Perception of the evolution of prosody in the French broadcast news style

Philippe Boula de Mareüil¹, Albert Rilliard¹, Alexandre Allauzen¹,²

¹ LIMSI-CNRS, Orsay, France
² Université Paris-Sud 11, Orsay, France
{mareuil,rilliard,allauzen}@limsi.fr

Abstract
This study makes use of advances in automatic speech processing to analyse French audiovisual archives and the perception of the journalistic style evolution regarding prosody. Three perceptual experiments were run, using prosody transplantation, delexicalisation and imitation. Results show that the fundamental frequency and duration correlates of prosody enable old-fashioned recordings to be distinguished from more recent ones. The higher the pitch is and the more there are pitch movements on syllables which may be interpreted as word-initially stressed, the more speech samples are perceived as dating back to the 40s or the 50s.

1. Introduction
The perception of the French broadcast news (BN) journalistic style is addressed in a diachronic perspective, focusing on prosody and the evolution it may have undergone since the 1940s. There is considerable variation, not only in terms of articulatory settings but also as far as prosody is concerned between Gaumont-Pathé cinematographic news dating back to World War II and radio or TV news recorded half a century later. In particular, objective measurements enabled us to quantify a decrease of mean pitch, word-initial stress (in a clitic–polysyllabic word context) and penultimate lengthening (before a pause) as time goes by [1]. The question as to whether the prosodic difference is perceivable is the purpose of the present study.

Three perceptual experiments were designed, based on prosody transplantation, delexicalisation and imitation. The overall goal of this work is to better understand the parameters which allow us to characterise an announcer style that dates back to the forties, for instance, and contrast it with a current one. The perception of the French broadcast news (BN) journalistic style is addressed in a diachronic perspective, focusing on prosody and the evolution it may have undergone since the 1940s. There is considerable variation, not only in terms of articulatory settings but also as far as prosody is concerned between Gaumont-Pathé cinematographic news dating back to World War II and radio or TV news recorded half a century later. In particular, objective measurements enabled us to quantify a decrease of mean pitch, word-initial stress (in a clitic–polysyllabic word context) and penultimate lengthening (before a pause) as time goes by [1]. The question as to whether the prosodic difference is perceivable is the purpose of the present study.

Three perceptual experiments were designed, based on prosody transplantation, delexicalisation and imitation. The overall goal of this work is to better understand the parameters which allow us to characterise an announcer style that dates back to the forties, for instance, and contrast it with a current one. The perception of the original prosody but not to the meaning of the sentences. In the TTS stimuli, it was the opposite. Experiment 3 concentrates on the prosody of the 1940s and the 1950s to compare it with the prosody of a contemporary journalist who read the same sentences in his own style and imitating an old-fashioned style. A pairwise test was set up, still making use of prosody transplantation: the sentences of each pair had the semantic content but a different prosody (3 possibilities). Listeners were asked to indicate if the presentation order was older–newer or newer–older. Can they be misled by an (amateur) imitator, who on the sole basis of prosody transplanted stimuli, listeners had access to both lexical and prosodic information of the original stimuli, but not to the recording and voice quality reminiscent of the originals. They then listened to the original stimuli. The subjects of Experiment 2 (distinct from the previous ones) first listened to the corresponding prosody transplantations which had been delexicalised using the *sartanaj* method. They then listened to the sentences read with a neutral, “modern” prosody by a standard text-to-speech (TTS) synthesis system [7]. In the delexicalised stimuli, listeners thus had access to the original prosody but not to the meaning of the sentences. In the TTS stimuli, it was the opposite.

In Experiments 1 and 2, subjects were asked to assign a date (between 1940 and 1999) to the speech excerpt they listened to. Each of these pilot experiments consisted of two blocks, corresponding to different conditions. The subjects of Experiment 1 first listened to synthetic stimuli, the prosody of which was copied from audio archive utterances. With the prosody-transplanted stimuli, listeners had access to both lexical and prosodic information of the original stimuli, but not to the recording and voice quality reminiscent of the originals. They then listened to the original stimuli. The subjects of Experiment 2 (distinct from the previous ones) first listened to the corresponding prosody transplantations which had been delexicalised using the *sartanaj* method. They then listened to the sentences read with a neutral, “modern” prosody by a standard text-to-speech (TTS) synthesis system [7]. In the delexicalised stimuli, listeners thus had access to the original prosody but not to the meaning of the sentences. In the TTS stimuli, it was the opposite.

In Experiments 1 and 2, subjects were asked to assign a date (between 1940 and 1999) to the speech excerpt they listened to. Each of these pilot experiments consisted of two blocks, corresponding to different conditions. The subjects of Experiment 1 first listened to synthetic stimuli, the prosody of which was copied from audio archive utterances. With the prosody-transplanted stimuli, listeners had access to both lexical and prosodic information of the original stimuli, but not to the recording and voice quality reminiscent of the originals. They then listened to the original stimuli. The subjects of Experiment 2 (distinct from the previous ones) first listened to the corresponding prosody transplantations which had been delexicalised using the *sartanaj* method. They then listened to the sentences read with a neutral, “modern” prosody by a standard text-to-speech (TTS) synthesis system [7]. In the delexicalised stimuli, listeners thus had access to the original prosody but not to the meaning of the sentences. In the TTS stimuli, it was the opposite.

2. Corpus and methodology

2.1. Method
Our three experiments are based on prosody transplantation on a synthetic voice, having the content and the overall prosody of utterances vary. The prosody transplantation paradigm enables the fundamental frequency and duration correlates of prosody to be disentangled from recording conditions and voice quality effects. Even though we are aware of its shortcomings and the artefacts it entails, the method proved appropriate to sort out the contribution of prosody in different languages and situations [3][4]. Delexicalisation is intended to render the sentences unintelligible: here, every vowel was replaced by an [a], every plosive by a [t], every fricative by an [s], every nasal by an [n], every glide by a [j] and every liquid by an [x]. The latter was preferred to the [l] initially proposed by [5], who inspired this *sartanaj* mapping, because [tx] is a much more natural and frequent cluster than [tl]. This way of ensuring unintelligibility was also preferred to older methods like the reiteration of [ma ma] syllables because it is a little bit more ecological and preserves some rhythmic patterns, throughout phonotactics. As for impersonation, it has been used in various studies to investigate what is stable across speakers in intonation, which features are salient and which habits are hard to imitate (e.g. [6]).

In Experiments 1 and 2, subjects were asked to assign a date (between 1940 and 1999) to the speech excerpt they listened to. Each of these pilot experiments consisted of two blocks, corresponding to different conditions. The subjects of Experiment 1 first listened to synthetic stimuli, the prosody of which was copied from audio archive utterances. With the prosody-transplanted stimuli, listeners had access to both lexical and prosodic information of the original stimuli, but not to the recording and voice quality reminiscent of the originals. They then listened to the original stimuli. The subjects of Experiment 2 (distinct from the previous ones) first listened to the corresponding prosody transplantations which had been delexicalised using the *sartanaj* method. They then listened to the sentences read with a neutral, “modern” prosody by a standard text-to-speech (TTS) synthesis system [7]. In the delexicalised stimuli, listeners thus had access to the original prosody but not to the meaning of the sentences. In the TTS stimuli, it was the opposite.

Experiment 3 concentrates on the prosody of the 1940s and the 1950s to compare it with the prosody of a contemporary journalist who read the same sentences in his own style and imitating an old-fashioned style. A pairwise test was set up, still making use of prosody transplantation: the sentences of each pair had the semantic content but a different prosody (3 possibilities). Listeners were asked to indicate if the presentation order was older–newer or newer–older. Can they be misled by an (amateur) imitator, who on the sole basis of prosody would succeed in sounding more “original” than originals? Since imitators who are often prone to caricature are likely to exaggerate some cues, a positive answer to that question would yield insights into what is important in the perception of an old-fashioned prosody.

2.2. Corpus
The whole corpus we relied on mainly stems from the ECHO project described in [8] and [1]. For Experiments 1 and 2, 30 utterances were selected among male speakers from 10 hours of audiovisual BN archives, between 1941 and 1997: 5 utterances per decade, 10-second long on average. For Experiment 3, we kept 10 stimuli of Experiments 1 and 2 (5 from the 40s and 5 from the 50s), and added another 5 stimuli from war archives. A contemporary journalist was asked to normally read their orthographic transcripts and to imitate the 1940–1950 style, not listening to the original excerpts but relying on his own representations of the Gaumont-Pathé style.
Thus, there were 15 sentences × 3 styles: archive original (O), contemporary (C) and imitation (I). Their prosody was once more copied and imposed on a diphone voice, because it would have been too easy for listeners to identify the voice quality of the actual recordings. From the resulting prosody transplantations, 45 pairs of stimuli were built up: the order of appearance could be OC or CO, OI or IO, IC or IC, with a 1-second pause in between. The average duration of a stimulus pair was 18 seconds.

The prosody transplantation method and the diphone speech synthesis system which were used are described in [7] and [3]. Prosodic parameters are extracted from the original stimuli and grafted onto the diphone base derived from a male speaker whose prestored speech units are concatenated. The pitch and duration parameters are then handled thanks to the TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add) algorithm [8]. Energy is not processed, it is only normalised.

2.3. Prosodic analysis of the stimuli

The overall prosodic analysis is presented in [9]. It is based on fundamental frequency (F₀) measurements provided by the PRAAT software (http://www.praat.org) with standard settings in which F₀ values below 75 Hz appear as undefined. Experts in prosody were asked to pinpoint prominent syllables of the experiment corpus, but no consensus emerged.

To approach initial stress, clitic–polysyllabic word sequences were considered, clitic words being defined as belonging to a list of about 30 monosyllabic words among the most frequent tokens of the corpus. The pitch difference between the polysyllabic word-initial and the clitic vowels was then computed (e.g. between [a] and [e] in les artistes “the artists”); and the percentage of occurrences in which this difference is greater than 3 semitones was assumed to be a good estimate of initial stress acoustic correlates. Comparative results for the experiment corpus (168 clitic–polysyllabic word contexts) and the whole corpus (12,158 contexts) are shown in Table 1. In both cases, a decrease of what can be interpreted as initial stress is observed. Also, an outstanding decrease of mean pitch is noteworthy: roughly from 170 Hz in the 40s and the 50s to 140 Hz in the 80s and the 90s, for both the experiment and the whole corpus. There were too few prepausal contexts to investigate penultimate lengthening before a pause.

Table 1. Percentage of clitic–polysyllabic contexts in which the pitch rise is greater than 3 semitones

<table>
<thead>
<tr>
<th></th>
<th>40s–50s</th>
<th>60s–70s</th>
<th>80s–90s</th>
</tr>
</thead>
<tbody>
<tr>
<td>experiment corpus</td>
<td>41</td>
<td>27</td>
<td>24</td>
</tr>
<tr>
<td>whole corpus</td>
<td>28</td>
<td>21</td>
<td>18</td>
</tr>
</tbody>
</table>

In Experiment 3, the journalist spoke louder while imitating the Gaumont-Pathé style, but this is not accounted for in the prosody transplantation, contrary to what follows. The journalist’s mean pitch rises from 142 Hz in his contemporary style up to 173 Hz in imitation, while the mean pitch of the original stimuli is 171 Hz. The percentage of clitic–polysyllabic word contexts exhibiting a pitch rise greater than 3 semitones is also higher in imitation (37%) than in the contemporary style (28%) — this percentage is 47% on the 15 original stimuli selected in the 40s and the 50s. These figures, which have to be compared with those of Table 1, are consistent with a decrease of initial stress over the years.

A closer look at accentual patterns enables us to better understand these mean and more local pitch differences. Before a pause especially, initial stresses are much more marked in the imitation than in the contemporary style. An example is the utterance-final phrase un vif succès (“a great success”), where pitch rises up to 250 Hz in the I-style, whereas it does not exceed 190 Hz in the C-style.

A stylisation of pitch in which each vowel is defined by an initial target, a final target, and possibly an intermediate target is extended to other utterance-final accentual phrases in Fig 1. This figure matches vowel durations and discards consonants. Thus, it should not be viewed as a faithful F₀ contour but as a more readily readable visualisation of the style effect.

Figure 1: Stylised melody of utterance-final vowels in the stimuli of Experiment 3, the C-style (full red lines) and the I-style (dotted black lines). The y-axis reports F₀ values in Hz.

2.4. Tasks and participants

After a few examples to get familiarised with the type of speech to be judged, the subjects of Experiment 1 listened to 30 prosody transplantations and then the 30 original stimuli.

After a familiarisation phase, the subjects of Experiment 2 also listened to 60 stimuli. The first 30 stimuli are the delexicalised versions of the prosody transplantations of Experiment 1. As explained above, an example of a sentence beginning is [ta tata atsasja ta nasa tatra] for du côté accessoires de nouveaux progrès (“on the accessory side new advances”). The last 30 stimuli result from the non-delexicalised output of a TTS system with the same diphone voice and a prosody generated by rules [7].

In each block of Experiments 1 and 2 (prosody transplantation and original stimuli, delexicalisation and TTS prosody), the stimuli were presented in a random order which was different for each subject. Participants could listen to each stimulus as often as they wished, through a web-based interface. A slider allowed them to assign a date to each stimulus, but it was not possible to correct previous responses once a new stimulus was displayed.

In Experiment 3, the pair order was also randomised, and the interface to play stimuli was the same. Yet, listeners had to make a binary forced choice, to indicate for each pair if the utterance order of appearance was older–newer or newer–older. Each experiment lasted 20–30 minutes.

Twenty-six subjects (18 males, 8 females, aged 34 on average) took part in Experiment 1, twenty-nine subjects (22 males, 7 females, aged 31 on average) took part in Experiment 2 and another twenty-six subjects (12 males, 14 females, aged 32 on average) took part in Experiment 3 — they had not participated in the first experiment. They all had French as their mother tongue and had no known hearing impairment. Before the test, they were requested to rate their ability to distinguish old or recent recordings. In the first two experiments (respectively Experiment 3), this ability was self-estimated at 3 (respectively 3.5) on a 1–5 scale.

3. Perceptual results

3.1. Experiments 1 and 2

To analyse perceptual results, we first considered the responses of Experiment 1 per decade. Partially due to the difficulty of the task (judge the perceived date of a recording) and partly due to the unavoidable differences amongst the stimuli of a same decade, listeners show an important variability in their answers which cannot be accounted for if the results are described stimulus by stimulus. However, robust tendencies can be extracted from the data when stimuli
are grouped. From the results of Experiment 1, a cluster analysis was applied on the confusion matrices obtained by counting the number of stimuli of each decade recognised in this or that decade. A hierarchical agglomerative clustering was performed on both prosody transplantation and the original stimuli (using the Euclidean distance). Results are not reported due to lack of space, but they show that the 40s and the 50s break away from the other decades, and that the stimuli recorded during the 60s and the 70s are regrouped. The 90s receive good recognition scores (i.e. the perceived decade is the same one as the real one), with some confusion with the 80s. The opposition between the 40s–50s and the other decades is also found in the prosody transplantation condition of Experiment 1, but there is more confusion between the 60s–70s and the 80s–90s. Therefore, on the basis of all the information contained in the original stimuli, listeners seem to categorise the proposed utterances in three categories of twenty years each (hereafter referred to as epochs).

In the following, results were accordingly pooled for the 40s and the 50s, the 60s and the 70s, the 80s and the 90s. In Figure 2, the x-axis stands for the real dates and the y-axis stands for the perceived dates. The perceived dates of the 40s and the 50s prosody transplants are overestimated with respect to the original stimuli. In other words, the old-fashioned nature of these stimuli is better perceived when the very peculiar voice quality and recording quality are heard.

![Figure 2: Results of Experiments 1 and 2. The x-axis represents the real date (averaged for each epoch). The y-axis represents the averaged perceive dates for the original stimuli and prosody transplantation (Experiment 1), delexicalisation and the TTS prosody (Experiment 2).](image)

In both conditions, response patterns are satisfactorily increasing. Nevertheless, perception may have been biased by the understanding of the informational content. In spite of our efforts to select lexically-unmarked sequences, news items most often refer to topics which give indications to the listeners (e.g. war vs agriculture or females’ status). Experiment 2 was intended to rule out this drawback.

In Experiment 2, the task happened to be quite hard with the delexicalised samples, harder than on the sole basis of the thematic content — with the TTS prosody. Results suggest that the content is more reliable to date a speech excerpt than is prosody. Though, a smooth upward slope is observed in the perception curve is steeper from the 40s–50s to the 80s–90s than for the prosody-only stimuli. These observations show that the date perceived by the listeners consistently rises through the different epochs and that each type of information tested may participate in this perception: the recording and voice quality is the main cue, the informational content and then prosody are secondary cues.

When we compare the results of Experiments 1 and 2, we can see that, with respect to the original stimuli, the perceived dates of prosody transplants are slightly closer than are the perceived dates of the TTS prosody stimuli: e.g. 1981 (prosody transplantation) vs 1979 (TTS prosody) for the period spanning the 1980s and the 1990s, while the average perceived date is 1984 for the original stimuli. This holds for each epoch, even if for the one which covers the 1960s and the 1970s the perception of the TTS prosody stimuli (average date = 1971) is closer to reality (average date = 1968). A kind of default response (1970) may explain this behaviour.

### 3.2. Experiment 3

The listeners of Experiment 3 had to judge, between two utterances, the one that sounded older. For each type of pair involving the prosody transplantation of original, contemporary or imitation excerpts, results were rated 1 if the first sentence of the pair was judged the older, and 0 otherwise. Results were thus expressed in terms of percentages of “older–newer” answers. As the pairs were presented in both ways, they were all flipped so as to have the type of pair (OC, IC or OI) plus a parameter indicating the order of presentation. For example, an OC pair was coded “OC order 1”, a CO pair was coded “OC order 2”, and the results were expressed as the percentage of original stimuli perceived as older than contemporary stimuli. Results are represented in Table 2.

An ANOVA (completely randomised two-factorial design) was conducted on the answers of Experiment 3, coded as explained above (with the percentages of “older–newer” responses). The two fixed factors were the type of Pair presented to the listeners (3 levels: OC, IC, OI) and the Order of presentation of the pair (2 levels). Again, the significance level was set at 0.01.

The type of Pair had a significant effect on listeners’ answers ($F_{2,1554} = 73.02, p < 0.01$). The pair Order of presentation was not significant, as the interaction between Pair and Order: listeners answered “older–newer” as often as “newer–older”.

Subsequent post-hoc comparisons (Tukey’s HSD tests) with an α level of 0.01, performed on the Pair factor, show that OC and IC pairs form a homogeneous subset, but differ significantly from OI pairs. The prosody of imitations and original stimuli is perceived as older than that of contemporary

<table>
<thead>
<tr>
<th>Epoch</th>
<th>Proportion of older answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1940-1950</td>
<td>0.95</td>
</tr>
<tr>
<td>1950-1960</td>
<td>0.98</td>
</tr>
<tr>
<td>1960-1970</td>
<td>0.99</td>
</tr>
<tr>
<td>1970-1980</td>
<td>0.96</td>
</tr>
<tr>
<td>1980-1990</td>
<td>0.94</td>
</tr>
</tbody>
</table>

For Experiment 1, both factors have a significant effect: the perceived dates significantly increase with the Epoch ($F_{2,1554} = 721.13, p < 0.01$), and the mean judgments significantly differ between the two Types of stimuli ($F_{2,1554} = 13.23, p < 0.01$), mainly due to the more accurate rating of the oldest original stimuli. A significant interaction between the Epoch and the Type of stimulus was found ($F_{2,1554} = 47.56, p < 0.01$). The curves corresponding to the original stimuli and the stimuli with prosody transplantation (see Fig. 2) show that the main difference between the two types of stimuli across the 3 epochs rests in the 40s–50s epoch. For this war or post-war era, the original stimuli are well classified, whereas for the prosody-transplanted stimuli, the mean perceived date is 1960.
studies in over 80% of cases. Between imitations and original stimuli, responses are more balanced: they behave in the same way when their prosody is compared with that of contemporary stimuli, even though the original prosody is found slightly older when compared with the imitation prosody. The latter result is in keeping with the decreasing tendency to word-initial stress reported in 2.3, greater in the O-style than in the I-style and greater in the I-style than in the C-style. However, this tendency is to take into account with caution, since a chi-square test shows no significant difference between the results obtained by the listeners on OI pairs and a random distribution. In other words, the journalist succeeded quite well in imitating the Gaumont-Pathé prosody.

Table 2. Results of Experiment 3: the percentage of answers is given for each type of pair in both orders of presentation.

<table>
<thead>
<tr>
<th></th>
<th>OC</th>
<th>IC</th>
<th>OI</th>
<th>%younger–older</th>
<th>CO</th>
<th>CI</th>
<th>IO</th>
</tr>
</thead>
<tbody>
<tr>
<td>%older–newer</td>
<td>87</td>
<td>88</td>
<td>52</td>
<td>79</td>
<td>88</td>
<td>59</td>
<td></td>
</tr>
</tbody>
</table>

In their comments (expressed in non-specialist terms), some listeners of Experiment 3 indicated that they relied on pitch range and height, finality, syllable duration and pauses. A higher pitch and more pitch movements, especially at the end of utterances, are associated to old recordings. Inversely, a more monotonous intonation is judged as more recent, which is in accordance with our acoustic analyses.

A more in-depth analysis of the listeners’ answers was carried out for each pair of prosody-transplanted stimuli. While mean results for OC comparisons show a clear tendency to judge the originals as older than the contemporary stimuli, for two samples listeners could not decide which one of the C-style or the O-style was the older. Moreover, with the same two original stimuli in OI comparisons as well as in a third OI pair, the imitations were clearly perceived as older than the originals (in over 70% of answers, whereas mean results for these comparisons are close to chance). These three original stimuli are interesting because they do not convey very marked prosodic cues, contrary to the corresponding imitations.

In order to match these subjective ratings and acoustic parameters, a prosodic analysis of all the stimuli of Experiment 3 was performed. Since there are too few occurrences to count clitic–nonclitic pitch differences and prepausal contexts for each utterance, prosodic variation was quantified in terms of mean pitch, mean syllable duration and 9th decile of all FO values (extracted with Praat).

Correlations between perception results and overall prosodic measurements were computed by calculating the difference of these prosodic measurements in each stimulus pair: e.g. O-style mean FO1–I-style mean FO2 (in semitones), for each OI pair in its order of presentation. Correlations with perception are \( r = 0.89 \) for mean pitch difference, \( r = 0.38 \) for mean syllable duration difference, and \( r = 0.77 \) for the 9th FO decile difference. These results point out that, whereas the correlation with syllable durations is weak, the higher the pitch of an utterance (on average and in terms of FO peaks), the older its perception. The perceptual salience of high FO peaks may be related to the presence of more word-initial stress in old-fashioned sentences. These final findings should be taken with caution, as they concern only a few speech samples, but they corroborate the prosodic analyses performed on the corpus [1].

### 4. Discussion

A prosody transplantation-based methodology was proposed, the application of which successfully evidenced audible diachronic differences in the French journalistic style prosody. Some of these differences are attributable to a decrease of mean pitch and word-initial stress, for which quite robust cues were found. We do not claim that we are facing a change in progress which goes beyond the scope of journalist style, but since the latter is often associated to word-initial stress (e.g. [2]), this decrease is an interesting outcome.

We saw that acoustically (especially prosodically) and perceptually the 40s and the 50s depart from the following decades. This tendency was confirmed by the imitation-based experiment we undertook, which showed good correlations between perception and prosodic parameters. In the sixties, there were no dramatic breakthroughs in sound capture techniques but rather in recording restitution and broadcasting techniques [10]. The advent of the transistor, the massive use of the tape, the popularisation of TV (as later on the liberalisation of the FM band in the eighties) have modified listening practices which become more democratic. Journalists no longer speak to living-rooms or movie theatres but to particular (especially young) listeners. They no longer talk to an audience in the solemn way of learned teachers or like village fair advertisers. They get closer to conversations between peers, which is facilitated by close-talking microphones. The technique thus impacts the sociology of the audience, the representation of the audience which journalists have, and therefore the journalists’ speaking style. This adaptation is interesting in the light of communication models such as Bell’s [11] audience design.

Further investigation is required to explore time-related phenomena. In particular, an extension of this study to broadcast news of the early 21st century is foreseen (and scheduled).

### 5. Acknowledgements

This work was partially financed by OSEO under the Quero program. Our thanks go to Laurent Vinet and the French Institut National de l’Audiovisuel (INA) for the audiovisual corpus and its transcription which were made available for the EURODELPHES and ECHO projects. We express our gratitude to Benoît Hervieu, the journalist who was recorded and to all the subjects who volunteered to participate in the listening tests.

### 6. References


