of speech modification goals

Since maintaining intelligibility is our key focus in this review, one way to look at speech production changes in response to context is in terms of how they might be expected to benefit the listener. In subsequent sections we group both human and algorithmic modifications under the following four categories, which are ordered approximately from low- to high-level.

Promoting audibility Here, the aim is to improve the audibility of the target speech by reducing the energetic masking effect of the noise caused by the interaction of speech and noise at the level of the auditory periphery.

Both simultaneous and non-simultaneous masking are well-described by computational models [e.g., 208, 70] which allow prediction of the benefit of speech modifications. Changing the spectral slope of speech in a masker-dependent manner is an example of an audibility-promoting speech modification.

Increasing coherence Audible fragments of speech which escape masking need to be allocated to a single ongoing linguistic interpretation. This requirement can be challenging when the masker is itself speech, leading to a form of informational masking which increases as the source and target get more similar [42]. One approach to coherence-enhancing speech modification is to attach information to the target speech to ease the process of grouping audible fragments together (e.g., by presenting it from a common spatial direction). Conversely, modifications may attempt to increase the distance between the target and masker to enable their separation into distinct auditory streams [39] (e.g., by changing the f_0 to avoid collision with the background source).

Enhancing linguistic information Information-bearing elements of speech are encoded with a high degree of redundancy, allowing speech to be degraded in various ways before intelligibility declines. For instance, at low levels, syllable-rate temporal modulations are present across the entire spectrum, leading to a resistance against masking, while at higher levels tactics such as lexical repetition or rephrasing also constitute redundant encoding. Modifications in this category aim to actively increase redundancy (e.g., synthesising speech with multiple cues to voicing; match interlocutor word choices), or better approximate canonical forms of speech.

Decreasing cognitive effort Interpreting speech output is often just one of several tasks which listeners are engaged in. Modification approaches might seek to minimise the cognitive effort associated with speech processing. This could involve simplifying the message or modifying it to better meet listeners' expectations, or, in the context of a dialog system, employing back channels to signal agreement or otherwise.

Note that the same modification might help intelligibility in more than one of the above categories. For example, the use of more prototypical vowels could help to increase coherence, by permitting sequential grouping of speech from the same talker, and at the same time decrease the cognitive effort required to process unfamiliar forms of speech [97].

3. Communicative situations which induce speech production modifications

Types of speech that are produced with the goal of improving intelligibility are commonly referred to as 'clear speech' or 'hyper-speech'. However, these labels cover a great variety of situations and speech types that differ in intention. First, there is a distinction between the 'inadvertent or natural clear speech' [31, 165] produced by certain speakers who are intrinsically more intelligible than others (inter-speaker variability), and 'deliberately clear speech' produced by one speaker in adaptation to a perturbed situation of communication or to a listener with reduced comprehension (intra-speaker variability). Most studies of deliberately clear speech use an elicitation procedure in which speakers are asked to imagine themselves speaking to a hearing impaired person [49, 236, 228, 299, 34] or

to a non-native listener [206, 178, 207, 300]. By extension, the term 'clear speech' has been used to designate any kind of hyper-articulated speech that aims at improving speech intelligibility or at preserving it in adverse listening conditions. [298] provides a comprehensive review of clear speech.

For the current review, we avoid using the term 'clear speech' in favour of a more detailed categorisation based on types of speech induced by either interlocutor- or environment-related factors.

3.1. Interlocutor-related speech types

Interlocutor-induced modifications occur in speech addressed to listeners (or machines) who are perceived by the talker to have intrinsically reduced comprehension (i.e., regardless of listening environment), and includes the following categories:

- Infant Directed Speech (IDS), also called 'motherese', 'parentese', 'babytalk' or Child Directed Speech (CDS) [271, 214, 222, 178, 47, 75]
- Speech addressed to children with learning disabilities [36]
- Hearing-Impaired Directed Speech (HIDS) [235, 236, 263, 299, 137, 34, 35, 171]
- Hyper-visual speech and speech addressed to deaf persons [24, 18]
- Foreign Directed Speech (FDS) addressed to non native listeners [102, 206, 221, 26, 260, 270, 300, 179, 258]
- Machine Directed Speech (MDS) [44, 219, 45, 199]
- Speech used when correcting/amending [19]
- Pet Directed Speech (PDS) [47, 265, 17]

3.2. Environment-related speech changes

Environmental modifiations are those which occur when audibility (including self-audibility) is affected by – or perceived to be affected by – the distorting effects of additive, channel or convolutional noise:

- Speech produced in noise, known as 'Lombard speech' (LS; [183]) [278, 146, 48, 226, 58, 115, 128] or speech addressed to a listener whose audition is perturbed by noise [49, 128]
- Speech produced in filtered or reverberant environments [43, 230] or addressed to a listener through a distorted transmission channel [68, 128].
- Speech addressed to a distant person [176, 297, 81, 51, 104, 230, 103, 105]

3.3. Issues

The subdivision into interlocutor- and environment-related contexts is by no means absolute. Speech directed to individuals with hearing impairment might share characteristics with speech produced in response to a perceived loss of audibility caused by noise, for example. At the same time, speech induced by interlocutor and environment-related contexts cannot always be treated as similar or even comparable. Interlocutor-related speech changes typically possess adaptations – a slower speech rate, longer and more frequent pauses, exaggerated articulation – that might be considered as communicative strategies that help the listener to retrieve and decode phonetic cues. In contrast, speech changes induced by environmental factors are primarily characterised by an increase in vocal intensity (accompanied by changes in fundamental frequency and spectrum) compared to speech produced at a comfortable level [254, 262, 293, 279]. Unlike interlocutor-induced speech, the primary goal of increased vocal intensity is not necessarily to enhance the clarity of phonetic cues, but to preserve their audibility, a strategy that can even be detrimental to their clarity [239, 240, 255, 146, 57]. As a consequence, complementary strategies are often observed in adverse conditions whose aim is to simplify lexical and semantic decoding of the message.

Speech adaptations are not only motivated by the improvement of speech intelligibility, but also by social and affective goals. For example, some of the modifications observed in IDS have no effect on speech intelligibility but instead may express affect or aim at getting the infant's attention [89, 295]. Thus, Pet Directed Speech (PDS) demonstrates expressive intonation patterns like IDS, but no similar hyper-articulation [47, 265, 17]. Conversely, FDS and MDS sometimes exhibit exaggerated segmental and prosodic cues similar to IDS, but lack expressive intonation [44, 221, 219, 26, 300, 45] and in the case of 'acted' MDS reduced f_0 range has been observed [199].

Another social adaptation to consider in the framework of speech enhancement is that of alignment (also referred to as speech accommodation or phonetic convergence). Indeed, it is now well known that people tend to partially imitate each other when they interact, at the postural level as well as at respiratory, phonetic or lexical levels [212, 224, 74, 151, 11, 14, 202, 12, 13]. This 'convergence' phenomenon – and its counterpart, divergence – is believed to be primarily driven by social motivations, such as the desire to yield positive affective evaluation by the interlocutor [118]. Nevertheless, this adaptation can also be considered as a strategy of speech enhancement, as it helps to harmonise phonetic and lexical repertoires between the speech partners and may therefore contribute towards successful communication [238].

4. Speech production changes in modified speech

In this section we identify and categorise, according to the taxonomy introduced in section 2, the large body of previous work which has attempted to describe how speech changes as a response to interlocutor and environmental factors. The identification of speech production changes with each of these levels of processing is not perfect: certain speech production modifications can be expected to improve intelligibility by acting at more than one of these levels (e.g., a slower speech rate can increase the likelihood of salient information escaping masking, decrease cognitive

load and at the same time contribute to more canonical speech cues). In these cases, we have described the speech modification at what we consider to be the processing level where the change is most effective.

4.1. Promoting audibility

Preserving and promoting audibility is the main problem in perturbed environments (noise, distance, filtering). To compensate for such environments it has been argued that talkers aim to 'expand speech sonority' [21] by increasing their vocal intensity, reallocating speech energy in the frequency domain and enhancing speech modulation over time. Such strategies can in fact be detrimental to the clarity of phonetic cues [239, 255, 146].

4.1.1. Increased intensity

A global increase of speech intensity is observed in LS [e.g., 312, 172, 79, 278, 32, 146, 48] and HIDS [236], as well as in speech produced at distance [310, 201, 104, 230, 105], with vocal intensity increasing as a quasi-linear function of noise level or distance to the listener [168, 161, 172, 79, 73, 115, 104, 230].

4.1.2. Raised f_0 and spectrum shift towards medium frequencies

The increase in vocal intensity is accompanied by a raised f_0 [278, 30] and a flatter spectral tilt, resulting in enhanced speech energy in medium and high frequencies [241, 236, 273, 278, 203, 146, 290, 48, 242]. These modifications in average f_0 and spectrum can simply be considered as direct consequences of an increase in vocal effort [293, 279]. Nevertheless, the sensitivity of the human ear to sound pressure level is frequency-dependent and is at its maximum around 3 kHz, so that these modifications in f_0 and spectrum may result from an attempt to improve audibility. Furthermore, speech produced in noise displays not only a flatter spectral tilt, but also a specific boost in the amplitude of higher formants around 3-4 kHz [114], similar to the spectrum 'ring' observed in stage actors [22], which may help in projecting the voice to a distant listener. Although speakers who are intrinsically more intelligible than others do not necessarily produce speech with greater vocal intensity, they typically demonstrate enhanced speech energy in the 1-3 kHz frequency band compared to less intelligible speakers [165].

Although IDS is not produced with increased vocal effort but rather at comfortable level [19], it is also commonly characterised by a significant rise in average f_0 [36]. This increase depends on the child's age, gender and hearing experience [216, 157, 319]. While f_0 increases are observed in a great variety of languages, both tonal [292, 319] or non-tonal [274, 245, 126, 19], there is some evidence that they are not a universal feature of babytalk [250]. A similar f_0 increase is also observed in PDS [47] but not in FDS [221, 300], supporting the idea that a higher f_0 may have an affective or attentional function rather than one of speech enhancement.

4.1.3. Enhanced voiced sounds

Vowels and voiced consonants with formant structure (nasals, liquids) are the most robust sounds to energetic masking by a broadband noise [146] and to sound attenuation with distance [233]. In LS and speech produced at distance, speakers increase the intensity and duration of vowels more than consonants [79, 273, 278, 203, 146, 48, 148, 33, 114], but sonorants are no more enhanced than voiceless consonants. Conversely, HIDS is characterized by

a relative boost in intensity for consonants [236, 36], especially for voiceless consonants [49]. Interestingly, IDS also has a higher vowel to consonant ratio (in both intensity and duration) compared to conversational speech [227], even though in that case there is no masking to compensate for. Other formant-related changes observed include increased formant amplitudes in HIDS [236] and narrower formant bandwidths in clear forms of speech [165].

4.1.4. Spectro-temporal energy reallocation

To a first approximation, the amount of energetic masking is determined by level differences across time and frequency between speech and masker. When speaking in noise, a speaker might therefore use either a *boosting* or *bypass* strategy to decrease the amount of energetic masking. Boosting would entail increasing speech level in those time-frequency regions where the background noise would otherwise be *more* intense than the speech, while the bypass strategy would operate by shifting energy concentrations in time and frequency to regions where the background noise is *less* intense. Of course, any reallocation of speech energy is constrained by both articulatory constraints and the need to preserve phonetic cues.

Speech adaptation in noisy environments is indeed significantly influenced by the type of noise [79, 203, 291, 110, 186, 144, 114]. At comparable sound pressure levels (in dB SPL), white and broadband maskers induce greater increases in vocal intensity, average f_0 and energy above 1kHz than multi-talker babble noise [114]. However, when noise types are compared at similar perceived loudness (in dB A), white noise, speech shaped noise, music noise or driving noise do not induce significantly different adaptations in vocal effort [147, 291, 144].

Some studies [203, 147] provided evidence for boosting strategies, demonstrating greater increases of speech energy in frequency bands containing high levels of masker energy. In a single talker experiment comparing speech adaptation to broadband noises filtered by different band-pass filters, the increase in vocal intensity varied with noise spectral tilt for a constant masker level [147].

Other studies have suggested that bypass strategies are operative, showing increases in spectral center of gravity (CoG) when speaking in low-pass noises (multi-babble noise, driving noise or low-pass filtered broadband noise) [110, 186, 187, 114]. There is also evidence that some speakers specifically adjust their f_0 and F1 in local energy minima of a multi-talker noise [110, 114]. However, speakers do not decrease the CoG of their speech spectrum when speaking in high-pass filtered noises [187], calling into question the existence of *active* bypass strategies. At a temporal level, talkers also reduce the overlaps between speech and background in fluctuating noise backgrounds (competing talker and speech modulated noise) [58].

4.2. Increasing coherence

In contrast to *perceptual* studies of factors which promote the formation of coherent auditory descriptions of objects [39, 25], relatively little work has been done to explore what changes talkers make (or are capable of making) to increase the coherence of their speech in the face of competing sound sources. Some related work is reviewed here, but even in these cases it is far from clear whether the talker's goal is to facilitate speech separation and perceptual organisation for listeners.

4.2.1. Increased speech modulation in amplitude and frequency

It is known that a voice is better detected in – and in some cases segregated from – an intense background when it demonstrates large amplitude dynamics [29, 135], temporal fluctuations [140] and increased frequency modulation (e.g., large vibrato, enhanced intonation) [191, 140]. Both LS and naturally-produced clear speech typically exhibit enhanced low-frequency modulations of the intensity envelope [165, 114].

Speakers exaggerate f_0 modulation in many forms of modified speech. Larger pitch excursions are observed in IDS [121, 91, 126, 217, 274, 313, 246, 159], FDS [221, 300] and LS [33, 110, 114]. In IDS, this enhancement of pitch modulations varies as a function of the child's age and gender [274, 158, 216]. In tone languages, these supra-segmental modifications of f_0 are likely to interfere with tone clarity. When talking to infants, speakers appears to prioritise exaggerated intonation contours, while reducing tonal information [221, 223]. On the contrary, in LS and FDS, speakers enhance tonal contrasts in priority to supra-segmental information [221, 223, 326]. Exaggeration of f_0 modulation is observed in IDS for a number of European languages, Japanese and tone languages [126, 92, 169, 158, 84]. However, further examination reveals some cross-linguistic variations in the use of pitch range [92], in the specific enhancement of prosodic cues to phrase and utterance boundaries [95, 113, 315, 115], and in the preservation of rhythm specificities [227].

4.2.2. Informational masking

Informational masking (IM) refers to any reduction in intelligibility once energetic masking in the auditory periphery has been accounted for. IM has many facets, including the difficulty of determining which audible components belong to the target source, a problem which is especially acute when the target and masker are similar in properties such as f_0 or vocal tract length. In principle, a talker might reduce IM through modifications that increase the distance between their own speech and that of the background in a number of ways such as adopting a different f_0 range or mean vocal output level to avoid clashing with the masker, or by simply changing spatial location to provide binaural cues for an interlocutor. However, the amount of informational masking has not yet been demonstrated to have any significant effect on speech adaptation to a noisy environment. Indeed, Lombard effects are very similar in the presence of competing speech and speech-modulated noise [58]. In a multitalker background, vocal effort increases with the number of talkers [186], i.e., with the amount of energetic masking in spite of the *decrease* in IM that accompanies an increase in the number of talkers. Talkers also make temporal adjustments in the face of potential informational maskers, but do so by decreasing the amount of temporal overlap with the masker [58, 10], which can be expected to increase IM. It seems likely that for a talker, minimising energetic masking takes priority over reducing IM.

4.3. Enhancing linguistic information

4.3.1. Segmental cues: vowels

In the acoustic space of the first two formant frequencies (F1 and F2), the distance between vowel categories increases in HIDS [49, 236, 87, 165], IDS [323, 248, 4, 169, 34, 47, 156], speech addressed to children with learning disabilities [36], FDS [206, 207, 300] and MDS [44, 219, 45]. In IDS, this expansion varies with child age and

development [248]. At the articulatory level, lip and jaw movements are exaggerated: vowels are articulated with globally more open and spread lips [236, 178, 125], with greater peak velocities of lip movements [193].

However, not all vowels in an utterance are hyper-articulated, but only those in stressed syllables [1]. Furthermore, vowel space modification in IDS, FDS and MDS does not appear to be part of an homogeneous expansion of the whole vowel system. Instead, only the extrema of the vowel system /a,i,u/ are hyper-articulated, whereas internal phonemic contrasts like [i-ɪ] are not enhanced, and may be reduced [67]. Furthermore, speaking clearly neither reduces the within-category dispersion nor increases the degree of coarticulation [34].

Hyper-articulation is also observed in speech produced in response to environmental changes, but differs from that obserbed in IDS, FDS or MDS. In LS [278, 32, 110, 111], speech produced at distance [176] and prosodic focus [21], the main articulatory modification consists in a global increase of jaw and lip aperture for all categories of vowels [262, 111] and a global amplification of lip opening and closing gestures [154, 71, 112, 96]. As a result, the vowel system is shifted towards higher first formant frequencies rather than exhibiting a global expansion in (F1, F2) space. Despite this shift, most speakers demonstrate an enhanced contrast between open and closed vowels, along the F1 dimension, as well as enhanced visible contrasts in lip opening, rounding and protrusion [145, 111]. At low noise levels, speakers are also able to enhance the contrast in F2 between front and back vowels, resulting in an increased vowel space as for IDS, FDS or MDS [145, 231]. At moderate and high noise levels, however, the F2 contrast is systematically reduced [231], exhibiting decreased rather than exaggerated tongue movements and lip spreading [111, 116].

Finally, the duration contrast between tense /i,u/ and lax /ɪ,ʊ/ vowels is enhanced in IDS for languages such as English, in which vowel duration is an additional discriminative cue [160].

Contrary to the 'hyperspace' hypothesis [142], the density of a vowel system does not appear to influence the degree of its expansion in modified forms of speech. Thus, vowel hyper-articulation in IDS is comparable in English, Spanish [34], Croatian [269], and Finnish [123, 124], although these last three languages have fewer vowels than English.

4.3.2. Segmental cues: consonants

Changes in voice onset time (VOT) for plosives have been reported in HIDS [49, 236], IDS [15, 281, 80, 280, 283] and LS [131]. VOT changes enhance the contrast between voiced and voiceless plosives for some speakers only [189, 283] who rely on this cue as listeners [149]. Other speakers, who rely more on f_0 at the onset of the following vowel for this discrimination, enhance instead that contrast in clear speech [149]. Changes in the VOT of plosive consonants depend more on speaker than language [123, 124].

Fricative consonants are produced with enhanced contrast in the first spectral moment in MDS [190] and speech addressed to older infants (12-14 months) but not younger infants (4-6 months) [66]. However, speakers do not enhance this contrast in noise [131] and may in fact reduce it [231].

Information-bearing formant transitions are generally longer in HIDS [49], while both citation and clear speech

exhibit steeper F2 transitions [207].

Finally, vowel quality and length are additional cues to discriminate voiced and voiceless consonants in some languages, such as English, in which vowels are shortened when followed by voiceless consonants, as compared to when followed by a voiced one. However, this phonological contrast is not enhanced in FDS, and has been found to decrease in LS [258].

4.3.3. Other segmental cues: tones and signs

The enhancement of audible contrasts between phonological categories extend to tone languages. Like vowels, tone contrasts are exaggerated in both IDS [180, 181, 319, 182, 318, 46] and FDS [221], and tone enhancement varies with a child's age and development [182]. In contrast, tones are not more contrasted when speaking in noise but instead produced at higher pitch [326]. The amplification of phonetic gestures also extends to sign languages, in which hand movements are exaggerated when a deaf parent interact with their deaf infant compared to when communicating with deaf adults [82, 192].

4.3.4. Prosodic cues

In addition to the possible coherence-related function of f_0 modulations described in section 4.2.1, raised f_0 and exaggerated f_0 modulation can have different linguistic and communicative functions. Our focus is on their role – in combination with other prosodic cues such as syllable lengthening, variations in vocal intensity – in speech parsing. Several studies support the idea that speakers enhance segmentation cues in modified forms of speech. First, speakers exaggerate syllabification in noise, speech produced at a distance and in HIDS, both by increasing the modulation of the syllable's intensity envelope and by inserting inter-syllabic pauses [110, 114]. Likewise, speakers globally insert more and longer pauses in HIDS [236, 138, 139], to children with learning disabilities [36, 164], to infants [92] or in noise [115]. These pauses are inserted especially before words starting with weak syllables [68] and before phrase boundaries [92, 138]. In IDS, exaggeration of pause duration decreases with the age of the infant [159].

In many languages, syllable lengthening is a boundary cue to the end of a word, phrase or utterance, with greater lengthening associated with higher prosodic levels [20, 143]. In IDS, HIDS and LS, the final vowels or syllables of words [2], phrases [95, 113, 115, 159] and sentences [95, 54, 113, 115] are further lengthened compared to that observed in conversational speech.

Pitch rises or falls complement syllable lengthening to mark lexical, phrase or utterance boundaries. Thus, a pitch lowering marks the end of a declarative utterance in many languages. In French, a bi-tonal LH* (low-high) primary accent marks the end of accentual phrases within the utterance [143], while in Japanese, a pitch rise marks their beginning [309]. In French, a secondary LHi (low-high initial) accent can be found within an accentual phrase [143] and its 'elbow' in the f_0 contour marks the boundary between the beginning of a content word and its preceding determiner [314]. In IDS and LS, speakers exaggerate these intonation cues to sentence, phrase and word boundaries [221, 95, 113, 315, 115].

Lexical stress in languages such as English can also be used as a perceptual cue to word segmentation [93]. Parents

enhance lexical stress when speaking to infants older than 11 months by lending additional stress to strong syllables [308].

4.4. Decreasing cognitive effort

The goal of communication is not to produce speech sounds and gestures but to use them to transmit a message. Thus, improving the audibility and the phonetic clarity of speech sounds and gestures is not the only way to improve speech intelligibility. Another potential strategy is to reduce message complexity by approaches such as utterance simplification, increasing redundancy over time or over different modalities, or by attracting the listener's attention to those parts of the utterance containing the most important information.

4.4.1. Slower speech rate

A slower speech rate is typically observed in IDS [126, 313, 246], FDS [221, 300], MDS [44, 219, 45], HIDS [236], speech addressed to children with learning disabilities [36], and speech produced at distance [310, 105]. Such a decrease in speech rate is also generally observed in LS [241, 146, 148, 155], although this is not true of all speakers [155, 110]. The reduction in the rate at which information is transmitted also applies to sign languages: deaf parents sign at a significantly slower rate when they address their deaf infants compared to the rate used to interact with deaf adult friends [82, 192].

However, a global decrease in speech rate does not necessarily reflect an attempt to improve the reception of phonetic information. Indeed, rate reduction comes mainly from more frequent and longer pauses within the speech stream, with a lesser role for speech segment lengthening [236]. Furthermore, not all segments benefit from speech rate reductions: vowels are lengthened much more than consonants [273, 146, 48] and in LS consonants may even be shortened [146, 48, 155, 186, 114].

4.4.2. Reduction in utterance complexity

IDS is characterised by a reduced vocabulary [50, 249] and by more frequent use of common and short words [234, 249, 324]. The lexicon addressed to children with Down syndrome also presents fewer function words but more onomatopoeias, and demonstrates less lexical variability than the lexicon addressed to typically developing children of the same age [324].

Similarly, speech addressed to hearing impaired children [50, 138, 139], to children with Downs syndrome [324] and to non-native adult listeners [184, 185] has shorter sentences, with less syntactic and morphological complexity than for children with normal hearing and typical development. In response to mis-comprehension, talkers tend to reformulate their utterance using a simpler syntactic structure [304, 303].

Parents talk to their young children with incomplete sentences, with skipped prepositions, pronouns and articles, again mirroring utterances produced by the children themselves [324]. This kind of imitation is supported by other evidences of syntactic [38] and semantic [117] co-ordination between adult speakers in dialogue interaction [although contrary observations have been reported, e.g., 132].

4.4.3. Emphasising important or novel information

At a lexical level, IDS exaggerates f_0 peaks even more on focus words [90] and enhances the existing contrast in f_0 and intensity between content and function words [75], to the point that function words are often skipped [324]. In noise too, words bearing information about agents, objects or locations are stressed in comparison to verbs or determiners [113, 110, 226].

The first occurrence of a word is produced with longer duration than subsequent repetitions of this same word [101, 100, 23]. Similarly, frequent words in a language are produced with shorter duration and reduced articulation, compared to less frequent words [327, 23]. In IDS, parents emphasise new words relative to the rest of the utterance [94].

4.4.4. Increase in redundancy

In both spoken and sign languages, parents repeat their own utterances [50] or the same signs [82, 192] significantly more frequently when addressing to HI children, than normal hearing children or deaf adults.

In adverse noisy environments, some speakers enhance more their lip articulatory movements when their interlocutor can see them, compared to when the interlocutor can only hear them [96]. However, not all speakers demonstrate this strategy [116].

5. Perceptual consequences of speech modification

While speech production changes as a function of communicative situation, there is no guarantee that the observed modifications have a beneficial effect on speech communication itself, and especially on message intelligibility. We detail below evidence of communicative benefits from modifications, first organised by modification context and subsequently by the speech characteristic believed to be responsible for the benefit.

5.1. Benefits as a function of modification type

HIDS benefits speech perception by listeners with severe hearing loss both in quiet [235] and in simulated noisy/reverberant environments [228, 299]. However, in adverse listening conditions, while HIDS improves intelligibility globally, it does not appear to improve the recognition and discrimination of segments [87]. HIDS also benefits elderly listeners with moderate hearing loss in noise and reverberation [263], and in both audio and audiovisual domains [134]. However, no relationship was found between the amount of benefit and the degree of hearing loss in elderly listeners [134]. HIDS also improves speech perception by normal-hearing listeners in noise/reverberation [49, 228, 299, 87, 163, 86], in audio-only, audiovisual and visual-only domains [107, 106, 133, 108], at both a global and segmental level [235]. The intelligibility benefit of HIDS increases with language experience, being larger for native than non-native listeners [35], and for adults compared to school-age children [36]. Independent of the type of listener, benefits typically increases with the level of acoustic degradation of the environment [228].

LS globally improves the intelligibility of words and sentences in the audio domain [76, 278, 242, 53, 186], although the degree of Lombard gain varies both with the type and level of the noise used to induce the Lombard

effect [186]. No evidence was found to support the idea that LS utterances would be more intelligible when perceived in the same background noise as the one in which they were produced [186]. LS also benefits non-native listeners to a similar degree as native listeners, but when presented to the former group in noise-free conditions results in a small *loss* in intelligibility [57]. At the segment level, all studies apart from [146] showed an increased intelligibility of vowels and consonants produced in noise in the audio domain. However, the intelligibility gain from the inclusion of visual information is weaker in LS that in speech produced in silence [53, 71, 305].

Concerning IDS, the use of exaggerated acoustic characteristics in the early preverbal period is assumed to primarily aim at attracting the infant's attention and encouraging interaction [257, 256, 274]. Indeed, numerous studies have shown that infants pay more attention to IDS [e.g., 88, 316, 229, 150, 62, 78] and demonstrate increased brain activity when listening to it [325, 211]. It seems probable that f_0 variations are mainly responsible for this increased attention, as babies show similar preference for sinewaves reproducing IDS intonation [63]. IDS is also believed to facilitate the recognition and discrimination of phonetic cues by infants, and to help them develop phonological representations. Indeed, infants' discrimination of sound categories appears to be directly related to their parents' vowel space area [180] and more globally to the distribution of acoustic cues they are exposed to [198, 197]. The exaggeration of audio and visual contrasts in IDS has been shown to improve infants' discrimination of vowels [296] and to enable them perceive non-native phonemes that they do not perceive in Adult Directed Speech form [218]. Finally, IDS also helps young children recognize words [272] and maintain that recognition over the longer term [266]. The perceptual benefits of IDS for other listeners is controversial: non-native listeners benefitted from IDS in their acquisition of new words [120], an observation that [321] failed to replicate.

5.2. Perceptual effects of modified speech

For the specific perceptual effects of clear speech, see also the review in [298].

5.2.1. Intensity increases

The increased intensity of forms such as LS naturally explains part of the intelligibility gain over speech produced in quiet, but other aspects are also responsible given that a significant Lombard gain remains even when intensity differences are removed [76, 278, 146, 186]. The increase of vocal intensity can affect intelligibility in other ways. Increased vocal effort is accompanied by an increased f_0, a flattening of spectral tilt in the medium-high frequency region, by larger mouth and jaw movements, and by a higher F1 frequency [254, 262, 293, 279]. While spectral tilt changes can be beneficial (see below), some of the concomitant acoustic modifications are detrimental to segment intelligibility [239, 255].

5.2.2. Spectral energy distribution

Shifts in the spectral energy distribution towards the mid-frequency region, as observed in LS, is very effective in enhancing intelligibility in noise [267, 187]. However, the artificial boost of speech energy between 1 kHz and 3 kHz, as observed in the most intelligible talkers, accounts poorly for their intelligibility [166].

5.2.3. Changes in mean f_0 and f_0 contour

The increased average f_0 observed in LS, HIDS and speech produced at a distance does not contribute to the intelligibility benefits brought by these kinds of speech in quiet or in noise [31, 37, 162, 8, 129, 16, 188, 199]. Synthesis of natural f_0 contour variations can improve intelligibility in comparison to a flat contour [174, 173, 311]. However, the wider f_0 range observed in IDS, LS and FDS does not improve infants' ability to recognize words [272] and appears not to contribute to the intelligibility benefits of these forms of speech [162, 199].

5.2.4. Slower speech rate

Although clear speech can be produced at a 'normal' speech rate, the clear speech intelligibility benefit is modulated by speech rate: increased by a slower speech rate, and decreased by a faster speech rate [163]. Speech rate is one of the modifications of IDS that significantly improve infants' ability to recognize words [272] and that may explain some of the prosody-related intelligibility gain of IDS, FDS, HIDS and LS [199]. However, the artificial slowing of speech does not lead to significant intelligibility improvements [261, 122, 205, 237, 299, 213, 61]. No systematic correlation has been found between speech rate and the intrinsic intelligibility of speakers [64, 31, 37, 129].

5.2.5. Pausing

The more frequent insertion of pauses has been related to increased intelligibility [236, 36]. However, the artificial insertion of pauses in conversational speech is not an effective speech enhancement technique [299] and in one study led to reductions in keyword scores for sentence material [287], probably due to the disruption of listeners' expectations.

5.2.6. Vowel enhancement

Based on observations of *written* English [264, 175], consonants are commonly thought to carry more information about sentence intelligibility than vowels. However, apart from [220], it has been found that vowel information enables better recovery of spoken utterances than consonant information [55, 153, 98, 99]. Additionally, the narrowing of formant bandwidths has been related to the intrinsic intelligibility of speakers [165].

5.2.7. Vowel space changes

Vowel space expansion in the audio domain and the exaggeration of visible articulatory movements of the lips is related to improved recognition and discrimination of vowels, syllables and words by infants [180, 272] and in noise [107, 133, 18, 87, 108, 252]. However, the intrinsic intelligibility of a speaker does not appear to be systematically related to greater vowel dispersion [31, 37, 130, 129]. Artificial boosting of the amplitude of F2 and F3 has been found to improve intelligibility in noise for normal hearing listeners, but not for those with hearing impairment [162].

5.2.8. VOT

A clear relation has been established between longer VOT and the intelligibility benefits of clear speech at normal speaking rates [165]. However, variations in VOT do not have any effect on the intrinsic intelligibility of speakers [31] nor on the intelligibility benefits of HIDS [204, 200].

6. Speech modification algorithms

Our review of algorithmic modifications to speech follows the taxonomy introduced in section 2 used to classify human speech modifications. Unsurprisingly perhaps, most of the available algorithms for improving speech intelligibility operate on the speech signal, and therefore, in the main, aim to improve audibility in noise (section 6.2). There are some examples of modifications at higher levels though, especially in the case of spoken dialogue systems. Before examining the algorithms themselves (sections 6.2-6.5), we discuss some of the assumptions and constraints which apply to these techniques.

6.1. Preliminaries

6.1.1. Assumptions

Although sharing the aim of maintaining a target level of intelligibility under challenging conditions, speech modification differs from speech enhancement in that it assumes the existence of a noise-free speech signal. For example, while a considerable amount of work has been done to improve the intelligibility of speech through hearing aids, they of course operate on a speech-and-noise mixture – speech enhancement – so fall out of scope for this review. However, where such algorithms specifically target speech, and could be expected to offer intelligibility gains if applied to clean speech before mixing with noise, we included them (although their performance in this scenario may not have been tested). It must also be noted that methods that improve intelligibility for the hearing-impaired will not necessarily help normal-hearing subjects.

The kind of speech modifications we consider also differ from active noise cancellation approaches in that we assume the noise signal cannot be modified. However, it should be noted that modifications to the speech signal can be designed to mask information in the masker (e.g., the allocation of energy to the speech signal in spectro-temporal regions whose effect is to mask masker transients). This idea is explored further in [9].

6.1.2. Constraints

Two constraints are assumed in most modification algorithms. While increasing speech level is included for completeness, in general the modifications of interest are those that, on average, operate without level or loudness increases. Secondly, durational adjustments (e.g., segment lengthening) are possible, but without resulting in an excessive increase in overall duration.

Application of these constraints leads to side-effects and tradeoffs that are not always apparent when considering speech modifications in isolation. For example, boosting formants leads to reduced energy away from formants, and lengthening vowels or inserting pauses requires increases in speech rate elsewhere. More generally, modifications that are designed to be beneficial in noise may distort speech when applied in quiet conditions, and might even reduce intelligibility for those groups (e.g., non-native, hearing-impaired or young listeners) whose performance is already below ceiling in noise-free conditions (see, e.g., [57]). Our focus in this review is on intelligibility rather than naturalness.

6.1.3. Types of speech output

The specific modifications that can be applied to speech will depend on the type of speech itself. For delayed live speech, the range of modifications is limited in practice to those changes that can be applied to short segments of the speech signal (e.g., range compression or spectral filtering). For pre-recorded natural speech, content and style information can be extracted offline, allowing for a greater range of modifications at deployment time (e.g., vowel space expansion, f_0 range expansion). For speech generated from text, a full spectrum of modifications is available, including high-level modifications such as choice of syntactic form or lexical substitution. The boundaries between live-recorded and recorded-synthetic are not always clear cut. For instance, if sufficient delay is permitted, higher-level linguistic information could in principle be extracted from live speech. Similarly, lexical substitutions might be considered for pre-recorded natural speech by generating words as a result of adapting synthesis models on fragments of the pre-recorded speech.

6.1.4. Knowledge of the context

Knowing something about the context (e.g., masker or target listener) can, in principle, permit the use of modifications that are tailored to the specific masking pattern or known listener deficits. For some application domains there is a realistic prospect of acquiring detailed noise estimates, either online or offline (e.g., transport-related noise in railway stations where the output devices are fixed) while for others the options are more limited (e.g., environmental noise for mobile devices). Modification candidates differ in the degree of detail required in estimates of context. For instance, for masking noise, this might range from high resolution spectral information in successive time-frames through long-term masker spectral profiles to complete noise-independence. For target listeners, variables such as first language or age might be required to tune a given type of modification.

6.2. Promoting audibility

This approach to improving intelligibility is the most obvious and the methods in this category are some of the simplest to implement, with many being possible in real time with low latency.

6.2.1. Increase intensity

Global increases in intensity may be placed under the listener's control, or may be made automatically. When the attack and decay times of dynamic amplitude compression (section 6.2.5) are very long (seconds rather than milliseconds), it may be known as Automatic Gain Control (AGC) and can be widely found on personal music players and in-car audio systems to even-out the volume of different material or to adjust the volume according to an external control signal. Thus, AGC can be made noise-dependent, with the amount of intensity increase being set in accordance with ambient noise, or directly from another source, such as a vehicle's speed. This approach may be employed on mobile phones, enabling the speech received to remain audible to the listener as the noise environment changes. Whilst this does reduce the dynamic range of the signal, AGC is better categorised as producing relatively slow adjustments to overall intensity rather than rapid short-term modifications.

The system described in [225] modified the amplitude of salient words (by which they simply mean content words) using a fixed increase of 4 dB. Although this was motivated by experimental evidence from human speech, no experimental results are provided as to the effectiveness of this modification on the intelligibility of the synthetic speech.

6.2.2. Decrease spectral tilt

Simple high-pass filtering will decrease spectral tilt. If this is followed by energy normalisation back to the original signal energy, this also effectively boosts transients (section 6.2.7) because low-frequency energy is reduced.

[328] uses spectral shaping that includes a decrease in spectral tilt to improve the intelligibility of speech in noise. This is then combined with dynamic range compression (DRC – section 6.2.5). Combining spectral shaping with DRC (and, crucially, in that order) makes a lot of sense, since in unmodified speech the bulk of the signal energy is in the low frequencies and this consumes much of the available headroom when boosting the amplitude of the waveform. This has long been recognised in music production, where it is common to split signals into several non-overlapping frequency bands and to apply DRC (known there as compression) to each band separately (and with different thresholds and ratios), before re-combining them. Similarly, high-pass filtering is a standard procedure in audio recording, production and broadcast, removing those low frequency components that would consume headroom (of either the transmission channel, storage medium, or reproduction apparatus) but contribute little to perceived quality.

While hearing aid processing in general is out of scope of this paper, the findings reported in [72] are still of interest. It is noted that patients' preferred settings for quality are not the same as those that give maximum intelligibility. This is relevant to the design of intelligibility-boosting algorithms because it implies that some loss of quality may reasonably be incurred when maximising intelligibility gain. Encouragingly, [72] also found that there were two simple strategies that worked for everyone, independent of patterns of individual hearing loss: a uniform boost (i.e., the increasing intensity strategy), or a 6dB/octave high-pass filter (i.e., reducing spectral tilt). However, the evidence is not completely clear, with [136] finding no significant differences for high/low frequency boosts in hearing aids.

6.2.3. Narrow formant bandwidths

As [215] points out, even simple high-pass filtering effectively boosts the second formant relative to the first formant, thus enhancing intelligibility due to the greater importance of F2 frequency. More targeted, specific formant enhancement of course requires reasonably accurate estimates of formant frequencies, which will always be prone to error. [28] describes real-time formant peak sharpening and tests it on hearing-impaired listeners, producing small gains in intelligibility. Enhancement is performed using variable-centre-frequency bandpass filters, and it is noted that choosing the bandwidth involves a compromise between wide filters offering little enhancement versus narrow filters that can lead to sudden amplitude variation as a consequence of harmonic peaks moving in and out of bandwidth (and we presume, although not mentioned in the paper, also because of errors in formant tracking). [3] follows on from [28] and demonstrates modest benefits to normal hearing listeners too in the presence of multi-talker babble. The

benefits were only for vowel perception.

6.2.4. Sparsify energy

Although there do not appear to be algorithms that deliberately sparsify speech signals, the vocoders used in speech coding and statistical parametric speech synthesis do this unintentionally: fine detail is removed, harmonic structure may be over-emphasised and useful redundancy is lost through the processes of spectral envelope estimation and modelling. It seems unlikely that this has positive benefits on intelligibility and, indeed, it is plausible that it is this lack of redundancy that makes synthetic speech intelligibility degrade much more rapidly than that of natural speech as SNR worsens. [28] proposes one form of sparsification in which speech is resynthesised using a very simple two-formant filter and suggests there could be benefits for listeners with poor frequency resolution, although this was not tested.

Intriguingly, an attempt to find optimal static spectral weightings [288], described below (section 6.2.9), *did* produce sparse representations which improved speech intelligibility by around 2 dB in a large-scale evaluation [60].

6.2.5. Dynamic amplitude compression

Depending on the application and the attack and decay times employed, dynamic amplitude compression (or Dynamic Range Compression – DRC) goes under a variety of names. Slow-response systems are commonly known as Automatic Gain Control. In audio and especially music production, with generally fast (1-10 msec) attack times, it is widely called 'compression' or occasionally 'companding'.

Both an AGC, which aims to provide a standard speech level, and an automatic level compensator (ALC), which modifies the speech level in response to ambient noise level, are using in the 'interphone' (headset-based communication for face-to-face situations with extremely high ambient noise levels, such as inside a noisy military vehicle) described by [294]. The AGC reduces the dynamic range of the amplitude of the speaker's speech to compensate for variations in their speech – a simple form of speech modification – whilst the ALC promotes audibility in the listener's varying noise conditions – a simple form of listener-dependent behaviour.

Reducing the dynamic range of signals is ubiquitous in broadcast and recorded media production, for both music and speech. In such a compressor, when the level of the input signal exceeds a user-defined threshold, its amplitude starts to be attenuated at a user-defined ratio. Inertia is applied to the attenuation control via controllable attack and release times.

One problem with all forms of dynamic amplitude compression is that any background noise present in speech pauses will be boosted in amplitude. Short attack and long release times can mitigate this, but very short attack times mean that transients will be eliminated. For speech signals this can lead to the loss of initial consonants: [294] suggest that users say an 'expendable word' at the start of each speech period! Modern systems are able to employ small delays to implement lookahead in order to avoid this mistreatment of transients. Actively boosting transients has been used to increase intelligibility (section 6.2.7). As described in section 6.2.2, [328] uses spectral shaping followed by dynamic range compression to improve the intelligibility of speech in noise.

Extreme forms of dynamic amplitude compression can be found. With the threshold set just below the clipping level and a very high compression ratio, one obtains a 'brick-wall limiter'. This is essentially what is used in [215] (their figure 2). If too much gain is applied, the waveform becomes clipped (in either analogue or digital domains). It has been found that this may slightly enhance intelligibility in some circumstances [243] although [167] shows that a fast limiter is more effective than clipping.

6.2.6. Increase/decrease f_0

There are many methods for modifying the fundamental frequency of speech, either directly in the time domain (e.g., using TD-PSOLA [210]), or on a coded representation of speech, as in Dudley's vocoder [77], or using one of many modern methods such as harmonic-plus-noise models [276] or the STRAIGHT vocoder [152]. Whilst these methods are widely employed to manipulate f_0 for the purposes of prosodic modifications, few automatic systems have the express intention of increasing intelligibility. A number of investigations into the effects of f_0 modification [188, 302] have found no benefits. [225], already mentioned in section 6.2.1 modified the f_0 of content words using a fixed increase of 20 Hz. However, its effectiveness was not evaluated.

Recently, [306] modified speech by seeking the change in f_0 which optimised a 'glimpse' proportion measure of enegetic masking [56]. In general, the outcome was a reduction in f_0, probably caused by the increased number of resolved harmonics that such a change produces. While the objective intelligibility measure predicted a modest gain, a later listener-based evaluation [59] demonstrated modest losses in actual intelligibility.

6.2.7. Boost transients

[247] found transient-boosting methods can improve intelligibility of speech in noise (recordings of an aircraft auxiliary power unit) presented at various SNRs from -30 dB to -10 dB, and that gains were still seen when combining this with active noise-cancelling headphones. Although the best-performing method was inspired by a dynamic time-varying filter approach, in order to obtain real-time performance this was approximated as a simple fixed filter, which amounts to attenuation below 700Hz and mid- and high-frequency boosting. So, although the method does boost transients, as any high-pass filtering would, it is not actually detecting them in the speech. A more sophisticated method in the same study using wavelet decomposition of the speech – which did target transient regions more specifically than the fixed filter – offered lesser gains. Distortions introduced by that method may have counteracted the benefits of transient boosting. From the results available in [247] and from simple spectral shaping, it is not possible to separate out the effects of a general high-frequency boost from specific localised transient boosting. [322] also found that increasing the intensity of transients relative to steady-state regions increased intelligibility.

6.2.8. Steady-state suppression

Pre-processing of the speech signal by steady-state suppression has been attempted in order to reduce the smearing of energetic speech parts (e.g., vowels) into subsequent segments, which results in forward masking in reverberant environments. However, a modulation filtering technique operating at a syllable rate produced no clear intelligibility benefit [6, 170].

6.2.9. Spectral profile modification

Spectral profile modification is a generalisation of high-pass filtering, spectral tilt or centre of gravity changes and is typically the outcome of an optimisation procedure based on avoiding energetic masking by a known noise. The response obtained during optimisation depends, of course, on the spectral resolution. In a relatively low-resolution approach, [285] sought the optimal linear time-invariant filter which maximised an approximation to the Speech Intelligibility Index [5]. In another low-resolution technique, [232] reallocated energy by modifying the speech with a simple filterbank, adjusting the filter gains to maximise intelligibility as measured using automatic speech recognition. [289] also sought a static spectral filter but at a much higher resolution based on a 55-channel gammatone auditory filterbank. Their optimisation process employed a genetic algorithm which attempted to maximise the glimpse proportion [56] across a corpus of speech, independently for different maskers and SNRs. A striking outcome was the finding, hinted at earlier, that as the SNR decreases, the optimal filter became increasingly sparse, focusing the boosting of energy in 3 or 4 widely-separated frequency bands.

6.2.10. Energy reallocation

Energy reallocation is more general still, denoting techniques that shift energy in both frequency and time. A study by [286] compares various manually-designed energy reallocation strategies and found that boosting the SNR of selected frequency bands gave the largest gains but also tended to reduce the speech quality the most (both measured only with objective measures). As in the method in [259], it is most effective to carefully choose to improve the SNR only of regions where improvement is possible: that is, to avoid processing clean speech or speech that is much lower amplitude than the competing noise signal such that no amount of boosting will improve its audibility.

In [284], energy is reallocated in energy over time and frequency guided by a spectro-temporal auditory model. The temporal aspects of the processing result in transient enhancement (see section 6.2.7). Improvements in intelligibility are claimed, although only objective measures are presented. [268] reallocated energy across time only, from voiced regions to unvoiced regions, under a constant energy constraint (per word), compared this with a high-pass filter that reallocated energy across frequency only, and obtained intelligibility improvements for around half of the 16 speakers compared in their listening test.

The approach of [301] reallocates energy within frames via manipulation of a cepstral representation of the spectral envelope, in order to optimise a computational model of intelligibility based on the number of glimpses of the speech signal [56].

6.3. Increasing coherence

While there have been numerous laboratory studies within the domain of auditory scene analysis which demonstrate the value – or otherwise – of specific organisational cues such as those which encode fundamental frequency [see 69, for a review], there have been surprisingly few explicit engineering attempts to modify speech with the aim of increasing coherence, perhaps due to the paucity of observations of coherence-inducing modifications in speech production.

We speculate that synthetic speech generated using waveform concatenation has impaired coherence due to use of recorded units in mismatching contexts, and that this could be one reason that this type of synthetic speech is typically less intelligible than natural speech whereas synthetic speech generated using statistical models driving a vocoder can be as intelligible as natural speech [320], even under certain additive noise conditions [282, 59].

6.4. Enhancing linguistic information

While in practice it is straightforward for a text-to-speech system to enhance salient linguistic information (e.g., via lexical repetition or adding references to previous information) very few studies have been carried out to explore the benefits of this class of modification. For instance, it is common in limited-domain automatic speech recognition systems, where the application design permits, to select a word vocabulary that minimises recognition errors [244]. This can be achieved by choosing words that are less confusable [65] or that are robust to speaker effects such as disfluency or prosodic variation [119]. This could easily be applied in the case of speech output systems.

6.4.1. Vowel space changes

[209] used conventional model adaptation methods from HMM-based speech synthesis to modify vowels, making them either more (hypo-articulated) or less (hyper-articulated) close to the neutral vowel. Intelligibility (only measured objectively using SII and not in a listening test) was successfully improved for the hyper-articulated case.

6.5. Decreasing cognitive effort

6.5.1. Durational changes

As with f_0 modifications, many algorithms are available for modifying speech rate and pause durations, globally or at a finer level [210, 276, 152]. However, the evidence for possible benefits is mixed, as reviewed earlier (section 5.2.4). One system [9] that did produce clear benefits – more than 4 dB – in the presence of a fluctuating masker used durational expansion to shift salient speech information (defined by the Cochlear-scaled Spectral Entropy metric; [275]) in time to avoid epochs of more intense noise. However, the same duration changes did not produce gains when mixed with a stationary masker, suggesting that the noise avoidance rather than slower speech rate was responsible for the benefits.

6.5.2. Pause insertion

[287] explored the insertion of pauses at word boundaries in order to avoid intense masker epochs, under a constant-duration constraint which meant that speech rate was increased to accommodate pause insertion. However, this led to a significant reduction in intelligibility, probably due to a reduction in predictability of word boundaries in noise.

6.5.3. Message simplification

Natural language generation systems can adjust the length and complexity of the sentences they produce, to target certain listener types. This is a large field and we do not attempt a survey of it here. As two illustrative examples, [141] manipulated the complexity of referring expressions to suit different end users and [251] showed that more complex sentences (containing more information) led to better task completion.

6.5.4. Emphasise information-bearing elements

One study that did take attempt to enhance linguistically-relevant information was in the system described by [225], where the simple modifications to f_0, amplitude and duration performed were applied only to content words, thus emphasising the main information bearing elements of the sentence. No detailed evaluation was provided to demonstrate whether this is effective although a 7% intelligibility benefit is mentioned.

7. Summary

Table 1 catalogues the 46 speech modifications which have emerged during the course of this review, identifying in each case whether it has been observed in a talker's speech production, whether it forms the basis for a speech modification algorithms, and finally if – to the best of our knowledge – it is beneficial to listeners.

One striking feature evident from this catalogue is lack of behavioural studies on possible benefits of certain types of modification, especially those at the levels of cognitive effort. Cognitive load is known to affect both speech perception [194, 196] – where, for instance, it changes the balance between the roles on acoustic and lexical factors in word segmentation – and in production [e.g., 85, 83]. See [195] for a recent review of the effect of adverse conditions, including cognitive load, on speech perception.

The absence of speech modification algorithms aimed at increasing coherence has already been noted. In fact, examination of the auditory scene analysis literature suggests many possible avenues for making speech more robust in the presence of competing sources. For instance, contrasting mean intensity between a target and background speech signal might help a listener's scene analysis task, even when this places the target speech at a negative SNR [41]. Similarly, there have been many studies of the benefit of f_0 differences in speech-on-speech mixtures (see [27, 7, 277]) and of the value of spatial separation between target and masker (see [40, 127]). Note that in some of these cases good estimates of masker properties will be required.

A further task for future studies is to incorporate higher levels of linguistic information into modified speech, something which requires access to the intended message and is therefore limited in practice to text-to-speech systems. Obvious approaches in dialog systems include the use of repetition, filled pauses and back-channels as well as the choice of words known to resist masking.

The speech modification approaches listed in table 1 focus on modifying speech to promote its intelligibility or to decrease the cognitive effort required to process it. However, speech modifications could, in principle, be used to reduce the detrimental effect of a masker. This might be achieved by reallocating excess speech energy in time and frequency in order to 'mask the masker', focusing on masker transients or, in the case of speech maskers, those epochs estimated to convey salient information.

It is tempting to use the proposals listed in table 1 as a menu from which arbitrary combinations can be selected. Indeed, positive results might be expected from the combination of spectral and temporal approaches, with additional benefits perhaps deriving from modifications targeted at the message level. However, further studies are needed

Table 1: Speech modifications

Modification	talkers	algorithms	beneficial	Notes
Promoting audibility				
Increase intensity	✓	✓	✓	
Decrease spectral tilt	✓	✓	✓	
Shift spectral centre of gravity	✓	✓	✓	
Narrow formant bandwidth	✓	✓	✓	
Sparsify energy		✓	✓	but not for f_0 [52]
Dynamic amplitude compression		✓	✓	
Increase f_0	✓	✓	✗	
Decrease f_0		✓	✗	
Boost transients		✓		
Change relative segment intensity	✓	✓		
Change relative segment duration	✓	✓		
Change pause duration		✓		
Steady-state suppression		✓	✗	can be harmful [59]
Spectral profile modification	✓	✓	✓	particularly for mid-freqs
Temporal energy reallocation	✓	✓	✓	
Spectro-temporal energy reallocation		✓	✓	
Increasing coherence				
Modulate amplitude	✓		✗	
Modulate f_0	✓		✗	
Flatten f_0 range	✓		✗	may have helped in [199]
Contrast with masker location			✓	
Contrast with masker f_0			✓	
Contrast with masker intensity			✓	
Contrast with interlocutor intensity			✓	
Change gender			✓	
Enhancing linguistic information				
More prototypical vowels	✓	✓	✓	
More prototypical consonants	✓		✓	
Expand vowel space	✓	✓	✓	
Enhance prosodic cues	✓			
Enhance word segmentation cues	✓			
Enhance discourse-level segments	✓			
Decrease neighbourhood density		✓		
Insert lexical/phrasal repetition	✓			for IDS
Decreasing cognitive effort				
Decrease speech rate	✓	✓	✗	
Insert pauses	✓	✓	✓	but not always [299, 287]
Insert additional disfluencies	✓		✗	
Simplify syntactic structure	✓	✓		
Emphasise new/information-bearing elements	✓	✓	✓	1st occurrence more intelligible [109]
Decrease word complexity	✓			
Increase word predictability			✓	e.g., [253, 307]
Decrease vocabulary size	✓			
Decrease sentence length	✓			for IDS
Add reference to previous information	✓			for IDS
Repeat interlocutor speech	✓			
Increase redundancy	✓		✓	e.g., visual cues
Provide backchannel feedback	✓			
Align with interlocutor	✓		✓	e.g., intensity, f_0, speech rate, lexical use, accent

on possible antagonistic effects of combining multiple modification methods. For example, while dynamic amplitude compression has proved to be a successful technique, applying it alongside other temporal energy reallocation methods which operate by unmasking weaker signal epochs may be counter-productive. Likewise, combinations that are not physiologically-coherent, such as increasing vocal effort and decreasing F1, might have negative effects on intelligibility as they may sound unnatural to a listener and consequently lead to attentional disturbance.

Finally, we note that since in many listening scenarios adequate intelligibility is maintained simply by increasing output level, the purpose of speech modification is often considered to reside in using any gain in dB to resist this tactic, i.e., 'turning down the volume'. However, the headroom gained by the modification can be spent in other ways. For example, instead of reducing output level, information rate might be increased by speeding up speech, or the need for repetition (in the case of public address systems) might be reduced.

Acknowledgements. The authors thank the EU Future and Emerging Technology (FET-OPEN) project the 'Listening Talker' for supporting the ideas which led to the preparation of this manuscript.

References

[1] Adriaans, F., & Swingley, D. (2012). Distributional learning of vowel categories is supported by prosody in infant-directed speech. In *CogSci* (pp. 72–77). Sapporo, Japan.
[2] Albin, D. D., & Echols, C. H. (1996). Stressed and word-final syllables in infant-directed speech. *Infant Behav. Dev.*, *19*, 401–418.
[3] Alcántara, J. I., Dooley, G. J., Blamey, P. J., & Seligman, P. M. (1994). Preliminary evaluation of a formant enhancement algorithm on the perception of speech in noise for normally hearing listeners. *Audiology*, *33*, 15–27.
[4] Andruski, J. E., & Kuhl, P. K. (1996). The acoustic structure of vowels in mothers' speech to infants and adults. In *ICSLP* (pp. 1545–1548).
[5] ANSI S3.5-1997 (1997). Methods for the calculation of the Speech Intelligibility Index.
[6] Arai, T., Kinoshita, K., Hodoshima, N., Kusumoto, A., & Kitamura, T. (2002). Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments. *Acoustical Science and Technology*, *23*, 229–232.
[7] Assmann, P. F. (1999). Fundamental frequency and the intelligibility of competing voices. In *ICPhS*.
[8] Assmann, P. F., Nearey, T. M., & Scott, J. M. (2002). Modeling the perception of frequency-shifted vowels. In *ICSLP* (pp. 425–428).
[9] Aubanel, V., & Cooke, M. (2013). Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. In *Interspeech* (p. submitted). Lyon, France.
[10] Aubanel, V., & Cooke, M. (under review). Strategies adopted by talkers faced with fluctuating and competing speech maskers. *J. Acoust. Soc. Am.*, .
[11] Aubanel, V., & Nguyen, N. (2010). Automatic recognition of regional phonological variation in conversational interaction. *Speech Comm.*, *52*, 577–586.
[12] Babel, M. (2011). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *J. Phon.*, *40*, 177–189.
[13] Babel, M., & Bulatov, D. (2012). The role of fundamental frequency in phonetic accommodation. *Lang. Speech*, *55*, 231–248.
[14] Bailly, G., & Lelong, A. (2010). Speech dominoes and phonetic convergence. In *Interspeech* (pp. 1153–1156). Makuhari, Japan.
[15] Baran, J. A., Laufer, M. Z., & Daniloff, R. (1977). Phonological contrastivity in conversation: A comparative study of voice onset time. *J. Phon.*, *5*, 339–350.
[16] Barker, J., & Cooke, M. (2007). Modelling speaker intelligibility in noise. *Speech Comm.*, *49*, 402–417.
[17] Batliner, A., Schuller, B., Schaeffler, S., & Steidl, S. (2008). Mothers, adults, children, pets — towards the acoustics of intimacy. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE.* (pp. 4497–4500).
[18] Beautemps, D., Borel, P., & Manolios, S. (1999). Hyper-articulated speech: Auditory and visual intelligibility. In *Eurospeech* (pp. 109–112).
[19] Beckford Wassink, A., Wright, R. A., & Franklin, A. D. (2007). Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. *J. Phon.*, *35*, 363–379.
[20] Beckman, M. E., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Ed.), *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III* (pp. 7–33). Cambridge University Press.
[21] Beckman, M. E., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. Docherty, & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 68–86). Cambridge University Press.
[22] Bele, I. V. (2006). The speaker's formant. *J. Voice*, *20*, 555–578.
[23] Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *J. Mem. Lang.*, *60*, 92–111.
[24] Benoit, C., Fuster-Duran, A., & Le Goff, B. (1996). An investigation of hypo- and hyper-speech in the visual modality. In *1st ETRW on speech production modeling* (pp. 237–240). Autrans.

[25] Best, V., Ozmeral, E., Kopko, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proc. Nat. Acad. Sci.*, *105*, 13173–13177.
[26] Biersack, S., Kempe, V., & Knapton, L. (2005). Fine-tuning speech registers: a comparison of the prosodic features of child-directed and foreigner-directed speech. In *European Conference on Speech Communication and Technology* (pp. 2401–2405).
[27] Bird, J., & Darwin, C. J. (1998). Effects of a difference in fundamental frequency in separating two sentences. In A. R. Palmer, A. Rees, A. Q. Summerfield, & R. Meddis (Eds.), *Psychophysical and Physiological Advances in Hearing* (pp. 263–269). London: Whurr.
[28] Blamey, P., Dooley, G., Alcantara, J., Gerin, E., & Seligman, P. (1993). Formant-based processing for hearing aids. *Speech Communication*, *13*, 453 – 461.
[29] Boike, K. T., & Souza, P. E. (2000). Effect of compression ratio on speech recognition and speech-quality ratings with wide dynamic range compression amplification. *J. Speech Lang. Hear. R.*, *43*, 456–468.
[30] Bond, Z. S., & Moore, T. J. (1990). A note on loud and Lombard speech. In *ICSLP* (pp. 969–972).
[31] Bond, Z. S., & Moore, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Comm.*, *14*, 325–337.
[32] Bond, Z. S., Moore, T. J., & Gable, B. (1989). Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask. *J. Acoust. Soc. Am.*, *85*, 907–912.
[33] Boril, H., & Pollak, P. (2005). Design and collection of Czech Lombard database. In *ICSLP* (pp. 1577–1580). Lisbon, Portugal.
[34] Bradlow, A. R. (2002). Confluent talker-and listener-oriented forces in clear speech production. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 241–274). Berlin: Mouton de Gruyter.
[35] Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *J. Acoust. Soc. Am.*, *112*, 272–284.
[36] Bradlow, A. R., Kraus, N., & Hayes, E. (2003). Speaking clearly for children with learning disabilities: Sentence perception in noise. *J. Speech Lang. Hear. R.*, *46*, 80–97.
[37] Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Comm.*, *20*, 255–272.
[38] Branigan, H., Pickering, M., & Cleland, A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*, B13–B25.
[39] Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.
[40] Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J. Acoust. Soc. Am.*, *83*, 1508–1516.
[41] Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.*, *109*, 1101–1109.
[42] Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers . *J. Acoust. Soc. Am.*, *100*, 2527–2538.
[43] Brunskog, J., Gade, A. C., Bellester, G. P., & Calbo, L. R. (2009). Increase in voice level and speaker comfort in lecture rooms. *J. Acoust. Soc. Am.*, *125*, 2072–2082.
[44] Burnham, D., Joeffry, S., & Rice, L. (2010). Computer-and human-directed speech before and after correction. In *SST* (pp. 13–17). Melbourne, Australia.
[45] Burnham, D., Joeffry, S., & Rice, L. (2010). "Does-Not-Compute": Vowel hyperarticulation in speech to an auditory-visual avatar. In *AVSP*. Japan.
[46] Burnham, D., Kim, J., Davis, C., Ciocca, V., Schoknecht, C., Kasisopa, B., & Luksaneeyanawin, S. (2011). Are tones phones? *J. Exp. Child Psychol.*, *108*, 693–712.
[47] Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*, 1435.
[48] Castellanos, A., Benedi, J. M., & Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Comm.*, *20*, 23–35.
[49] Chen, F. R. (1980). *Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level*. Ph.D. thesis MIT.
[50] Cheskin, A. (1981). The verbal environment provided by hearing mothers for their young deaf children. *J. Comm. Disord.*, *14*, 485–496.
[51] Cheyne, H. A., Kalgaonkar, K., Clements, M., & Zurek, P. (2009). Talker-to-listener distance effects on speech production and perception. *J. Acoust. Soc. Am.*, *126*, 2052–2060.
[52] Chládková, K., Boersma, P., & Podlipský, V. (2009). On-line formant shifting as a function of F0. In *Interspeech* (pp. 464–467).
[53] Chung, V., Mirante, N., Otten, J., & Vatikiotis-Bateson, E. (2005). Audiovisual processing of Lombard speech. In *AVSP* (pp. 55–56).
[54] Church, R., Bernhardt, B., Shi, R., & Pichora Fuller, K. (2005). Infantdirected speech: Final syllable lengthening and rate of speech. *J. Acoust. Soc. Am.*, *117*, 2429–2430.
[55] Cole, R. A., Yan, Y., Mak, B., Fanty, M., & Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. In *ICASSP* (pp. 853–856). IEEE.
[56] Cooke, M. (2006). A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, *119*, 1562–1573.
[57] Cooke, M., & García Lecumberri, M. L. (2012). The intelligibility of Lombard speech for non-native listeners. *J. Acoust. Soc. Am.*, *132*, 1120–1129.
[58] Cooke, M., & Lu, Y. (2010). Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. *J. Acoust. Soc. Am.*, *128*, 2059–2069.
[59] Cooke, M., Mayo, C., & Valentini-Botinhao, C. (submitted). Intelligibility-enhancing speech modifications: the Hurricane Challenge. In *Interspeech 2013*.
[60] Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., & Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Comm.*, *55*, 572–585.
[61] Cooke, M., Mayo, C., & Villegas, J. (submitted). The role of durational increases in the lombard speech intelligibility benefit. *J. Acoust. Soc. Am.*, .
[62] Cooper, R. P., Abraham, J., Berman, S., & Staska, M. (1997). The development of infants' preference for motherese. *Infant Behav. Dev.*, *20*, 477–488.

[63] Cooper, R. P., & Aslin, R. N. (1994). Developmental differences in infant attention to the spectral properties of infantdirected speech. *Child Dev.*, *65*, 1663–1677.
[64] Cox, R. M., Alexander, G. C., & Gilmore, C. (1987). Intelligibility of average talkers in typical listening environments. *J. Acoust. Soc. Am.*, *81*, 1598–1608.
[65] Cox, S., & Vinagre, L. (2004). Modelling of confusions in aircraft call-signs. *Speech Comm.*, (pp. 289–312).
[66] Cristià, A. (2010). Phonetic enhancement of sibilants in infant-directed speech. *J. Acoust. Soc. Am.*, *128*, 424–434.
[67] Cristia, A., & Seidl, A. (2013). The hyperarticulation hypothesis of infant-directed speech. *J. Child Lang.*, (pp. 1–22).
[68] Cutler, A., & Butterfield, S. (1990). Durational cues to word boundaries in clear speech. *Speech Comm.*, *9*, 485–495.
[69] Darwin, C. J. (2008). Listening to speech in the presence of other sounds. *Phil. Trans. R. Soc. B*, *363*, 1011–1021.
[70] Dau, T., Püschel, D., & Kohlrausch, A. (1996). A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am.*, *99*, 3615–3622.
[71] Davis, C., Kim, J., Grauwinkel, K., & Mixdorff, H. (2006). Lombard speech: Auditory (A), Visual (V) and AV effects. In *Speech Prosody* (pp. 248–252). Dresden, Germany.
[72] Davis, H., Hudgins, C. v., Marquis, R. J., Nichols, R. H. J., Peterson, G. E., Ross, D. A., & Stevens, S. S. (1946). The selection of hearing aids. *The Laryngoscope*, *3*, 85–115.
[73] Dejonckere, P. H., & Pepin, F. (1983). Etude de l'effet Lombard par la mesure du niveau sonore équivalent. *Folia Phoniatrica*, *35*, 310–315.
[74] Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, *64*, 145–173.
[75] Dodane, C., Blanc, J. M., & Dominey, P. F. (2006). Différenciation des mots de fonction et des mots de contenu par la prosodie: analyse d'un corpus trilingue de langue adressée à l'enfant et à l'adulte. In *Journées d'Etude sur la Parole* (pp. 255–258). Dinard, France.
[76] Dreher, J. J., & O'Neill, J. J. (1957). Effects of ambient noise on speaker intelligibility for words and phrases. *J. Acoust. Soc. Am.*, *29*, 1320–1323.
[77] Dudley, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America*, *11*, 169–177.
[78] Dunst, C. J., Gorman, E., & Hamby, D. W. (2012). Child-directed motionese with infants and toddlers with and without hearing impairments. *Center for Early Literacy Learning reviews*, *5*, 1–11.
[79] Egan, J. J. (1972). Psychoacoustics of the Lombard voice response. *Journal of Auditory Research*, *12*, 318–324.
[80] Englund, K. T. (2005). Voice onset time in infant directed speech over the first six months. *First language*, *25*, 219–234.
[81] Eriksson, A., & Traunmuller, H. (2002). Perception of vocal effort and distance from the speaker on the basis of vowel utterances. *Percept. Psychophys.*, *64*, 131–139.
[82] Erting, C., Prezioso, C., & O'Grady Hynes, M. (1990). The interactional context of deaf mother-infant communication. In V. Volterra, & C. Erting (Eds.), *From gesture to language in hearing and deaf children* (pp. 97–106). Washington, DC: Gallaudet University Press.
[83] Estival, D., & Molesworth, B. (2009). A study of EL2 pilots' radio communication in the general aviation environment. *Australian Review of Applied Linguistics*, *32*.
[84] Falk, S. (2011). Melodic versus intonational coding of communicative functions: A comparison of tonal contours in infant-directed song and speech. *Psychomusicology: Music, Mind and Brain*, *21*, 54.
[85] Farris, C., Trofimovich, P., Segalowitz, N., & Gatbonton, E. (2008). Air traffic communication in a second language: Implications of cognitive factors for training and assessment. *TESOL Quart.*, *42*, 397–410.
[86] Ferguson, S. H. (2004). Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners. *J. Acoust. Soc. Am.*, *116*, 2365–2373.
[87] Ferguson, S. H., & Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.*, *112*, 259–271.
[88] Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behav. Dev.*, *8*, 181–195.
[89] Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Dev.*, *60*, 1497–1510.
[90] Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Infant Behav. Dev.*, *10*, 279–293.
[91] Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Dev. Psychol.*, *20*, 104.
[92] Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *J. Child Lang.*, *16*, 477–501.
[93] Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quart.*, *39*, 399–423.
[94] Fisher, C., & Tokura, H. (1995). The given-new contract in speech to infants. *J. Mem. Lang.*, *34*, 287–310.
[95] Fisher, C., & Tokura, H. (1996). Acoustic cues to grammatical structure in infantdirected speech: crosslinguistic evidence. *Child Dev.*, *67*, 3192–3218.
[96] Fitzpatrick, M., Kim, J., & Davis, C. (2011). The effect of seeing the interlocutor on auditory and visual speech production in noise. *AVSP*, (pp. 31–35).
[97] Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *J. Exp. Psychol. Human.*, *32*, 1276–1293.
[98] Fogerty, D., & Kewley-Port, D. (2009). Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *J. Acoust. Soc. Am.*, *126*, 847–857.
[99] Fogerty, D., Kewley-Port, D., & Humes, L. E. (2012). The relative importance of consonant and vowel segments to the recognition of words and sentences: Effects of age and hearing loss. *J. Acoust. Soc. Am.*, *132*, 1667–1678.
[100] Fowler, C. A. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Lang. Speech*, *31*, 307–319.
[101] Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *J. Mem. Lang.*, *26*, 489–504.
[102] Freed, B. F. (1981). Foreigner talk, baby talk, native talk. *Int'l. J. Soc. Lang.*, *1981*, 19–40.

[103] Fux, T., Aubergé, V., Feng, G., & Zimpfer, V. (2012). Speaker's prosodic strategy for a large physical distance communication task. In *International Conference on Speech and Corpora*. Belo Horizonte.
[104] Fux, T., Feng, G., & Zimpfer, V. (2011). Relevant acoustic features of speech signals for natural-to-shouted voice transformation. In *Forum Acusticum*. Aalborg, Danemark.
[105] Fux, T., Feng, G., & Zimpfer, V. (2012). Natural-to-shouted voice transformation for distance cues of monosyllabic consonant-vowel-consonant words. *Acta Acust. United Ac.*, *98*, 839–843.
[106] Gagne, J. P. (1995). Auditory, visual, and audiovisual speech intelligibility for sentence-length stimuli: An investigation of conversational and clear speech. *Volta Review*, *97*, 33–51.
[107] Gagne, J. P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal-Academy Of Rehabilitative Audiology*, *27*, 135–158.
[108] Gagne, J. P., Rochette, A. J., & Charest, M. (2002). Auditory, visual and audiovisual clear speech. *Speech Comm.*, *37*, 213–230.
[109] Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *J. Mem. Lang.*, *62*, 35–51.
[110] Garnier, M. (2007). *Communiquer en environnement bruyant : de l'adaptation jusqu'au forçage vocal* . Ph.D. thesis Université Paris 6 Paris.
[111] Garnier, M. (2008). May speech modifications in noise contribute to enhance audio-visible cues to segment perception? In *AVSP*. Moreton Island, Australia.
[112] Garnier, M., Bailly, L., Dohen, M., Welby, P., & Lœvenbruck, H. (2006). An acoustic and articulatory study of Lombard speech: global effects on the utterance. In *Interspeech* (pp. 17–22).
[113] Garnier, M., Dohen, M., Loevenbruck, H., Welby, P., & Bailly, L. (2006). The Lombard effect: A physiological reflex or a controlled intelligibility enhancement? In *7th International Seminar on Speech Production* (pp. 255–262). Ubatuba, Brazil.
[114] Garnier, M., & Henrich, N. (). Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise. *Comput. Speech Lang.*, .
[115] Garnier, M., Henrich, N., & Dubois, D. (2010). Influence of sound immersion and communicative interaction on the Lombard effect. *J. Speech Lang. Hear. R.*, *53*, 588–608.
[116] Garnier, M., Ménard, L., & Richard, G. (2012). Effect of being seen on the production of visible speech cues. A pilot study on Lombard speech. In *Interspeech*. Portland, US.
[117] Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*, 181–218.
[118] Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: developments in applied sociolinguistics* (pp. 1–68). Cambridge University Press.
[119] Goldwater, S., Jurafsky, D., & Manning, C. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Comm.*, *52*, 181–200.
[120] Golinkoff, R. M., & Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing Chinese: Implications for language acquisition. *J. Child Lang.*, *22*, 703–726.
[121] Golinkoff, R. M., & Ames, G. J. (1979). A comparison of fathers' and mothers' speech with their young children. *Child Dev.*, (pp. 28–32).
[122] Gordon-Salant, S. (1986). Recognition of natural and time/intensity altered CVs by young and elderly subjects with normal hearing. *J. Acoust. Soc. Am.*, *80*, 1599–1607.
[123] Granlund, S., Baker, R., & Hazan, V. (2011). Acoustic-phonetic characteristics of clear speech in bilinguals. In *ICPhS* (pp. 763–766). Hong Kong.
[124] Granlund, S., Hazan, V., & Baker, R. (2012). An acoustic–phonetic comparison of the clear speaking styles of Finnish–English late bilinguals. *J. Phon.*, *40*, 509–520.
[125] Green, J. R., Nip, I. S. B., Wilson, E. M., Mefferd, A. S., & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *J. Speech Lang. Hear. R.*, *53*, 1529.
[126] Grieser, D. A. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Dev. Psychol.*, *24*, 14.
[127] Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *J. Acoust. Soc. Am.*, *115*, 833–843.
[128] Hazan, V., & Baker, R. (2011). Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *J. Acoust. Soc. Am.*, *130*, 2139–2152.
[129] Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *J. Acoust. Soc. Am.*, *116*, 3108–3118.
[130] Hazan, V., & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: speaker and listener effects. *Lang. Speech*, *43*, 273–294.
[131] Hazan, V. L., Grynpas, J., & Baker, R. (2012). Is clear speech tailored to counter the effect of specific adverse listening conditions? *JASA Express Letters*, *132*, EL371–EL377.
[132] Healey, P. G. T., Howes, C., & Purver, M. (2010). Does structural priming occur in ordinary conversation. In *Linguistic Evidence* (pp. 1–4).
[133] Helfer, K. S. (1997). Auditory and auditory-visual perception of clear and conversational speech. *J. Speech Lang. Hear. R.*, *40*, 432–443.
[134] Helfer, K. S. (1998). Auditory and auditory-visual recognition of clear and conversational speech by older adults. *J. Am. Acad. Audiol.*, *9*, 234–242.
[135] Hornsby, B. W. Y., & Ricketts, T. A. (2001). The effects of compression ratio, signal-to-noise ratio, and level on speech recognition in normal-hearing listeners. *J. Acoust. Soc. Am.*, *109*, 2964–2973.
[136] Horwitz, A. R., Turner, C. W., & Fabry, D. A. (1991). Effects of different frequency response strategies upon recognition and preference for audible speech stimuli. *J Speech Hear Res*, *34*, 1185–1196.
[137] Howell, P., & Bonnett, C. (1997). Speaking clearly for the hearing impaired: Intelligibility differences between clear and less clear speakers.

Eur J Disord Commun, *32*, 89–97.
[138] Imaizumi, S., Hamaguchi, S., & Deguchi, T. (1993). Vowel devoicing in teachers' speech directed to the hearing-impaired/normal-hearing children: Do teachers avoid devoicing to help hearing-impaired understand dialogue? *Communication Disorder Research*, *22*, 7–20.
[139] Imaizumi, S., Hayashi, A., & Deguchi, T. (1993). Planning in speech production: Listener adaptive characteristics. *Japan Journal of Logopedics and Phoniatrics*, *34*, 394–401.
[140] Ishizuka, K., & Aikawa, K. (2002). Effect of F0 fluctuation and amplitude modulation of natural vowels on vowel identification in noisy environments. In *ICSLP* (pp. 1633–1636). Denver, USA.
[141] Janarthanam, S., & Lemon, O. (2010). Learning to adapt to unknown users: referring expression generation in spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* ACL '10 (pp. 69–78). Morristown, NJ, USA: Association for Computational Linguistics.
[142] Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: phonetic targets are hyperarticulated. *Language*, *69*, 505–528.
[143] Jun, S. A., & Fougeron, C. (2000). A phonological model of French intonation. In A. Botinis (Ed.), *Intonation: Analysis, Modeling and Technology* (pp. 209–242). Dordrecht: Kluwer Academic Publishers.
[144] Jung, O. (2012). On the Lombard effect induced by vehicle interior driving noises, regarding sound pressure level and long-term average speech spectrum. *Acta Acust. United Ac.*, *98*, 334–341.
[145] Junqua, J. C. (1992). Acoustic and production pilot studies of speech vowels produced in noise. In *ICSLP* (pp. 811–814). Banff, Canada.
[146] Junqua, J. C. (1993). The Lombard reflex and its role on human listener and automatic speech recognizers. *J. Acoust. Soc. Am.*, *93*, 510–524.
[147] Junqua, J.-C., Fincke, S., & Field, K. (1998). Influence of the speaking style and the noise spectral tilt on the lombard reflex and automatic speech recognition. In *ICSLP* (pp. 467–470).
[148] Kadiri, N. (1998). *Conséquences d'un environnement bruité sur la production de la parole*. Ph.D. thesis Université Paul Sabatier Toulouse, France.
[149] Kang, K. H., & Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *J. Acoust. Soc. Am.*, *124*, 3909–3917.
[150] Kaplan, P. S., Jung, P.C., Ryther, J. S., & Zarlengo-Strouse, P. (1996). Infant-directed versus adult-directed speech as signals for faces. *Dev. Psychol.*, *32*, 880.
[151] Kappes, J., Baumgaertner, A., Peschke, C., & Ziegler, W. (2009). Unintended imitation in nonword repetition. *Brain Lang.*, *111*, 140–151.
[152] Kawahara, H., Masuda-Katsuse, I., & de Cheveign, A. (1999). Restructuring speech representations using a pitch-adaptive time¢frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, *27*, 187 – 207.
[153] Kewley-Port, D., Burkle, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *J. Acoust. Soc. Am.*, *122*, 2365–2375.
[154] Kim, J., Davis, C., Vignali, G., & Hill, H. (2005). A visual concomitant of the Lombard reflex. In *AVSP* (pp. 17–21). Vancouver, Canada.
[155] Kim, S. (2005). Durational characteristics of Korean Lombard speech. In *Interspeech* (pp. 2901–2904). Lisbon, Portugal.
[156] Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. In *J. Acoust. Soc. Am.* (pp. 2238–2246).
[157] Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy*, *4*, 85–110.
[158] Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2001). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behav. Dev.*, *24*, 372–392.
[159] Kondaurova, M. V., & Bergeson, T. R. (2011). The effects of age and infant hearing status on maternal use of prosodic cues for clause boundaries in speech. *J. Speech Lang. Hear. R.*, *54*, 740.
[160] Kondaurova, M. V., Bergeson, T. R., & Dilley, L. C. (2012). Effects of deafness on acoustic characteristics of American English tense/lax vowels in maternal speech to infants. *J. Acoust. Soc. Am.*, *132*, 1039–1049.
[161] Korn (1954). Effect of psychological feedback on conversationnal noise reduction in rooms. *J. Acoust. Soc. Am.*, *26*, 793–794.
[162] Krause, J. C. (2001). *Properties of naturally produced clear speech at normal rates and implications for intelligibility enhancement*. Ph.D. thesis MIT Cambridge, MA.
[163] Krause, J. C., & Braida, L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *J. Acoust. Soc. Am.*, *112*, 2165–2172.
[164] Krause, J. C., & Braida, L. D. (2003). Effects of listening environment on intelligibility of clear speech at normal speaking rates. *Iran. Audiol*, *2*, 39–47.
[165] Krause, J. C., & Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Am.*, *115*, 362–378.
[166] Krause, J. C., & Braida, L. D. (2009). Evaluating the role of spectral and envelope characteristics in the intelligibility advantage of clear speech. *J. Acoust. Soc. Am.*, *125(5)*, 3346–3357.
[167] Kretsinger, E. A., & Young, N. B. (1960). The use of fast limiting to improve the intelligibility of speech in noise. *Speech Monographs*, *27*, 63–69.
[168] Kryter, K. D. (1946). Effect of ear protective devices on the intelligibility of speech in noise. *J. Acoust. Soc. Am.*, *18*, 413–417.
[169] Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*, 684–686.
[170] Kusumoto, A., Arai, T., Kinoshita, K., Hodoshima, N., & Vaughan, N. (2005). Modulation enhancement of speech by a pre-processing algorithm for improving intelligibility in reverberant environments. *Speech Comm.*, *45*, 101–113.
[171] Lam, C., & Kitamura, C. (2012). Mommy, speak clearly: induced hearing loss shapes vowel hyperarticulation. *Developmental Sci.*, *15*, 212–221.
[172] Lane, H. L., Tranel, B., & Sisson, C. (1970). Regulation of voice communication by sensory dynamics. *J. Acoust. Soc. Am.*, *47*, 618–624.
[173] Laures, J., & Bunton, K. (2003). Perceptual effects of a flattened fundamental frequency at the sentence level under different listening

conditions. *J. Comm. Disord.*, *36*, 449–464.
[174] Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *J. Speech Lang. Hear. R.*, *42*, 1148–1156.
[175] Lee, H. W., Rayner, K., & Pollatsek, A. (2001). The relative contribution of consonants and vowels to word identification during reading. *J. Mem. Lang.*, *44*, 189–205.
[176] Lienard, J. S., & Di Benedetto, M. G. (1999). Effect of vocal effort on spectral properties of vowels. *J. Acoust. Soc. Am.*, *106*, 411–422.
[177] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Dordrecht, The Netherlands: Kluwer Academic Publishers.
[178] Lindblom, B., Brownlee, S., Davis, B., & Moon, S. J. (1992). Speech transforms. *Speech Comm.*, *11*, 357–368.
[179] Little, H. R. (2011). *Foreigner Directed Speech*. Ph.D. thesis The University of Edinburgh.
[180] Liu, H. M., Kuhl, P. K., & Tsao, F. M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Sci.*, *6*, F1–F10.
[181] Liu, H. M., Tsao, F. M., & Kuhl, P. K. (2007). Acoustic analysis of lexical tone in Mandarin infant-directed speech. *Dev. Psychol.*, *43*, 912.
[182] Liu, H. M., Tsao, F. M., & Kuhl, P. K. (2009). Age-related changes in acoustic modifications of Mandarin maternal speech to preverbal infants and five-year-old children: a longitudinal study. *J. Child Lang.*, *36*, 909–922.
[183] Lombard, E. (1911). Le signe d'élévation de la voix [The sign of the elevation of the voice]. *Annales des maladies de l'oreille et du larynx*, *37*, 101–119.
[184] Long, M. H. (1981). Input, interaction, and second-language acquisition. *Annals of the New York Academy of Sciences*, *379*, 259–278.
[185] Long, M. H. (1983). Linguistic and conversational adjustments to non-native speakers. *Studies in second language acquisition*, *5*, 177–193.
[186] Lu, Y., & Cooke, M. (2008). Speech production modifications produced by competing talkers, babble, and stationary noise. *J. Acoust. Soc. Am.*, *124*, 3261–3275.
[187] Lu, Y., & Cooke, M. (2009). Speech production modifications produced in the presence of low-pass and high-pass filtered noise. *J. Acoust. Soc. Am.*, *126*, 1495–1499.
[188] Lu, Y., & Cooke, M. (2009). The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise. *Speech Comm.*, *51*, 1253–1262.
[189] Malsheen, B. J. (1980). Two hypotheses for phonetic clarification in the speech of mothers to children. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology (Vol. 2). Perception* (pp. 173–184). Academic Press.
[190] Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *J. Acoust. Soc. Am.*, *125*, 3962–3973.
[191] Marin, C. M. H., & McAdams, S. (1991). Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width. *J. Acoust. Soc. Am.*, *89*, 341–351.
[192] Masataka, N. (1992). Motherese in a signed language. *Infant Behav. Dev.*, *15*, 453–460.
[193] Matthies, M., Perrier, P., Perkell, J. S., & Zandipour, M. (2001). Variation in anticipatory coarticulation with changes in clarity and rate. *J. Speech Lang. Hear. R.*, *44*, 340.
[194] Mattys, S., Brooks, J., & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology*, *59*, 203–243.
[195] Mattys, S., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Lang. Cognitive Proc.*, *27*, 953–978.
[196] Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *J. Memory and Language*, *65*, 145–160.
[197] Maye, J., Weiss, D. J., & Aslin, R. N. (2007). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Sci.*, *11*, 122–134.
[198] Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.
[199] Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. In *Interspeech* (p. 74). Portland, US.
[200] Metz, D. E., Samar, V. J., Schiavetti, N., Sitler, R. W., & Whitehead, R. L. (1985). Acoustic dimensions of hearing-impaired speakers' intelligibility. *J. Speech Lang. Hear. R.*, *28*, 345.
[201] Michael, D. D., Siegel, G. M., & Pick, H. L., Jr (1995). Effects of distance on vocal intensity. *J. Speech Lang. Hear. R.*, *38*, 1176.
[202] Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2010). Alignment to visual speech information. *Atten. Percept. Psycho.*, *72*, 1614–1625.
[203] Mokbel, C. (1992). *Reconnaissance de la parole dans le bruit: bruitage / débruitage*. Ph.D. thesis Ecole nationale supérieure des télécommunications Paris.
[204] Monsen, R. B. (1978). Toward measuring how well hearing-impaired children speak. *J. Speech Lang. Hear. R.*, *21*, 197.
[205] Montgomery, A. A., & Edge, R. A. (1988). Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults. *J. Speech Hear. Res.*, *31*, 386–393.
[206] Moon, S. J., & Lindblom, B. (1989). Formant undershoot in clear and citation-form speech: A second progress report. *STL-QPSR*, *30*, 121–123.
[207] Moon, S. J., & Lindblom, B. (1994). Interaction between duration, context, and speaking style in English stressed vowels. *J. Acoust. Soc. Am.*, *96*, 40–55.
[208] Moore, B. C. J. (2003). Temporal integration and context effects in hearing. *J. Phonetics*, *31*, 563–574.
[209] Moore, R. K., & Nicolao, M. (2011). Reactive speech synthesis: actively managing phonetic contrast along an H&H continuum. In *17th Int. Cong. Phonetic Sciences* (pp. 1422–1425).
[210] Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453 – 467.
[211] Naoi, N., Minagawa-Kawai, Y., Kobayashi, A., Takeuchi, K., Nakamura, K., Yamamoto, J., & Kojima, S. (2012). Cerebral responses to infant-directed speech and the effect of talker familiarity. *NeuroImage*, *59*, 1735–1744.
[212] Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *J. Pers. Soc. Psychol.*,

32, 790–804.

[213] Nejime, Y., & Moore, B. C. J. (1998). Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. *J. Acoust. Soc. Am.*, *103*, 572–576.

[214] Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: some effects and non-effects of maternal speech style. In C. E. Snow, & C. A. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 109–149). Cambridge University Press.

[215] Niederjohn, R. J., & Grotelueschen, J. H. (1976). The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression. *IEEE Trans. on Acoustics, Speech and Signal Processing*, *24*, 277–282.

[216] Niwano, K., & Sugai, K. (2002). Age-related change in Japanese maternal infant-directed speech and infant's vocal response. *Research and clinical center for child development Annual Report*, *24*, 13–24.

[217] Niwano, K., & Sugai, K. (2003). Maternal accomodation in infant-directed speech during mother's and twin-infants' vocal interactions. *Psychological reports*, *92*, 481–487.

[218] Ostroff, W. L. (2000). *Non-Linguistic Influences on Infants' Nonnative Phoneme Perception: Exaggerated Prosody and Visual Speech Information Improve Discrimination*. Ph.D. thesis Virginia Polytechnic Institute and State University.

[219] Oviatt, S., Levow, G. A., Moreton, E., & MacEachern, M. (1998). Modeling global and focal hyperarticulation during human-computer error resolution. *J. Acoust. Soc. Am.*, *104*, 3080–3098.

[220] Owren, M. J., & Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *J. Acoust. Soc. Am.*, *119*, 1727–1739.

[221] Papousek, M., & Hwang, S. F. C. (1991). Tone and intonation in Mandarin babytalk to presyllabic infants: Comparison with registers of adult conversation and foreign language instruction. *Applied Psycholinguistics*, *12*, 481–504.

[222] Papousek, M., Papoušek, H., & Bornstein, M. H. (1985). The naturalistic vocal environment of young infants: On the significance of homogeneity and variability in parental speech. *Social perception in infants*, (pp. 269–297).

[223] Papousek, M., Papoušek, H., & Symmes, D. (1991). The meanings of melodies in motherese in tone and stress languages. *Infant Behav. Dev.*, *14*, 415–440.

[224] Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *J. Acoust. Soc. Am.*, *119*, 2382–2393.

[225] Patel, R., Everett, M., & Sadikov, E. (2006). Loudmouth:: modifying text-to-speech synthesis in noise. In *ACM SIGACCESS Conference on Assistive Technologies: Proceedings of the 8 th international ACM SIGACCESS conference on Computers and accessibility* (pp. 227–228). volume 23.

[226] Patel, R., & Schell, K. W. (2008). The influence of linguistic content on the Lombard effect. *J. Speech Lang. Hear. R.*, *51*, 209–220.

[227] Payne, E., Post, B., Astruc, L., Prieto, P., & Vanrell, M. M. (2009). Rhythmic modification in child directed speech. *Oxford University Working Papers in Linguistics, Philology and Phonetics*, *12*, 123–144.

[228] Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *J. Acoust. Soc. Am.*, *95*, 1581–1592.

[229] Pegg, J. E., Werker, J. F., & McLeod, P. J. (1992). Preference for infant-directed over adult-directed speech: Evidence from 7-week-old infants. *Infant Behav. Dev.*, *15*, 325–345.

[230] Pelegrín-García, D., Smits, B., Brunskog, J., & Jeong, C. H. (2011). Vocal effort with changing talker-to-listener distance in different acoustic environments. *J. Acoust. Soc. Am.*, *129*, 1981–1990.

[231] Perkell, J. S., Denny, M., Lane, H., Guenther, F., Matthies, M. L., Tiede, M., Vick, J., Zandipour, M., & Burton, E. (2007). Effects of masking noise on vowel and sibilant contrasts in normal-hearing speakers and postlingually deafened cochlear implant users. *J. Acoust. Soc. Am.*, *121*, 505–518.

[232] Petkov, P., Kleijn, B., & Henter, G. (2012). Enhancing subjective speech intelligibility using a statistical model of speech. In *Proc. Interspeech*. Portland, USA.

[233] Peutz, V. M. A. (1971). Articulation loss of consonants as a criterion for speech transmission in rooms. *J. Audio Eng. Soc.*, *19*, 915–919.

[234] Phillips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. *Child Dev.*, *44*, 182–185.

[235] Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversationnal speech. *J. Speech Hear. Res.*, *28*, 96–103.

[236] Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *J. Speech Hear. Res.*, *29*, 434–446.

[237] Picheny, M. A., Durlach, N. I., & Braida, L. D. (1989). Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. *J. Speech Hear. Res.*, *32*, 600–603.

[238] Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research Language Computation*, *4*, 203–228.

[239] Pickett, J. (1956). Effects of vocal force on the intelligibility of speech sounds. *J. Acoust. Soc. Am.*, *28*, 902–905.

[240] Pickett, J. M., & Pollack, I. (1958). Intelligibility at high voice levels and the use of a megaphone. *J. Acoust. Soc. Am.*, *30*, 1100–1104.

[241] Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C., & Yuchtman, M. (1985). Some acoustic-phonetic correlates of speech produced in noise. In *ICASSP* (pp. 1581–1584). Tampa, Florida.

[242] Pittman, A. L., & Wiley, T. L. (2001). Recognition of speech produced in noise. *J. Speech Lang. Hear. R.*, *44*, 487–496.

[243] Pollack, I., & Pickett, J. M. (1959). Intelligibility of peak-clipped speech at high noise levels. *The Journal of the Acoustical Society of America*, *31*, 14–16.

[244] Pucher, M., Türk, A., Ajmera, J., & Fecher, N. (2007). Phonetic distance measures for speech recognition vocabulary and grammar optimization. *3rd Congress of the Alps Adria Acoustics Association*, (pp. 2–5).

[245] Pye, C. (1986). Quiché Mayan speech to children. *J. Child Lang.*, *13*, 85–100.

[246] Räsänen, O., Altosaar, T., & Laine, U. K. (2008). Comparison of prosodic features in Swedish and Finnish IDS/ADS speech. In *Nordic Prosody*.

[247] Rasetshwane, D., Boston, J., Durrant, J., Yoo, S., Li, C.-C., & Shaiman, S. (2011). Speech enhancement by combination of transient emphasis and noise cancelation. In *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), 2011 IEEE* (pp. 116–121).

[248] Ratner, N. B. (1984). Patterns of vowel modification in mother–child speech. *J. Child Lang.*, *11*, 557–578.
[249] Ratner, N. B. (1988). Patterns of parental vocabulary selection in speech to very young children. *J. Child Lang.*, *15*, 481–492.
[250] Ratner, N. B., & Pye, C. (1984). Higher pitch in BT is not universal: Acoustic evidence from Quiche Mayan. *J. Child Lang.*, *11*, 515–522.
[251] Rieser, V., Keizer, S., Lemon, O., & Liu, X. (2011). Adaptive information presentation for spoken dialogue systems: Evaluation with real users. In *13th European Workshop on Natural Language Generation* (pp. 102–109).
[252] Rogers, C. L., DeMasi, T. M., & Krause, J. C. (2010). Conversational and clear speech intelligibility of bVd syllables produced by native and non-native English speakers. *J. Acoust. Soc. Am.*, *128*, 410–423.
[253] van Rooij, J. C. G. M., & Plomp, R. (1991). The effect of linguistic entropy on speech perception in noise in young and elderly listeners. *J. Acoust. Soc. Am.*, *90*, 2985–2991.
[254] Rostolland, D. (1982). Acoustic features of shouted voice. *Acta Acust.*, *50*, 118–125.
[255] Rostolland, D., & Parant, C. (1973). Distorsion and intelligibility of shouted voice. In *Symposium Speech Intelligibility* (pp. 293–304).
[256] Ryan, M. L. (1978). Contour in context. *Recent advances in the psychology of language*, (pp. 237–251).
[257] Sachs, J. (1977). The adaptive significance of linguistic input to prelinguistic infants. In C. Snow, & C. A. Ferguson (Eds.), *Talking to children. Language input and acquisition* (pp. 51–61). Cambridge: Cambridge University Press.
[258] Sankowska, J., García Lecumberri, M. L., & Cooke, M. (2011). Interaction of intrinsic vowel and consonant durational correlates with foreigner directed speech. *PSiCL*, *47*, 109.
[259] Sauert, B., & Vary, P. (2006). Near end listening enhancement: Speech intelligibility improvement in noisy environments. In *Proc. ICASSP* (pp. 493–496). Toulouse, France.
[260] Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., & Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. *J. Acoust. Soc. Am.*, *121*, 3044.
[261] Schmitt, J. F. (1983). The effects of time compression and time expansion on passage comprehension by elderly listeners. *J. Speech Lang. Hear. R.*, *26*, 373.
[262] Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *J. Acoust. Soc. Am.*, *85*, 295–312.
[263] Schum, D. J. (1996). Intelligibility of clear and conversational speech of young and elderly talkers. *J. Am. Acad. Audiol.*, *7*, 212–218.
[264] Shimron, J. (1993). The role of vowels in reading: A review of studies of English and Hebrew. *Psychol. Bull.*, *114*, 52.
[265] Sims, V. K., & Chin, M. G. (2002). Responsiveness and perceived intelligence as predictors of speech addressed to cats. *Anthrozoos: A Multidisciplinary Journal of The Interactions of People & Animals*, *15*, 166–177.
[266] Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*, 654–666.
[267] Skowronski, M. D., & Harris, J. G. (2006). Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. *Speech Comm.*, *48*, 549–558.
[268] Skowronski, M. D., & Harris, J. G. (2006). Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments. *Speech Comm.*, *48*, 549–558.
[269] Smiljanic, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *J. Acoust. Soc. Am.*, *118*, 1677–1688.
[270] Smith, C. L. (2007). Prosodic accommodation by French speakers to a non-native interlocutor. In *ICPhS* (pp. 1081–1084).
[271] Snow, E. C., & Ferguson, C. (1977). Talking to children: Language input and acquisition. Cambridge: Cambridge University Press.
[272] Song, J. Y., Demuth, K., & Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *J. Acoust. Soc. Am.*, *128*, 389–400.
[273] Stanton, B. J., Jamieson, L. H., & Allen, G. D. (1988). Acoustics-phonetic analysis of loud and Lombard speech in simulated cockpit conditions. In *ICASSP* (pp. 331–334).
[274] Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *J. Child Lang.*, *10*, 1–15.
[275] Stilp, C., & Kluender, K. (2010). Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *P. Natl. Acad. Sci. USA*, *107*, 12387–12392.
[276] Stylianou, Y., Laroche, J., & Moulines, E. (1995). High-quality speech modification based on a harmonic + noise model. In *Proc. Eurospeech* (p. 451454). Madrid, Spain.
[277] Summers, R. J., Bailey, P. J., & Roberts, B. (2010). Effects of differences in fundamental frequency on across-formant grouping in speech perception. *J. Acoust. Soc. Am.*, *128*, 3667–3677.
[278] Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *J. Acoust. Soc. Am.*, *84*, 917–928.
[279] Sundberg, J., & Nordenberg, M. (2006). Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *J. Acoust. Soc. Am.*, *120*, 453–457.
[280] Sundberg, U. (2001). Consonant specification in infant-directed speech. Some preliminary results from a study of Voice Onset Time in speech to one-year-olds. In *Lund Working Papers in Linguistics* (pp. 148–151).
[281] Sundberg, U., & Lacerda, F. (1999). Voice onset time in speech directed to infants and adults. *Phonetica*, *56*, 186–199.
[282] Suni, A., Raitio, T., Vainio, M., & Alku, P. (2010). The GlottHMM speech synthesis entry for Blizzard Challenge 2010. In *Proc. Blizzard Challenge Workshop*. Kyoto, Japan.
[283] Synnestvedt, A., Bernstein Ratner, N., & Newman, R. (2010). Voice onset time in infant-directed speech at 7.5 and 11 months. *J. Acoust. Soc. Am.*, *127*, 1853.
[284] Taal, C. H., Hendriks, R. C., & Heusdens, R. (2012). A speech preprocessing strategy for intelligibility improvement in noise based on a perceptual distortion measure. In *Proc. ICASSP* (pp. 4061–4064).
[285] Taal, C. H., Jensen, J., & Leijon, A. (2013). On optimal linear filtering of speech for near-end listening enhancement. *IEEE Signal Proc. Let.*, *20*, 225–228.
[286] Tang, Y., & Cooke, M. (2010). Energy reallocation strategies for speech enhancement in known noise conditions. In *Proc. Interspeech* (pp. 1636–1639).

[287] Tang, Y., & Cooke, M. (2011). Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In *Interspeech* (pp. 345–348). Florence, Italy.
[288] Tang, Y., & Cooke, M. (2012). Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In *Interspeech*. Portland, USA.
[289] Tang, Y., & Cooke, M. (2012). Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In *Proc. Interspeech*. Portland, USA.
[290] Tartter, V. C., Gomes, H., & Litwin, E. (1993). Some acoustic effects of listening to noise on speech production. *J. Acoust. Soc. Am.*, *94*, 2437–2440.
[291] Ternström, S., Södersten, M., & Bohman, M. (2002). Cancellation of simulated environmental noise as a tool for measuring vocal performance during noise exposure. *J. Voice*, *16*, 195–206.
[292] Thanavisuth, C., & Luksaneeyanawin, S. (1998). Acoustic qualities of IDS and ADS in Thai. In *ICSLP* (pp. 445–448). Sydney, Australia.
[293] Titze, I. R. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *J. Acoust. Soc. Am.*, *85*, 901–906.
[294] Torick, E., & Allen, R. (1966). An interphone system for "hands-free" operation in high ambient noise. *Audio and Electroacoustics, IEEE Transactions on*, *14*, 168 – 173.
[295] Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, *11*, 188–195.
[296] Trainor, L. J., & Desjardins, R. N. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychon. B Rev.*, *9*, 335–340.
[297] Traunmuller, H., & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust. Soc. Am.*, *107*, 3438–3451.
[298] Uchanski, R. M. (2005). Clear speech. In D. B. Pisoni, & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 207–235). Oxford, UK: Blackwell.
[299] Uchanski, R. M., Choi, S. S., Braida, L. D., Reed, C. M., & Durlach, N. I. (1996). Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate. *J. Speech Hear. Res.*, *39*, 494–509.
[300] Uther, M., Knoll, M., & Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner- and infant-directed speech. *Speech Comm.*, *49*, 2–7.
[301] Valentini-Botinhao, C., Maia, R., Yamagishi, J., King, S., & Zen, H. (2012). Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise. In *Proc. ICASSP*. Kyoto, Japan.
[302] Valentini-Botinhao, C., Yamagishi, J., & King, S. (2011). Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise? In *Proc. Interspeech*. Florence, Italy.
[303] Valian, V. (1980). Listening and clarity of syntactic structure. *J. Phon.*, *8*, 327–334.
[304] Valian, V., & Wales, R. (1976). What's what: Talkers help listeners hear and understand by clarifying sentential relations. *Cognition*, *4*, 155–176.
[305] Vatikiotis-Bateson, E., Chung, V., Lutz, K., Mirante, N., Otten, J., & Tan, J. (2006). Auditory, but perhaps not visual processing of Lombard speech. *J. Acoust. Soc. Am.*, *119*, 3444.
[306] Villegas, J., & Cooke, M. (2012). Maximising objective speech intelligibility by local f0 modulation. In *Interspeech*. Portland, OR.
[307] Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *J. Mem. Lang.*, *40*, 374–408.
[308] Wang, Y., Cristia, A., & Seidl, A. (2012). Prosodic prominence in infant- and adult-directed speech. In *UD Conference on Stress and Accent*. Newark.
[309] Warner, N., Otake, T., & Arai, T. (2010). Intonational structure as a word-boundary cue in Tokyo Japanese. *Lang. Speech*, *53*, 107–131.
[310] Warren, R. M. (1968). Vocal compensation for change in distance. In *International Congress of Acoustics* (pp. 61–64).
[311] Watson, P. J., & Schlauch, R. S. (2008). The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours. *Am. J. Speech Lang. Pathol.*, *17*, 348–355.
[312] Webster, J. C., & Klumpp, R. G. (1962). Effects of ambient noise and nearby talkers on a face-to-face communication task. *J. Acoust. Soc. Am.*, *34*, 936–941.
[313] van de Weijer, J. (1997). Language input to a prelingual infant. In *GALA'97 Conference on Language Acquisition*. Edinburgh, Scotland.
[314] Welby, P. (2003). French intonational rises and their role in speech segmentation. In *Eurospeech* (pp. 2125–2128). Genève.
[315] Welby, P. (2006). Intonational differences in Lombard speech: looking beyond F0 range. In *Speech Prosody* (pp. 763–766).
[316] Werker, J. F., & McLeod, P. J. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *43*, 230.
[317] World Health Organisation (2011). Burden of disease from environmental noise: Quantification of healthy life years lost in Europe.
[318] Xu, N., & Burnham, D. (2010). Tone hyperarticulation and intonation in Cantonese infant directed speech. In *Speech Prosody*.
[319] Xu, N., Burnham, D., Kitamura, C., Hamme, H. V., & Son, R. V. (2007). Vowels and tones in infant directed speech: hyperarticulation for both, but different developmental patterns. In *Interspeech* (pp. 1877–1880). Antwerp, Belgium.
[320] Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., & Tokuda, K. (2008). The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard challenge. In *Proc. Blizzard Challenge Workshop*.
[321] Yang, H., & Chen, L. (2004). Infant-directed speech does not necessarily facilitate word learning for second language learners. *National Hsin-hua Industrial Vocational Senior High School*, .
[322] Yoo, S. D., Boston, J. R., El-Jaroudi, A., Li, C.-C., Durrant, J. D., Kovacyk, K., & Shaiman, S. (2007). Speech signal modification to increase intelligibility in noisy environments. *J. Acoust. Soc. Am.*, *122*, 1138–1149.
[323] Zajdó, K. (2006). Patterns of vowel space utilization in Hungarian caregiverese addressed to young children: An evaluation of the MIPhI model. In *7th International Seminar on Speech Production* (pp. 99–106). Belo Horizonte MG, Brasil.
[324] Zampini, L., Fasolo, M., & D'Odorico, L. (2012). Characteristics of maternal input to children with Down syndrome: A comparison with vocabulary size and chronological age-matched groups. *First language*, *32*, 324–342.

[325] Zangl, R., & Mills, D. L. (2007). Increased brain activity to infant-directed speech in 6-and 13-month-old infants. *Infancy*, *11*, 31–62.
[326] Zhao, Y., & Jurafsky, D. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. *J. Phon.*, *37*, 231–247.
[327] Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, *40*, 1–95.
[328] Zorilă, T. C., Kandia, V., & Stylianou, Y. (2012). Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In *Proc. Interspeech*. Portland, USA.